Inégalité, mobilité et hétérogénéité sur le marché du travail : Contribution Empiriques et Méthodiques

Stéphane Bonhomme

UNIVERSITÉ PARIS 1 – PANTHÉON SORBONNE
UFR SCIENCES ÉCONOMIQUES

THÈSE

Pour obtenir le grade de

DOCTEUR EN SCIENCES ECONOMIQUES

Présentée et soutenue publiquement par

Stéphane BONHOMME

le 23 Mai 2006

Inégalité, Mobilité et Hétérogénéité sur le Marché du Travail : Contributions Empiriques et Méthodologiques

Directeur de Thèse

Jean-Marc ROBIN, Professeur à l'Université Paris 1 - Panthéon Sorbonne

Jury :

Jean-Pierre Florens (Rapporteur), Professeur à l'Université Toulouse I

Eric Renault (Rapporteur), Professeur à la North Carolina University (USA)

Patrick Sevestre, Professeur à l'Université Paris1 - Panthéon Sorbonne

Manuel Arellano, Professeur au CEMFI (Madrid)

Année 2006

Remerciements

Mes premiers remerciements vont à Jean-Marc Robin pour sa formidable énergie dans le travail et ses conseils avisés. Je remercie aussi Grégory Jolivet pour son calme et son humour. Au CREST et au CEMFI, “juniors” et “seniors” m’ont été précieux pendant ces trois ans. Dans le désordre: Laurent Gobillon, Fabrice Murtin, Gema Zamarro, Luc Behagel, Guy Laroque, Per Engström, Patrick Gagliardini, Francis Kramarz, David Blau, Fabien Postel-Vinay, ont chacun contribué à ce travail. Je remercie l’ensemble des “thésards” des deux endroits. Un merci spécial à Manuel Arellano de m’avoir accueilli au CEMFI. Merci aussi et peut-être surtout à Pierre Cahuc d’avoir rendu cette aventure possible.

Finalement, alors que s’achève cette période de ma vie, je voudrais remercier Carmen qui m’accompagne. Et dire toute ma gratitude à ma mère, Christine, disparue il y a deux ans. Cette thèse lui est dédiée.

Résumé

Ce travail rassemble quatre essais consacrés à l'étude de l'hétérogénénité et des dynamiques individuelles sur le marché du travail. Le premier chapitre met en évidence le lien entre mobilité (ou inertie) salariale et le degré de persistence des inégalités. Nous employons une méthode statistique simple et originale pour étudier les trajectoires individuelles de salaires, et l'appliquons à des données françaises couvrant la période 1990-2002. Nous trouvons que la récession du début des années 1990 a été associée à une augmentation sensible des inégalités longitudinales.

Dans le deuxième chapitre, nous étudions l'effet de la mobilité entre emplois sur les corrélations entre salaires et caractéristiques non salariales. Dans notre modèle, de fortes préférences pour ces caractéristiques ne se traduisent pas nécessairement en corrélations négatives si les frictions de mobilité sont importantes. Sur données européennes, nous estimons de fortes préférences pour certaines caractéristiques telles que le type de travail ou la sécurité de l'emploi, ainsi que des differentiels de salaires très faibles entre niveaux d'aménités.

Les chapitres 3 et 4 introduisent une méthode de modélisation de l'hétérogénéité inobservée: l'analyse en composantes indépendantes. Celle-ci diffère de l'analyse en composantes principales en ce que les facteurs ne sont pas supposés simplement non corrélés, mais statistiquement indépendants. Cette hypothèse permet d'identifier les facteurs de manière non ambiguë. Nous appliquons notre méthode à des données de salaire de d'éducation en France pour l'année 1995. Nos résultats suggèrent une relation complexe et multidimensionnelle entre le niveau d'études et son rendement sur le marché du travail.

Mots clefs: Inegalité, mobilité, dynamique salariale, copules, différences compensatrices, salaires hédoniques, mobilité entre emplois, aménités, hétérogénéité inobservée, modèles à facteurs, moments d'ordre élevé, analyse en composantes indépendantes, estimation nonparamétrique, deconvolution, transformation de Fourier.

Abstract

This dissertation contains three essays devoted to the study of the heterogeneity and the dynamics of individuals on the labor market. The first chapter highlights the link between wage (im-)mobility and the persistence of wage inequality. We use a simple and original statistical method to analyze wage trajectories. We estimate the model on French data for the period 1990-2002. We find that the recession of the early nineties was associated to an increase in the level of longitudinal inequality.

In the second chapter, we study the effect of job-to-job mobility on the correlation between wages and non-wage job characteristics. In the model, strong preferences for these amenities do not necessarily translate into negative correlations if mobility costs are highly heterogeneous. On European data, we estimate strong preferences for some amenities such as the type of work or job security, together with small wage differentials between different levels of amenities.

In chapters 3 and 4, we introduce a method to model unobserved heterogeneity: Independent Component Analysis. The latter differ from Principal Component Analysis in that factors are assumed statistically independent, and not merely uncorrelated. This additional assumption yields unambiguous identification of factor loadings. We apply the method to French data on wage and education. Our results suggest that the relation between wages and education is complex and multidimensional.

Keywords: Inequality, mobility, earnings dynamics, copulas, compensating differentials, hedonic wages, job mobility, amenities, unobserved heterogeneity, factor models, high-order moments, Independent Component Analysis, nonparametric estimation, deconvolution, Fourier transform.

JEL codes: C13, C14, C33-35, D30, D63, J22, J31-33, J63, J64 and J81.

Introduction

La littérature microéconométrique contemporaine décrit le marché du travail comme un lieu où les dynamiques sont complexes et l'hétérogénéité omniprésente. Une image de l'économie à un moment donné est la résultante de choix intertemporels contraints effectués par des individus différents. Depuis une trentaine d'années, le recueil et l'exploitation de nouvelles données microéconométriques ont permis de dégager ces enseignements et de montrer leur pertinence empirique (voir Heckman, 2001, pour une synthèse). Les quatre textes rassemblés ici retiennent ce cadre général pour point de départ.

Les deux premiers chapitres soulignent l'importance de la mobilité, définie comme changement d'état au cours du temps, dans l'étude des inégalités. Dans le premier chapitre, la mobilité des salaires est mise en évidence comme un facteur affectant la persistence des inégalités salariales. Dans le chapitre 2, il est montré que la mobilité entre emplois peut avoir une forte influence sur le lien entre salaires et caractéristiques non salariales, et partant sur les inégalités en emploi.

Ces deux études insistent sur la nécessité de prendre en compte l'hétérogénéité inobservée, que les données habituellement utilisées ne permettent pas de mesurer. Dans les chapitres 3 et 4, une méthode de modélisation de l'hétérogénéité est introduite et analysée. L'application à des données de salaire et d'éducation suggère une relation complexe et multidimensionnelle entre le niveau d'études et son rendement sur le marché du travail.

Chapitre 1 : Mobilité des salaires et persistence des inégalités

Dans le premier chapitre, coécrit avec Jean-Marc Robin, nous explorons l'impact de la mobilité salariale sur les inégalités. Pour cela, nous introduisons une nouvelle méthodologie qui nous permet d'analyser inégalité et mobilité dans un cadre commun. Nous appliquons notre méthode aux données de l'Enquête-Emploi de l'INSEE pour la période 1990-2002.

Inégalité et mobilité

Depuis la fin des années 1970 et l'explosion des inégalités de salaire aux Etats-Unis, les publications académiques consacrées aux inégalités se sont multipliées (voir Levy et Murnane, 1992, et Katz et Autor, 2001). Cet effort de recherche, associé a des données de meilleure qualité, a permis de documenter avec précision les différences d'inégalité entre pays : les pays anglo-saxons (Angleterre, Canada) suivant l'évolution américaine, les pays d'Europe continentale (France, Allemagne) connaissant une certaine constance de inégalités au cours des dernières décennies. De plus, de nombreux travaux ont décrit les "gagnants" et les "perdants" de l'évolution des inégalités, et plusieurs explications ont été proposées et contrastées.[1] Dans leur majorité pourtant, les mesures d'inégalités utilisées habituellement dans ces travaux sont statiques. En effet, dans la plupart des articles consacrés aux inégalités de salaires ou de revenus, celles-ci sont mesurées à un instant donné. En conséquence, le degré de persistence des inégalités n'est pas pris en compte.

Des distributions de coupe...

Prenons l'exemple des inégalités salariales. Le graphique 1 représente la densité de distribution des salaires mensuels des hommes en emploi, mesurée pour la France et

[1] Voir par exemple, sur la question du rendement des caractéristiques individuelles inobservées, Juhn, Murphy et Pierce (1993) ainsi que le récent article de Lémieux (2006).

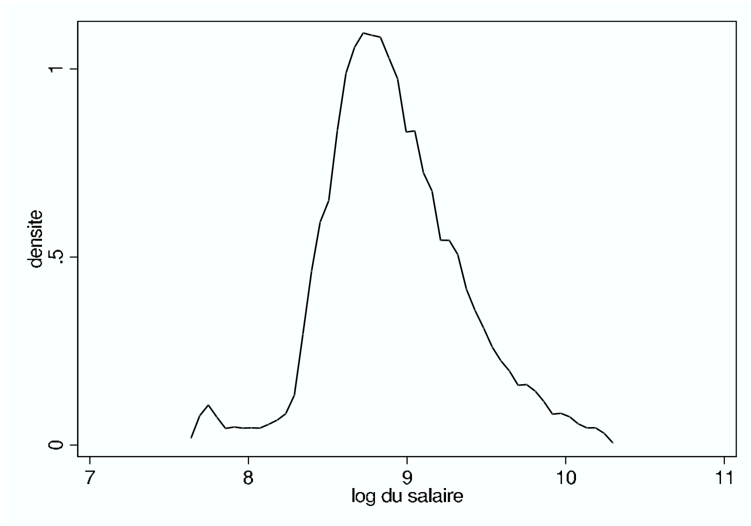


Figure 1: Densité de salaires (en log(Francs)), hommes en emploi, Enquête-Emploi 1995

l'année 1995.[2] Sur l'axe horizontal sont reportés les logarithmes des salaires, sur l'axe vertical leur densité de probabilité. Pour étudier les inégalités de salaires, l'économètre cherchera le plus souvent à décrire une telle distribution. Par exemple, il estimera l'importance relative des classes moyennes (salaires autour du salaire médian) par rapport aux "riches" et "pauvres". Il mesurera le niveau d'inégalité en calculant un ou plusieurs indices (Gini, variance...) qui quantifient la dispersion de cette distribution. Puis il utilisera ces indices pour documenter des évolutions, ou des différences entre pays ou régions par exemple.

... aux distributions longitudinales...

Pourtant, les conclusions fondées sur ces distributions de coupe peuvent être incomplètes, voire trompeuses. Imaginons, a titre d'illustration, deux économies, A et B, dont la distribution de salaires de coupe est invariante dans le temps, égale à celle représentée sur le graphique 1. Supposons que dans l'économie A, les individus gar-

[2]Les données de l'Enquête-Emploi, décrites dans le chapitre 1, ont servi à construire ce graphe ainsi que le graphique 3 plus bas.

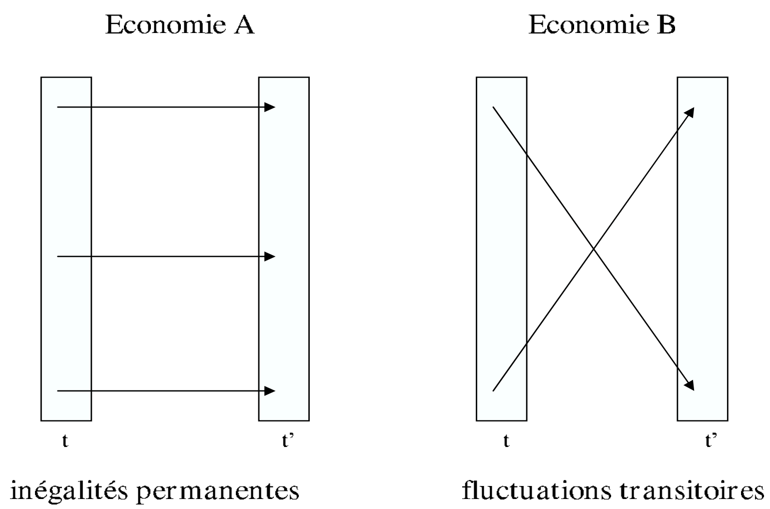


Figure 2: Inégalité et mobilité

dent le même salaire d'une date à l'autre. A l'opposé, dans l'économie B, les riches à l'année t deviennent pauvres à l'année $t+1$, puis de nouveau riches à $t+2$, de manière à ce qu'il y ait, d'une année à l'autre, alternance des positions relatives dans l'échelle des salaires. Ces deux situations sont représentées sur le graphique 2.

Il est intuitif que l'économie B est plus égale que A, même si son niveau d'inégalités de coupe est le même. Une manière d'illustrer cette intuition est de considérer la moyenne des revenus individuels sur une longue période. Dans l'économie B, chacun a le même revenu de long terme, ce qui signifie qu'il n'y a pas d'inégalités (intertemporelles) dans cette économie! La situation est très différente dans l'économie A, où le niveau longitudinal d'inégalités est égal au niveau de coupe.

... via la mesure de la mobilité

Le graphique 2 illustre le fait qu'une économie où la mobilité salariale est forte peut être intertemporellement moins inégale qu'une autre, à niveau d'inégalités de coupe identique. En d'autres termes, la mobilité des salaires affecte le degré de persistence des inégalités.[3]

[3]Il est à noter que cet argument suppose des individus neutres au risque. Voir sur ce point Gottschalk et Spolaore (2002).

Ces considérations prennent un relief particulier dans le cas de la France dans les années 1990-2000. En effet, sur cette période comme sur le siècle (Piketty, 2001), les inégalités salariales de coupe varient peu. Cette relative stabilité cache-t-elle une crise plus profonde, autour de la récession– exceptionnellement sévère– de 1993 ? L'évolution de la mobilité et des inégalités longitudinales sur la période révèle-t-elle une tendance plus marquée ?

Un outil : les copules

Pour tenter de répondre à ces questions, nous modélisons les trajectoires individuelles de salaires. Nous prenons en compte le niveau et les changements de salaire, ainsi que les épisodes de chômage. Ce dernier point est important pour décrire la situation d'un pays comme la France. Un ingrédient essentiel dans le cadre que nous proposons est la modélisation de la mobilité salariale.

Mobilité absolue et relative

Deux approches de la mobilité sont habituellement distinguées (Fields et Ok, 1999). La première modélise le lien entre les niveaux absolus de salaire à différentes dates. Une mesure habituelle de ce type de mobilité est la corrélation entre salaires à t et $t+1$. La deuxième approche, fréquemment utilisée en sociologie (par ex. McClendon, 1977) et en économie (Shorrocks, 1978, Buchinsky et Hunt, 1999), considère le lien entre niveaux relatifs de salaires. Formellement, notons Y_t le salaire à la date t pour un individu donné, et Y_{t+1} son salaire à $t+1$. L'approche relative de la mobilité étudie le lien entre $F_t(Y_t)$ et $F_{t+1}(Y_{t+1})$, où F_t et F_{t+1} sont les fonctions de répartition de Y_t et Y_{t+1} respectivement. Les quantités $F_t(Y_t)$ et $F_{t+1}(Y_{t+1})$, comprises entre 0 et 1, s'interprètent comme les rangs d'un individu dans les deux distributions de salaire. Dans notre contexte, cette seconde approche est plus naturelle que la modélisation des niveaux de salaires Y_t et Y_{t+1}, puisqu'elle permet de séparer la modélisation des inégalités de coupes de celle de la mobilité relative.

Modéliser la mobilité relative

Le graphique 3 présente les revenus relatifs en 1995 et 1996. Chaque point sur le graphique correspond à un individu. Sa position relative en 1995 est reportée sur l'axe horizontal, sa position en 1996 sur l'axe vertical. Autrement dit, le rang en 1995, $F_{1995}(Y_{1995})$, est reporté en abscisse, le rang en 1996, $F_{1996}(Y_{1996})$, en ordonnée. Par exemple, un individu situé sur la première diagonale a la même position dans l'échelle des salaires, relativement aux autres,[4] en 1995 et 1996. Une manière de résumer l'information contenue dans le graphique 3 est de diviser l'ensemble des positions relatives (l'intervalle entre 0 et 1) en quintiles ou déciles, et de calculer une matrices de transition entre ces états. Cette approche a été souvent appliquée pour étudier la mobilité relative. Les matrices de transition ont l'avantage de permettre une modélisation fine (non nécessairement linéaire) des corrélations. Lorsque le nombre d'états tend vers l'infini, la description des données est parfaite. L'objet statistique que l'on obtient alors est appelé une copule.

Qu'est-ce qu'une copule?

Une copule décrit la densité jointe des positions relatives. Cet outil est souvent utilisé en finances pour modéliser une dépendance fine entre risques de défaut (Nelsen, 1998). Une des contributions du chapitre 1 est de montrer que les copules peuvent être utiles pour modéliser la mobilité relative des salaires en économie du travail. En effet, la densité des niveaux de salaires est par définition égale au produit des densités marginales et de la copule des positions relatives. Formellement, si $f_{t,t+1}$ dénote la densité du couple de salaires (Y_t, Y_{t+1}), et f_t et f_{t+1} dénotent les densités marginales, alors :

$$f_{t,t+1}(y_t, y_{t+1}) = f_t(y_t) f_{t+1}(y_{t+1}) c_{t,t+1}(F_t(Y_t), F_{t+1}(Y_{t+1})),$$

oû $c_{t,t+1}$ est la (densité de copule) associée au couple de salaires. Une manière de traduire ce résulat est de dire que pour décrire une trajectoire absolue de salaires, il

[4] On a contrôlé de l'effet de l'âge pour tracer le graphique 3.

suffit de décrire une trajectoire relative, une fois les distributions de coupe prises en compte. Ceci implique que les inégalités de coupe et la mobilité relative peuvent être modélisées indépendemment les unes des autres. Cette propriété donne au chercheur une grande flexibilité dans son choix de formes fonctionnelles.

La littérature donne des indications claires sur la manière de modéliser les distributions marginales (log-normale, Pareto...). Quelle forme choisir pour la copule? Le graphique 3 présente une structure caractéristique : les observations tendent à se concentrer autour de la première diagonale, et le carré semble, en première approximation, symmétrique par rapport à ses deux diagonales. Ce résultat qualitatif est remarquablement constant, que l'on varie la période d'étude ou que l'on conditionne par des caractéristiques individuelles comme l'éducation. Ces régularités suggèrent que la structure de la mobilité entre t et $t+1$ peut être capturée par une spécification très simple. De fait, nous testons plusieurs familles paramétriques de copules, et trouvons qu'un paramètre est suffisant pour décrire la mobilité. La famille de copule que nous utilisons est due à Plackett (1965). Nous obtenons ainsi une modèle des trajectoires salariales à trois paramètres : deux paramètres décrivent la moyenne et la variance des salaires en coupe, et un unique paramètre décrit la mobilité relative au sein de ces échelles salariales. Nous complétons le modèle en introduisant l'état chômage.

Tenir compte de l'hétérogénéité inobservée

Le modèle de base suppose que les individus sont homogènes, dans la mesure où ils font face aux mêmes chocs de salaires et d'emploi. Cependant, pour modéliser finement les trajectoires sur le marché du travail, il est nécessaire de prendre en compte l'hétérogénéité individuelle.

Inégalité, mobilité et hétérogénéité

Une première manière d'introduire l'hétérogénéité individuelle dans le modèle est de considérer certaines caractéristiques renseignées dans les données, comme le niveau

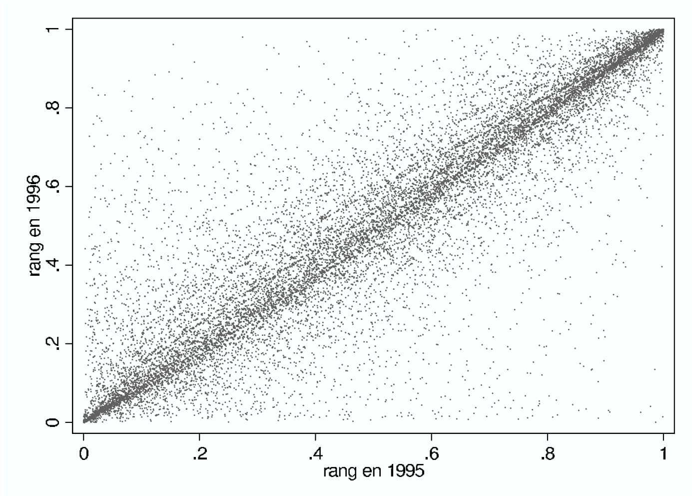


Figure 3: Mobilité relative entre 1995 et 1996, hommes en emploi, Enquête-Emploi

d'éducation et l'âge. Il est aussi important de tenir compte de différences individuelles qui ne sont pas renseignées dans les données, comme l'ont montré de nombreux travaux microéconométriques en économie du travail. Des notions comme l'"habileté" ou la "motivation", de même que des chocs permanents subits par les individus avant leur entrée sur le marché du travail, sont difficiles à quantifier, et sont absents des données auxquelles les chercheurs ont accès. Ces dimensions de l'hétérogénéité individuelle sont rassemblées sous le nom d'"hétérogénéité inobservée". L'accès à des données de panel, longitudinales, permet de la modéliser.

Le lien entre hétérogénéité inobservée et inégalités a été rarement souligné. Dans un article important, Gottschalk et Moffit (1994) tentent de mesurer les parts "permanentes" et "transitoires" des inégalités. Leur mesure de "revenu permanent" est une moyenne des salaires sur une période de dix ans. Autrement dit, la part "permanente" du salaire est la contribution d'un effet fixe individuel, c'est à dire de l'hétérogénéité individuelle. La part "transitoire" est la différence entre cet effet fixe et le salaire courant. Gottschalk et Moffit décomposent ensuite des indices usuels d'inégalité, et trouvent que deux tiers des inégalités mesurées en coupe sont permanentes, le dernier tiers étant transitoire. L'article de Gottschalk et Moffit montre qu'une modélisation fine des inégalités, qui souligne le degré de persistence des différences salariales, doit prendre en compte l'hétérogénéité inobservée.

D'autre part, la mobilité individuelle est elle aussi hétérogène. Dans le modèle à trois paramètres présenté plus haut, les trajectoires sont supposées markoviennes d'ordre 1. En d'autres termes, deux individus ayant les mêmes salaires à la période t font face à la même distribution de salaires en $t + 1$, indépendemment de leur salaire en $t - 1$. Cette remarque vaut également pour la dynamique emploi/chômage. Cependant, de nombreux travaux sur données longitudinales ont montré que cette hypothèse n'est le plus souvent pas satisfaite (par ex. Fougère et Kamionka, 1992). Une manière de modéliser cette dépendance additionnelle au niveau agrégé est de supposer la présence d'hétérogénéité au niveau individuel. Dans notre travail, nous nous inspirons du modèle dit du *Mover-Stayer* de Blumen, Kogan et McCarthy (1955), qui distingue des individus "mobiles" et des individus "stables", pour modéliser des différences dans les trajectoires relatives de salaires.

Modéliser l'hétérogénéité par des mélanges

Nous modélisons deux types d'hétérogénéité. Le premier type affecte les salaires et leur dynamique. Le second type affecte uniquement la mobilité relative. Techniquement, introduire une hétérogénéité bivariée dans le modèle peut compliquer singulièrement l'estimation des paramètres. Dans notre contexte, cependant, ce n'est pas le cas. En effet, le modèle à trois paramètres se généralise quasi trivialement à un mélange discret. Ainsi, nous distinguons plusieurs groupes d'individus, dont chacun possède son propre ensemble de paramètres : deux paramètres pour la distribution des salaire de coupe dans laquelle la trajectoire s'inscrit, et un paramètre pour la mobilité relative au sein de cette distribution particulière. L'algorithme séquentiel d'estimation que nous développons est simple à programmer et très rapide. Cette simplicité d'utilisation fournit, à notre sens, une autre motivation pour utiliser les copules afin de modéliser des trajectoires individuelles.

La France dans les années 1990-2002

Nous estimons le modèle sur les données de l'Enquête-Emploi de l'INSEE, pour les années 1990-2002.

Les données : un panel court

Pour analyser la question des inégalités longitudinales, nous avons besoin de données satisfaisant deux critères *a priori* difficilement conciliables. Premièrement, la dimension de panel est essentielle, pour prendre en compte la dynamique et l'hétérogénéité inobservée. Deuxièmement, les données doivent être représentatives de la population, si l'on veut pouvoir décrire les inégalités longitudinales pour toute une population, de la même manière que dans les études de coupe. L'Enquête-Emploi remplit dans une certaine mesure ces deux conditions. C'est un panel rotatif d'une durée de trois ans. Chaque année, un tiers de l'échantillon est remplacé. Cette caractéristique permet au panel de rester représentatif de la population, tout en suivant les individus sur une certaine durée. Le prix à payer pour la représentativité est la faible dimension longitudinale des données, trois observations annuelles permettant tout juste de prendre en compte l'hétérogénéité inobservée et la dynamique salariale.

Résultats

Dans l'estimation, nous distinguons trois groupes d'hétérogénéité du premier type (distributions de salaire) et deux groupes d'hétérogénéité du deuxième type (mobilité relative). La première hétérogénéité est fortement corrélée à l'éducation.[5] Cette hétérogénéité affecte les salaires ainsi que la probabilité de chômage. Le second type d'hétérogénéité, qui classe les individus en “mobiles” et “stables”, est en revanche peu corrélé aux autres variables de l'Enquête-Emploi, indiquant que cette hétérogénéité est essentiellement inobservée.

[5]Nous montrons que cette hétérogénéité est aussi très corrélée au croisement entre éducation et cohorte, en cohérence avec l'augmentation massive du nombre de diplômés depuis cinquante ans.

Un fort impact du chômage : Pour mesurer, en première approximation, le poids du chômage dans les inégalités, nous attribuons un revenu virtuel à chaque chômeur, égal à 60% du salaire qu'il aurait reçu s'il avait été employé (*ratio de remplacement*), ce dernier étant calculé à partir du modèle. Nous calculons ensuite le degré d'inégalité salariale de coupe, et le comparons au niveau d'inégalité prenant en compte ces "revenus virtuels" perçus au chômage. Nous trouvons que l'introduction du chômage a un effet dramatique sur les inégalités, qui augmentent de plus de 20% en niveau. De plus, la récession de 1993 a un effet plus prononcé lorsque le chômage est pris en compte, l'augmentation des inégalités étant de 20% après cette date.

Evolution de la mobilité relative : Nous calculons aussi un indice de mobilité relative, la corrélation de Spearman. Nos résultats témoignent d'une baisse de la mobilité simultanée à la hausse des inégalités de coupe, après 1993.

Persistence des inégalités : Nous mesurons ensuite l'impact de l'évolution des inégalités de coupe et de la mobilité sur les inégalités longitudinales. Nous trouvons une hausse plus prononcée qu'en coupe après 1993, 25% d'augmentation contre 10% en coupe. Après 1995, comme en coupe, les inégalités longitudinales reviennent progressivement à leur niveau de 1990. En niveau, les inégalités longitudinales représentent 80% de leur niveau de coupe. Nous décomposons aussi ces inégalités en trois composantes : "permanente" (hétérogénéité individuelle), "transitoire" (chocs sans effet durable) et "persistente" (effets durables des chocs de salaires et d'emploi). Nous trouvons que 60% des inégalités longitudinales est de nature permanente, le reste étant la résultante des chocs passés.

Ces résultats illustrent l'importance de la prise en compte des dynamiques sous-jacentes aux inégalités, et particulièrement du poids et de la persistence du chômage.

Chapitre 2 : Pertinence des caractéristiques non-salariales

Dans le deuxième chapitre, coécrit avec Grégory Jolivet, nous tentons de mesurer l'importance que les travailleurs attachent aux caractéristiques non salariales, ou aménités, de leur emploi. Notre approche exploite les changements volontaires d'un emploi à l'autre pour identifier ces préférences. Notre analyse est appliquée aux données du Panel Européen des Ménages (ECHP) pour la période 1994-2001.

Comment révéler les préférences individuelles pour les aménités?

Salaires et conditions de travail sont-ils positivement ou négativement corrélés? Interprétée à la lumière des discussions du premier chapitre sur la pertinence des indices d'inégalités usuels, cette question peut se reformuler ainsi : Les inégalités salariales sous-estiment ou surestiment-elles le niveau d'inégalités dans l'emploi (Hamermesh, 1999)?

Différences compensatrices et prix hédoniques La théorie des différences compensatrices énoncée par Adam Smith (1776) fournit un cadre pour penser cette question. Selon Smith le travail d'un mineur, pénible et salissant, doit être mieux payé toutes choses égales par ailleurs. L'argument repose sur l'offre et la demande : si le prix du travail de mineur est trop faible, à cause des mauvaises conditions de travail et d'un salaire insuffisant, alors l'offre de travail diminuera, ce qui contraindra les entreprises minières à augmenter les salaires.

Une manière de formaliser cette intuition est introduite par Rosen (1974). Dans son modèle de prix hédoniques, consommateurs et producteurs sont hétérogènes et s'apparient suivant leur degré de préférence pour les caractéristiques des biens. Appliqué au marché du travail, ce modèle a deux implications principales qui rejoignent celles de la théorie des différences compensatrices : premièrement, salaires et aménités sont négativement corrélés à l'equilibre; deuxièmement, l'intensité de la corrélation permet de mesurer les préférences individuelles pour les aménités. En conséquence, les préférences pour les caractéristiques des emplois peuvent être identifiées et quantifiées à partir de données de salaires et d'aménités.

Régressions hédoniques de salaires L'approche initiée par Rosen a motivé de nombreux travaux empiriques ayant pour but de mesurer les differences compensatrices pour certaines aménités ainsi que les Proportions Marginales à Payer des travailleurs pour ces caractéristiques. Dans le cadre de la théorie, ces deux quantités sont égales, et peuvent être mesurées par de simples régressions de salaires. A partir des travaux de Thaler et Rosen (1975), cette approche a été très souvent appliquée. Dans l'ensemble, pourtant, les résultats obtenus ne sont pas complètement convainquants. Les corrélations entre salaire et aménité sont souvent faibles, de l'ordre de 5%, et souvent du signe contraire à celui prédit par la théorie (Brown, 1980).

Une approche complémentaire : exploiter les changements d'emplois volontaires

Des problèmes de données et de méthode ont été invoqués pour expliquer ces résultats (Duncan et Holmlund, 1983). Une seconde explication, liée aux imperfections du marché du travail, a été proposée plus récemment.

Frictions sur le marché du travail : Les conclusions du modèle hédonique de Rosen dépendent de manière cruciale de l'hypothèse d'un marché du travail parfaitement concurrentiel. En présence de frictions, en effet, les préférences ne se traduisent pas nécessaire-ment en coupe. En d'autres termes, il se peut que l'on n'observe pas de différences compensatrices pour une certaine aménité, alors même que le prix que les travailleurs attachent à celle-ci est élevé. Hwang, Mortensen et Reed (1998), et Lang et Majumdar (2004) contruisent des modèles où l'on peut même observer une corrélation positive entre salaire et aménité, à l'opposé de la prédiction de la théorie.

Notre approche : Dans le chapitre 2, nous tentons de mesurer les préférences individuelles pour diverses caractéristiques non salariales en tenant compte des imperfections du marché du travail. Nous insistons en particulier sur l'importance des coûts de mobilité. En effet, si la mobilité est coûteuse il n'est plus clair que les préférences puissent être mesurées à partir d'une coupe de salaires et d'aménités. Dans ce cas, d'autres données sont nécessaires. Notre approche exploite les transitions volontaires d'emploi à emploi. L'idée est la suivante : si les travailleurs attachent une valeur élevée à leurs conditions de travail, par exemple, alors on doit observer qu'ils arbitrent entre salaire et aménité lorsqu'ils décident de changer d'emploi. Par exemple, certains emplois offrant un salaire plus faible peuvent être acceptés parce qu'ils offrent de meilleures conditions de travail.

Nous utilisons des données sur la durée des emplois d'origine, ainsi que sur les caractéristiques (salariales et non salariales) des emplois de destination. Ainsi, nous sommes en mesure de quantifier l'effet du salaire et des aménités sur la probabilité de changer d'emploi, mais aussi de mesurer les contraintes sur la mobilité et leur effet sur la corrélation entre salaires et aménités dans les nouveaux emplois. Nous pouvons ainsi réconcilier les résultats des régressions de salaires hédoniques (faibles corrélations en coupe) avec les fortes préférences estimées directement à partir de la durée des emplois (Gronberg et Reed, 1994). Dans le cadre que nous proposons, une forte Proportion Marginale à Payer pour une aménité n'est pas incompatible avec des corrélations non significatives en coupe, si les coûts de mobilité sont élevés et hétérogènes.

Un modèle dynamique des transitions

Pour quantifier ces effets, nous construisons un modèle dynamique du marché du travail. Les trajectoires de salaires et d'aménités sont modélisées, ainsi que la dynamique emploi-chômage. De plus, nous distinguons entre changements d'emplois volontaires (le travailleur dit avoir trouvé un meilleur emploi) et contraints (par exemple : licenciements).

Hétérogénéité inobservée : Notre modèle tient compte de l'hétérogénéité inobservée. La littérature insiste sur les différences de productivité entre travailleurs (Hwang, Reed et Hubbard, 1992). Nous modélisons cette hétérogénéité comme un effet fixe, constant sur la durée de l'emploi, qui intervient dans les équations de salaires et d'aménités. L'idée est que les appariements entre firmes et travailleurs les plus productifs seront caractérisés par des salaires plus élevés et de meilleures aménités. D'autre part, les données d'aménités que nous utilisons se présentent sous une forme subjective. Il s'agit, pour le travailleur, de dire s'il ou elle est satisfait(e) de son emploi dans telle ou telle dimension (conditions de travail, sécurité de l'emploi, etc...). De telles données sont vraisemblablement entachées d'un "biais de subjectivité" (Duncan et Holmlund, 1983). Par exemple, différents individus peuvent attacher des sens différents aux termes "satisfait" ou "très satisfait". Pour cette raison, nous introduisons une seconde hétérogénéité, elle aussi constante sur la durée de l'emploi, qui intervient uniquement dans les équations d'aménités.

Modéliser les changements d'emploi volontaires

Dans ce chapitre, nous identifions les préférences individuelles pour les aménités à partir des changements d'emploi volontaires. Notre modélisation suppose que les travailleurs en emploi reçoivent une offre qu'ils peuvent accepter ou refuser. Ces offres sont composées d'un salaire et de plusieurs caractéristiques non salariales. Nous permettons à ces différentes composantes d'être corrélées entre elles. Ensuite, le travailleur arbitre entre les avantages de l'offre qui lui est proposée et les caractéristiques de son emploi présent. Dans sa décision, le poids relatif qu'il attache à l'offre d'aménité vis-à-vis du salaire offert est notre mesure de sa Proportion Marginale à Payer (PMP) pour cette aménité.

Dans ce cadre, nous distinguons trois éléments cruciaux dans l'analyse. Premièrement, la PMP, qui est une mesure des préférences individuelles. Deuxièmement, la corrélation entre salaire et aménité dans les offres d'emploi, qui représente un effet "demande de travail". Par exemple, une corrélation négative peut refléter le fait que l'aménité est couteuse à produire pour l'entreprise. Le dernier élément reflète les contraintes sur la mobilité. Nous mesurons ce dernier terme comme le poids, en termes de variance, des offres de salaire dans la décision de changer d'emploi. Intuitivement, si la mobilité individuelle est coûteuse, alors ce terme est faible.

Différences compensatrices et coûts de mobilité

Le graphique 4 illustre le lien statistique entre salaire et aménité dans deux économies différentes. Les deux courbes sur chaque graphique représentent les densités de salaires associées à de "bonnes" ($a = 1$) et "mauvaises"($a = 0$) aménités, après un changement d'emploi volontaire. Dans les deux cas, les préférences individuelles pour l'aménité sont fortes. Mais les deux économies diffèrent relativement à leur mobilité. Le graphique de droite présente un cas où la mobilité emploi-emploi est peu coûteuse, où le salaire a un fort pouvoir explicatif, en termes de variance, sur la decision de changer d'emploi. Dans ce cas, la courbe correspondant à de "bonnes" aménités est clairement décalée vers la gauche par rapport à la deuxième courbe. Ceci signifie que, du fait de l'arbitrage individuel entre salaire et aménité lors du changement d'emploi, les "mauvaises" aménités sont associées à une différence compensatrice. La situation est très différente sur le graphique de gauche, où la mobilité est associée à des coûts élevés et hétérogènes. Dans ce cas, les deux courbes sont pratiquement confondues et salaire et aménité ne sont pas corrélés.

Le problème de sélection

Modéliser les changements d'emploi volontaires permet d'une part de mesurer les préférences des travailleurs pour les aménités, et d'autre part de mettre en évidence le lien entre ces préférences et les corrélations de coupe. Cependant, la nature par définition endogène de la mobilité, décidée en fonction d'offres de salaires qui ne sont observables que si elles sont acceptées, rend l'identification du modèle délicate. En d'autres termes, cette approche se heurte à un problème de sélection (Manski, 1989).

Pour résoudre ce problème, nous utilisons d'autres transitions sur le marché du

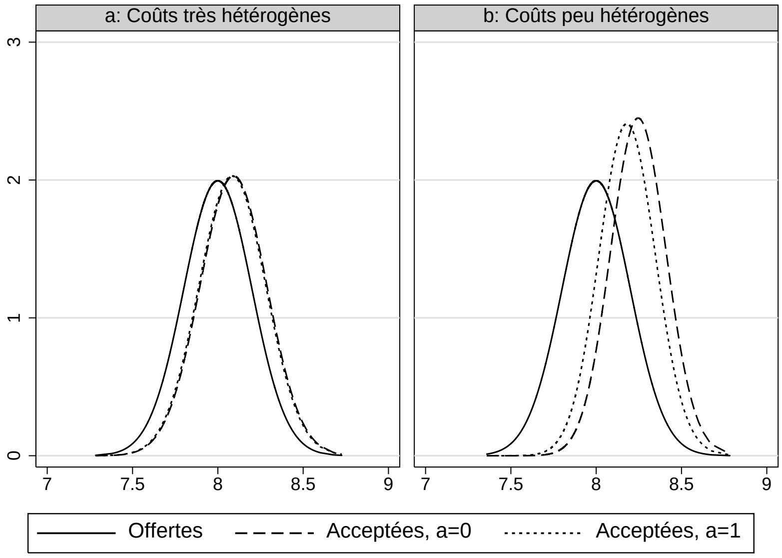


Figure 4: Différences compensatrices et couts de mobilité

travail, avec l'idée qu'elles sont informatives sur les offres reçues par les travailleurs. Idéalement, si ceux-ci étaient contraints d'accepter certaines de ces offres, alors nous pourrions utiliser ces transitions exogènes pour identifier la distribution des offres. Cette information pourrait ensuite permettre de mesurer les préférences des travailleurs et le reste des paramètres intervenant dans la decision de mobilité. Clairement, les données utilisées habituellement en microéconométrie– et l'ECHP ne fait pas exception –ne sont pas assez précises pour isoler de telles transitions.

Néanmoins, nos données nous permettent de distinguer plusieurs types de transitions associés à différents types de contraintes. Par exemple, la cause déclarée (par le travailleur) d'un changement d'emploi peut être la fin d'un contrat de travail ou un licenciement, mais aussi la naissance d'un enfant. Notre approche consiste à estimer le modèle sous l'hypothèse qu'un certain type de transitions est exogène du point du vue du travailleur, et permet d'identifer la distribution des offres d'emploi qu'il reçoit. Nous testons plusieurs types de transitions et vérifions que les résultats obtenus varient peu. Cette méthode s'apparente à l'utilisation de plusieurs groupes de contôles dans la littérature sur les effets de traitement, où varier la définition du groupe de contrôle permet de tester la robustesse des résultats (Meyer, 1995).

Evidence en Europe : similitudes et contrastes

Nous estimons le modèle sur les données de l'ECHP, pour neuf pays et la période 1994-2001. Les pays que nous incluons dans l'analyse couvrent un large spectre en termes de types de marché du travail. Les taux bruts de mobilité emploi-emploi, par exemple, varient entre 4% en France et en Italie, jusqu'à 7% en Irlande et Pays Bas et 10% au Danemark. Nous considérons cinq aménités (binaires) : "type de travail", "conditions de travail", "horaires de travail", "distance au lieu de travail" et "sécurité de l'emploi".

De fortes préférences pour les aménités...

Les statistiques descriptives montrent que les changements d'emploi volontaires sont en moyenne associés à une augmentation de salaire. Cependant, entre 30% et 40% des transitions se traduisent par des baisses de salaires. D'autre part, les transitions volontaires sont en moyenne associées à une amélioration du niveau d'aménité. Cette évidence suggère que les aménités, comme le salaire, interviennent dans la décision de changer d'emploi. Nos résultats confirment cette intuition. Plus précisément, nous estimons des Proportions Marginales à Payer positives pour la plupart des aménités. Une exception est l'aménité "horaires de travail" pour laquelle nous ne trouvons pas de préférences significatives. Les PMP les plus élévées sont obtenues pour le "type de travail" et la "sécurité de l'emploi". Au Danemark, par exemple, nous trouvons des PMP de l'ordre de 30% pour ces deux aménités. De plus, ces conclusions sont qualitativement similaires lorsqu'on varie le type de transition choisi pour résoudre le problème de sélection.

... mais des frictions importantes

Simultanément, nous estimons le pouvoir explicatif des salaires dans la décision de mobilité, que nous interprétons comme reflétant l'intensité des contraintes sur la mobilité individuelle. Nous obtenons des résultats très faibles, de l'ordre de 5% ou moins dans tous les pays (12% en Autriche). Ces résultats indiquent la présence de fortes contraintes sur les décisions de mobilité. Notre analyse implique qu'en ce cas les préférences individuelles pour les aménités peuvent être invisibles en coupe. En effet, nous estimons simultanément des corrélations très faibles entre salaire et aménités pour les individus ayant changé d'emploi volontairement. Malgré de fortes préférences pour certaines aménités, qui induisent des arbitrages entre salaire et aménités lors des changements d'emploi, nos résultats indiquent une quasi absence de différences compensatrices en coupe.

Pays latins *versus* pays nordiques?

Trois conclusions émergent : nous constatons de fortes préférences pour au moins deux des aménités que l'on considère (type de travail et sécurité), de fortes contraintes sur la mobilité emploi-emploi, et de faibles différences compensatrices. Ces conclusions ne sont pas sensibles aux tests de robustesse que nous avons effectués. De plus, elles sont remarquablement constantes entre les pays. En affinant l'analyse, néanmoins, nous distinguons clairement deux groupes de pays : les pays "nordiques" (Danemark, Pays Bas, Finlande, Autriche, Irlande) et "latins" (France, Italie, Espagne, Portugal). Ces derniers se caractérisent par des frictions de mobilité plus importantes, ainsi que par des préférences pour les aménités très élevées dans certains cas (PMP supérieures à 50%) et mal estimées.

Nous trouvons deux raisons à ces différences. La première est liée à la méthode que nous utilisons pour traiter le problème de sélection. Il apparait que, dans les pays "latins", les transitions contraintes décidées par l'employeur (tels les licenciements) sont de pauvres groupes de contrôle. Lorsque nous excluons ces transitions des transitions contraintes de référence, nous obtenons des résultats beacoup plus comparables entre pays “latins” et “nordiques”. La seconde raison renvoie aux différences de sexe : estimer le modèle sur les sous-échantillons d’hommes et de femmes ne produit pas de différences notables dans les pays “nordiques”. En revanche, l’effet sur les résultats dans les pays “latins” est fort, et tend une fois encore à rapprocher les estimations des PMP entre les deux groupes de pays.

Chapitres 3 et 4 : L'analyse factorielle en composantes indépendantes, une approche méthodologique

Les troisième et quatrième chapitres, tous deux coécrits avec Jean-Marc Robin, présentent une contribution de type méthodologique. Nous étudions l'identification des modèles linéaires à facteurs indépendants, et développons des méthodes simples et générales d'estimation. Nous appliquons notre méthode à un modèle simple de rendements de l'éducation.

Un outil pour décrire et interpréter les données

Les chercheurs en sciences humaines et sociales ont très souvent recours à des modèles à facteurs linéaires pour décrire et interpréter les données. Les premiers modèles de ce type sont apparus avec Spearman au début du siècle. Spearman (1904) suppose que l'intelligence humaine est gouvernée par une composante, qu'il appelle le facteur “g”. Cette composante étant par nature inobservable, il construit un modèle simple où le facteur “g” intervient dans plusieurs tests de mesure de l'intelligence (Quotient Intellectuel, tests verbaux...). Disposer de mesures répétées dont chacune est informative relativement au facteur sous-jacent permet de mesurer celui-ci. Cette approche a ensuite été généralisée à des modèles multi-facteurs, à partir des travaux de Thurstone arguant de la nature multidimensionnelle de l'intelligence.

L'utilité principale de l'analyse factorielle est sa capacité à résumer l'information contenue dans les données, et à en aider l'interprétation. Une méthode souvent utilisée est l'Analyse en Composantes Principales (ACP) où la matrice de variance-covariance des données est décrite par un faible nombre de facteurs. *Ex-post*, on peut reconstruire ces facteurs inobservés et les corréler avec d'autres variables, ce qui permet de les interpréter. Cette approche est très largement utilisée en finance dans le cadre des modèles fondés sur la théorie des prix arbitrés (Arbitrage Pricing Theory ou APT) de Ross (1976) pour modéliser l'évolution de cours de titres. Elle est aussi très présente dans l'étude de la structure à terme des taux d'intérêt (par exemple Piazzesi, 2003). En macroéconomie, l'analyse factorielle est utilisée pour prédire l'évolution de l'activité économique (Stock et Watson, 2002). Un autre domaine d'application est celui des modèles VAR structurels (par ex. Blanchard et Quah, 1989).

En microéconométrie, les modèles linéaires à erreur de mesure sont des exemples de modèles à facteur unique. Les modèles à facteurs sont aussi souvent utilisés pour modéliser l'hétérogénéité inobservée. Un exemple est fourni par les modèles à effets aléatoires en données de panel. Dans une récente application, Carneiro, Hansen et Heckman (2003) estiment un modèle de choix d'éducation dans lequel l'hétérogénéité inobservée joue un rôle central, puisqu'elle permet de justifier l'hypothèse classique d'indépendance entre revenus potentiels et traitement. Ces auteurs utilisent une structure factorielle pour modéliser cette hétérogénéité.

Limites de l'analyse en composantes principales

Dans ces nombreux exemples, l'identification et l'estimation des paramètres du modèles, ou facteurs de pondérations (*factor loadings* en anglais), repose sur la matrice de variance-covariance des données. Comme l'ont montré Anderson et Rubin (1956), cette caractéristique a deux conséquences sur l'identification de ces modèles.

L'ambiguité liée au choix d'une rotation : Premièrement, les facteurs de pondération sont en général non identifiés de manière unique. Considérons à titre d'illustration un modèle du type :

$$Y = \Lambda X,$$

où il y a L facteurs $Y_1, ..., Y_L$, K facteurs $X_1, ..., X_K$ et une matrice Λ de $L \times K$ facteurs de pondération. La matrice de variance covariance des données s'écrit :

$$\mathrm{Var}\,(Y) = \Lambda\,\mathrm{Var}\,(X)\,\Lambda^T = \Lambda\Lambda^T,$$

où nous adoptons la notation $()^T$ pour indiquer la transposée d'une matrice, et où les facteurs sont supposés non corrélés et de variance 1. Il est clair que la matrice Λ des facteurs de pondération n'est en général pas définie de manière unique, puisqu'on peut la multiplier par une matrice orthogonale sans changer l'expression de la matrice de variance-covariance des données.

Ainsi, en général il est nécessaire de fixer une telle matrice orthogonale, appelée rotation dans la littérature, pour définir les coefficients de manière non ambiguë. Par exemple, on peut prendre Λ égale à la décomposition de Cholesky de la matrice $\mathrm{Var}(Y)$, ce qui revient à supposer Λ triangulaire. Cette convention est souvent adoptée dans la littérature sur les modèles VAR structurels. En statistique, d'autres choix de matrices de rotation ont été proposés (Darton, 1980). Pourtant, la nature arbitraire du choix de la rotation peut sembler insatisfaisante, sachant que l'interprétation des facteurs dépend dans une très large mesure de ce choix.

Le nombre de facteurs : Une seconde limite de l'analyse factorielle traditionnelle est liée à la présence de facteurs spécifiques, ou erreurs, dans le modèle. Ainsi, si on suppose

$$Y = \Lambda X + U, \tag{1}$$

où à chaque Y_ℓ correspond une erreur U_ℓ, et les erreurs sont non corrélées entre elles et non corrélées aux facteurs, alors un simple calcul montre que le nombre de facteurs X_k identifiables est sévèrement limité. Par exemple, il faut au moins trois mesures (observées) pour identifier un unique facteur dans ce modèle.

Une approche alternative : l'analyse en composantes indépendantes

Dans les chapitres 3 et 4, nous considérons des modèles à facteurs linéaires du type (1) où les facteurs et les erreurs sont non seulement non corrélés, mais statistiquement indépendants. Nous montrons que l'hypothèse d'indépendance permet, sous des hypothèse générales, d'obtenir l'identification non ambiguë de la matrice Λ des facteurs de pondération. D'autre part, elle nous permet aussi d'identifier les densités de distribution des facteurs et erreurs. Tous les éléments du modèle sont alors explicités.

Des antécédents en économétrie : Un cas particulier du modèle que nous considérons est le sujet de deux articles historiques, de Geary (1942) et Reiersol (1950). Ces deux auteurs se placent dans le cadre d'un modèle à erreur de mesure indépendant, qui s'écrit comme un modèle à facteur unique. Reiersol prouve que deux mesures sont en générale suffisantes pour identifier les facteurs de pondération, et permettre de corriger de l'erreur de mesure sans utiliser d'instrument auxiliaire. L'hypothèse fondamentale, outre l'indépendance, est la non-normalité du facteur. Il est à noter qu'aucune autre hypothèse n'est faite sur la forme fonctionnelle des densités de distribution des facteurs et erreurs. Dans ce sens, les résultats de Reiersol, ainsi que ceux que nous présentons dans le chapitre 3, sont semi-paramétriques. L'intuition derrière l'hypothèse de non normalité est la suivante : les restrictions sur les covariances impliquées par le modèle ne sont pas suffisantes pour obtenir l'identification. Nous sommes alors dans le cas de l'analyse factorielle traditionnelle, où une troisième mesure est nécessaire pour identifier les paramètres d'intérêt. De plus, dans le cas où le facteur est normal, les restrictions sur les moments d'ordre élevé ne sont pas informatives. D'où la non identification dans ce cas. Si le facteur n'est pas normal, en revanche, alors les moments d'ordre supérieur à trois fournissent une information supplémentaire qui permet l'identification. Geary (1942) construit plusieurs estimateurs explicites des facteurs de pondération qui s'écrivent comme une combinaison de moments d'ordre 2, 3 et 4 des données.

Une littérature récente : L'idée originale de Geary et Reiersol a été utilisée à plusieurs reprise en économétrie, dans le cas de modèles à erreur de mesure. Dans le chapitre 3, nous généralisons cette idée au cas de modèles multi-facteurs. Notre approche est reliée à une littérature en rapide expansion dans le domaine de l'extraction des signaux, l'Analyse en Composantes Indépendantes (ACI). Formellement, le type de modèle considéré dans cette littérature est le même que celui que nous considérons ici. De nombreuses méthodes ont été développées dans le cadre de l'ACI pour estimer les coefficients de la matrice Λ dans un modèle sans erreurs. Une telle hypothèse représente une limite sérieuse en sciences sociales où il est important de tenir compte des erreurs de mesures ainsi que des facteurs spécifiques à chaque mesure. Dans le chapitre 3, nous généralisons une méthode très utilisée en ACI au cas où les erreurs ne sont pas négligeables. L'algorithme d'estimation de la matrice Λ que nous introduisons, que nous appelons quasi-JADE, est construit comme une généralisation de l'algorithme JADE de Cardoso et Souloumiac (1993). Il est à noter que les résultats que nous démontrons sont nouveaux, y compris relativement à la littérature consacrée à l'ACI.

Quelques travaux sur les distributions : Dans le chapitre 4, nous considérons le problème de l'identification et l'estimation non-paramétriques des densités de distribution des facteurs et erreurs. Ici encore, le cas d'un unique facteur a été traité dans la littérature comme une généralisation relativement directe des méthodes de déconvolution de Carroll et Hall (1988) et Stefanski et Carroll (1990) par exemple. Ainsi, Li et Vuong (1998) s'inspirent d'un résultat de Kotlarski (1967) pour construire des estimateurs convergents des densités de X, U_1 et U_2 dans le modèle

$$\begin{aligned} Y_1 &= X + U_1, \\ Y_2 &= X + U_2. \end{aligned}$$

Nous contribuons à cette littérature en étendant l'estimateur de Li et Vuong au cas de modèles multi-facteurs généraux. A notre connaissance, la littérature consacrée à l'ACI n'a pas jusqu'ici proposé de méthode non-paramétrique d'estimation des densités.

Méthodes pour l'identification et estimation

Les méthodes que nous mettons en oeuvre pour d'une part démontrer l'identification des paramètres et des densités, et d'autre part en construire des estimateurs, font intervenir des techniques assez complexes. Je donne ici juste les idées principales, et renvoie le lecteur intéressé aux textes eux-mêmes.

Les facteurs de pondération

Dans le troisième chapitre, nous développons des méthodes algébriques pour estimer la matrice Λ, fondées sur les moments d'ordre 2, 3 et 4 des données, comme dans l'article original de Geary (1942). Nous montrons qu'en utilisant les nombreuses restrictions suridentifiantes impliquées par l'hypothèse d'indépendance statistique, l'imprécision bien connue de l'estimation de ces moments en échantillon fini diminue fortement. Nous obtenons un algorithme aux performances encourageantes sur de grands échantillons (plus de 1000 observations).

Moments d'ordre élevés et décompositions algébriques Notre approche repose sur une réécriture des restrictions de moments sous forme matricielle. Précisément, nous considérons certaines matrices de moments des données, que l'on peut estimer directement par moyennes empiriques, dont la structure factorielle implique qu'elles s'écrivent comme la somme de deux composantes.

• La première composante fait intervenir les moments (d'ordre 2 à 4) des erreurs. Nous montrons qu'elle satisfait une restriction linéaire en fonction des moments des données. Ainsi, elle peut être estimée par moindres carrés ordinaires. La première étape de l'algorithme quasi-JADE que nous proposons consiste à estimer les premiers moments des erreurs.

• Une fois la première composante estimée, la seconde se calcule directement par différence avec la matrice de moments des données. La structure factorielle implique alors que la matrice Λ, qui intervient dans la seconde composante, est solution d'un problème de diagonalisation jointe. Nous utilisons l'algorithme proposé par Cardoso et Souloumiac (1993) pour effecuer cette diagonalisation dans la deuxième étape de notre algorithme et estimer Λ.

Il est à noter que nous estimons la matrice de facteurs de pondération sous l'hypothèse que le nombre de facteurs est connu. Nous montrons aussi comment estimer ce dernier, en testant le rang de certaines matrices de moments des données.

Propriétés asymptotiques et simulations de Monte-Carlo : Notre algorithme utilise la linéarité et l'indépendance du modèle à facteurs pour obtenir une solution simple à un problème extrêmement non-linéaire. D'autre part, il est à noter que notre approche utilise tous les moments d'ordre 2,3 et 4 des données. Techniquement, les restrictions suridentifiantes se traduisent en un grand nombre de matrices à diagonaliser simultanément. Nous montrons que cette propriété tend à diminuer la variance asymptotique, et en particulier la probabilité que la solution soit proche d'un cas de racines multiples.

Nous étudions en détail les propriétés de notre algorithme à distance finie. Nous obtenons des résultats remarquablement précis, comparés à l'imprécision des moments emipiques des moments d'ordre 4 sur lesquels sont fondés les calculs. Nous trouvons aussi que les performances de l'algorithme sont très sensibles à la non-normalité des données : si les facteurs ont des moments d'ordre 3 et 4 proches de ceux de la distribution normale, alors les estimations deviennent très bruitées.

Les densités

Dans le quatrième chapitre, nous supposons la matrice Λ connue. Nous construisons un estimateur des densités fondé sur des déconvolutions et intégrations. Nous étudions certaines de ses propriétés asymptotiques.

Généraliser les méthodes de déconvolution : L'idée centrale est d'utiliser les restrictions que la structure factorielle implique sur la fonction caractéristique (FC) des données. Précisément, nous considérons la fonction génératrice des cumulants (FGC), qui est égale au logarithme de la FC. La FGC des données s'écrit comme une fonction linéaire des FGC des facteurs et erreurs évaluées en certains points. Nous différentions deux fois ces restrictions et obtenons un système matriciel que l'on peut en général inverser, si le nombre de facteurs n'est pas trop grand. Nous estimons ensuite les FC des facteurs et erreurs par double intégration, et leur densité par déconvolution.

Propriétés de l'estimateur : La dernière étape (déconvolution) requiert le choix d'un paramètre de troncature, car l'intégrale ne converge pas en général. Pour cette raison, ce paramètre –qui doit tendre vers l'infini avec la taille de l'échantillon – ne doit pas tendre trop vite. Cette caractéristique de l'estimateur produit un taux de convergence qui peut être extrêmement lent dans certains cas (Carroll et Hall, 1988, et Fan, 1991). Nous démontrons la convergence de notre estimateur sous des hypothèses générales, en particulier concernant le support des distributions. Nous dérivons le taux de convergence dans certains cas, et trouvons effectivement qu'il peut être logarithmique (Horowitz et Markatou, 1996).

Nos simulations de Monte Carlo confirment qu'en général la convergence est lente. D'autre part, nous trouvons que la forme de la densité à estimer a un fort impact sur les performances de l'estimateur, au moins dans les exemples que nous considérons. En particulier, des densités très asymétriques ou très kurtiques semblent particulièrement difficiles à estimer. Il est intéressant de considérer ce résultat à la lumière des analyses du chapitre trois, où nous montrons que des moments d'ordre 3 et 4 élevés sont cruciaux pour bien estimer la matrice Λ.

Une application aux rendements de l'éducation

Les dernières sections des chapitres 3 et 4 présentent une application des estimateurs à un exemple simple de rendements de l'éducation. Nous utilisons l'Enquête-Emploi pour l'année 1995. Les salaires horaires sont notés Y, et le nombre d'année d'études est D. Les travaux cherchant à mesurer l'effet (causal) de D sur Y sont innombrables dans la littérature (Griliches, 1977 et Card, 2001, sont deux revues classiques).

Une seconde mesure de l'éducation comme instrument : Nous construisons une seconde mesure de niveau d'éducation, que nous notons D^*. Pour cela, nous considérons une variable qui indique qualitativement le plus haut niveau obtenu (par exemple : baccalauréat), et nous rendons cette variable continue en faisant correspondre à chaque catégorie le nombre d'années d'études médian qui a été nécessaire pour l'atteindre. Ainsi, la différence entre cette variable continue, D^*, et le nombre d'années d'études D reflète en particulier les redoublements, mais aussi les changements d'orientation, etc...

Régresser directement le salaire horaire Y sur le nombre d'années d'études D par moindres carrés donne un rendement de 4.4%. Si le régresseur D est mesuré avec erreur, il est connu que ce coefficient ne reflète pas le "vrai" rendement des études. Disposer d'une seconde mesure D^* permet d'estimer ce dernier par variables instrumentales. Nous obtenons un coefficient de 6.6% par cette méthode. Ce résultat indique que les moindres carrés sous-estiment le rendement de l'éducation. De nombreuses études, utilisant d'autres types d'instruments comme des variations dans l'offre d'éducation, obtiennent une conclusion similaire.

Un deuxième facteur : Nous interprétons le modèle à erreur de mesure comme un modèle à facteur unique. En imposant l'indépendance des erreurs et du facteur, nous pouvons appliquer les méthodes développées dans les chapitres 3 et 4 pour estimer ce modèle. Nous trouvons des résultats très semblables à ceux obtenus par variables instrumentales. Nous testons ensuite la présence d'un second facteur. L'analyse factorielle classique peut au plus identifier un facteur dans la relation entre le salaire horaire Y et les deux mesures d'éducation D et D^*. nous montrons que sous l'hypothèse d'indépendance, un second facteur est identifié. Nous estimons ce dernier par quasi-JADE, et trouvons que le second facteur est positivement corrélé au nombre d'années d'études et négativement à l'éducation. Ce résultat suggère qu'un certain type d'investissement en éducation est associé à un rendement négatif sur le marché du travail. Nous utilisons ensuite les méthodes du chapitre 4 pour estimer les densités des deux facteurs, que nous trouvons significativement non-normales. Ces estimations nous permettent de prédire la valeur des deux facteurs pour chaque individu dans l'échantillon. Nous trouvons que le facteur positivement relié au salaire, de loin le facteur dominant en terme de variance, est aussi positivement corrélé à plusieurs autres variables comme le fait de travailler dans le secteur public, la catégorie socio-professionnelle (CSP) de l'individu ainsi que la CSP du père. Le second facteur, en revanche, est négativement corrélé aux deux premières de ces variables.

Contents

Chapter 1

Assessing Lifetime Earnings Inequality Using Short Panels with an Application to France, 1990-2002

1.1 Introduction

In this chapter, we propose a methodology for decomposing life-time earnings inequality into permanent, persistent and transitory components, which can be applied when individual earnings are recorded over short periods of time. Short panels severely limit the possibilities for identifying the stochastic processes governing individual earnings mobility together with the vast amount of individual heterogeneity that mixes the earnings processes. However, if short panels render the identification of models with many parameters difficult, compared to long panel surveys as the PSID, which are anyway rare, they offer large sample sizes and limited attrition. It is therefore useful to develop methodologies that can be applied to rotating panels such as the French Labour Force Survey that we use in this chapter. This is the very first aim of this chapter.

To complete this goal, we exploit advances in the statistical literature on the parametric specification of copulas. If X_t and X_{t+1} are two consecutive earnings with marginal distributions F_t and F_{t+1}, the copula function (respectively, copula density) of (X_t, X_{t+1}) is the cdf (resp. pdf) of $(F_t(X_t), F_{t+1}(X_{t+1}))$, that is the distribution of the ranks of X_t and X_{t+1} in their respective marginal distributions. The joint density of (X_t, X_{t+1}) is then the product of the marginal densities of X_t and X_{t+1} and the copula density. The applied statistical literature offers many parsimonious parametric specifications; some of them, we shall see, fit the data exceptionally well.

This way of modelling earnings processes differs from the more familiar linear decomposition of individual log wages into a fixed effect, a deterministic trend, a random walk and a moving average.[1] This is hardly feasible if only three or four years of individual-temporal data are available, as one must first difference out the fixed effect

[1]See for example the survey of Alvarez, Browning and Ejrnaes (2001).

and use IV for estimation with lagged values of right-hand side variables as instruments. Moreover, if linear models admittedly provide a good description of earnings dynamics on average, there is currently no evidence that the autoregressive correlations are well rendered throughout the entire distribution. Meghir and Pistaferri (2004) is one of the very few attempts at improving the description of higher-order earnings dynamics in such models and they provide definite evidence of autoregressive heteroskedasticity.

The need for better statistical measures of association between two random variables (usually different stock returns) and the fact that stock returns usually exhibit second order dynamics precisely explain why copulas are so popular in financial econometrics. But to our knowledge, copulas have not been used in labour economics so far. One objective of this chapter is to investigate the use of copulas to model earnings dynamics.

Once the problem of fitting earnings mobility has been reduced to reasonable dimension, it becomes possible to address the question of how much mixing—that is how much heterogeneity, observed and unobserved—there is in the data. Since the seminal work of Gottshalk and Moffitt (1994), we know that earnings inequality reflects more than persistent shocks. Genuine non ergodicity calls for additional mixing. Moreover, there is clear evidence that a large share of individual heterogeneity in earnings processes is driven by unobserved individual characteristics. Yet, most of the papers on relative earnings mobility base inference on transition probability matrices and rule out unobserved heterogeneity. Furthermore, if one believes that the study of earnings inequality can be pursued independently of the study of relative earnings mobility, this means that one is willing to assume that the unobserved heterogeneity that is driving cross-sectional distributions is independent of the heterogeneity in the dynamics. This is obviously questionable. Constructing and estimating a model of earnings dynamics with lots of unobserved heterogeneity—i.e. the individual heterogeneity that remains once education and experience and gender and cohort have been accounted for—constitutes the second aim of this chapter.

Using the copula technology to reduce dimensionality, the model can be nonparametric as far as unobserved heterogeneity is concerned. We therefore use a discrete mixture approach as in Heckman and Singer (1984). The population is composed of several groups, each group facing a specific earnings process. By proceeding this way, we do not a priori specify which parameter is heterogenous.

We apply our methodology to the French Labour Force Survey data between 1990 and 2002.[2] A third of the surveyed individuals being replaced every year, we can construct 10 samples of three-year individual employment/earnings trajectories. This allows us to study how earnings inequality and earnings mobility have changed over the last decade of the twentieth century in France. We find rather small changes in earnings inequality over that period. Earnings inequality, unemployment risk and immobility seem to increase in business cycle busts and to decrease in booms. Altogether, these three risks are driving a 25% rise in the variance of lifecycle log-earnings between 1990 (a boom) and 1995 (a bust). This is significantly more than the 10% rise in the cross-sectional earnings variance that occurs between 1990 and 1995. Importantly, we show how to decompose earnings inequality into separate permanent, transitory and persistent components in this setup. We find that the variance of intertemporal log-earnings remains a very significant share of the variance of one-year log-earnings, whatever the horizon over which they are computed (80% in the limit).

[2]Before 1990, earnings were not precisely recorded. Moreover, we use the Labor Force Survey instead of administrative earnings data because we need to precisely monitor the sequence of employment and unemployment spells to analyse earnings inequality over the lifecycle.

Moreover, about 60% of the variance of intertemporal log-earnings is due to individual heterogeneity (permanent inequality) and 33% is persistent, leaving a mere 7% to residual transitory variance.

The remainder of the chapter is organized as follows. Section 1.2 is devoted to the construction of the model. We first introduce the concept of copula. We then review several popular parametric families of copulas and test their ability to fit earnings data. Section 1.3 develops a particular specification of the model and presents estimates using data drawn from the French Labour Force Survey, 1990-2002. Section 1.4 analyzes the time changes and the anatomy of earnings inequality and relative mobility in France over the last decade of the twentieth century. Section 1.5 simulates individual trajectories and studies the inequality of individual sums of log-earnings flows. The last section concludes.

1.2 Copula models for earnings dynamics

The studies of earnings inequality document changes in the sequence of marginal earnings distributions and the studies of earnings mobility usually focus on the dynamics of individual *relative positions* in this sequence of marginal distributions. Although it is never referred to under this name, the statistical notion of copula is thus at the very heart of the literature on earnings inequality and mobility.

1.2.1 Copulas

Let X and Y be two random variables with cdf's F^X and F^Y, respectively, and let F be the cdf of (X, Y), the copula of (X, Y) is the cdf (denote it as C) of $\left(F^X(X), F^Y(Y)\right)$:

$$F(x,y) = C\left[F^X(x), F^Y(y)\right], \quad \forall (x,y). \tag{1.1}$$

The copula C is unique if the marginal cdf's F^X, F^Y are continuous. One can also define the copula density of (X, Y) as $c(u, v) \equiv \frac{\partial^2 C(u,v)}{\partial u \partial v}$. Because $F^X(X)$ and $F^Y(Y)$ have uniform distributions, $c(u, v)$ is the density of $(F^X(X), F^Y(Y))$ at (u, v) and it is also the conditional density of $F^Y(Y)$ at point v given $F^X(X) = u$.

Differentiation of equation (1.1) splits the joint density of X and Y, $f(x, y)$, into the product of marginal densities $f^X(x)$ and $f^Y(y)$ and of the copula density, $c(u, v)$:

$$f(x, y) = f^X(x) f^Y(y) c \left[F^X(x), F^Y(y) \right]. \tag{1.2}$$

Let (Y_t, Y_{t+1}) be a couple of subsequent earnings. The above theory implies that the joint density of (Y_t, Y_{t+1}) can be decomposed into the product of marginal densities times the copula density. Compared to standard earnings dynamics models (e.g. ARMA models), the copula approach does not impose parametric restrictions on cross-sectional variances and covariances across time, as marginal distributions and copula densities are parametrically independent. Whether or not this characteristic is supported by economic theory requires additional research. Yet, one can think of a market equilibrium which time path is determined by exogenous aggregate shocks. Individuals are heterogenous and the distribution of heterogeneity is time-invariant. It is then likely that individuals' positions will not change if the cross-sectional distribution is modified by an aggregate shock. Moreover, any dynamics of individual positions within the equilibrium distributions, which may reflect individual-specific shocks or other market imperfections at the source of random individual instability (search-matching frictions, asymmetric information, etc.), could be exogenous to the process of macroeconomic shocks on the steady-state equilibrium distributions.

Lastly, most studies of relative earnings mobility start by categorizing earnings into quantile intervals and then compute the probabilities of transitions between quantile intervals. These transition probability matrices are nonparametric approximations of the underlying true copula. It is of course possible to obtain smooth nonparametric estimates. Unfortunately, the nonparametric estimation of bivariate density functions requires large amounts of data which may be available for applications in finance but are usually not available in labour economics. Furthermore, the curse of dimensionality makes it almost infeasible to incorporate observed—not to mention unobserved—heterogeneity in practice.

1.2.2 Parametric copulas

The statistical literature offers a large choice of parametric specifications of copulas. Does there exist one that fits the dynamics of earnings ranks well? We now investigate this question using earnings data drawn from the French Labour Force Surveys of 1990 and 1991. Appendix A.1 details the copula families that we consider.

To remain as much agnostic as possible as far as the form of the marginal distributions is concerned, we first regress log-earnings on education and experience separately for 1990 and 1991 and compute the empirical cumulative distribution functions of the residuals. Then, we estimate the different copula parameters by maximizing the likelihood of the empirical ranks. Finally, we categorize the residuals by quintiles and compare the empirical transition probability matrix to the predictions from the different copula estimates.

Table 1.1 shows the actual and predicted transition probability matrices for the Gaussian, Plackett, Frank, Gumbel, Joe, Clayton, FGM and Log copulas. Several specifications are able to capture the general form of the observed transition matrix. In particular, there should be more inertia at the top and bottom of the distributions of ranks than in the middle. Plackett's (1965) one-parameter family of copulas fits transition probabilities better than all other families.[3] The Euclidian distance d

[3]We refer to Appendix A.2 for a presentation of the Plackett copula and its properties.

between predicted and actual matrices is .08 for this copula, while the Gaussian has: $d = .30$. We checked the robustness of this result by varying samples and controls. Plackett's copula always provided the best fit.

$$\underset{\text{Observed}}{\begin{pmatrix} .68 & .21 & .08 & .03 & .00 \\ .20 & .50 & .22 & .06 & .02 \\ .07 & .21 & .47 & .20 & .05 \\ .03 & .06 & .19 & .53 & .19 \\ .02 & .02 & .04 & .18 & .74 \end{pmatrix}} \quad \underset{\text{Gaussian} \quad (d = .30)}{\begin{pmatrix} .66 & .25 & .08 & .01 & .00 \\ .25 & .37 & .26 & .11 & .01 \\ .08 & .26 & .32 & .26 & .08 \\ .01 & .11 & .26 & .37 & .25 \\ .00 & .01 & .08 & .25 & .66 \end{pmatrix}} \quad \underset{\text{Plackett} \quad (d = .08)}{\begin{pmatrix} .73 & .20 & .04 & .02 & .01 \\ .20 & .52 & .21 & .05 & .02 \\ .04 & .21 & .50 & .21 & .04 \\ .02 & .05 & .21 & .52 & .20 \\ .01 & .02 & .04 & .20 & .73 \end{pmatrix}}$$

$$\underset{\text{Frank} \quad (d = .19)}{\begin{pmatrix} .68 & .26 & .05 & .01 & .00 \\ .26 & .45 & .23 & .05 & .01 \\ .05 & .23 & .44 & .23 & .05 \\ .01 & .05 & .23 & .45 & .26 \\ .00 & .01 & .05 & .26 & .68 \end{pmatrix}} \quad \underset{\text{Gumbel} \quad (d = .22)}{\begin{pmatrix} .63 & .26 & .09 & .02 & .00 \\ .26 & .39 & .25 & .09 & .01 \\ .09 & .25 & .37 & .25 & .04 \\ .02 & .09 & .25 & .45 & .19 \\ .00 & .01 & .04 & .19 & .76 \end{pmatrix}} \quad \underset{\text{Joe} \quad (d = .41)}{\begin{pmatrix} .46 & .31 & .17 & .06 & .00 \\ .31 & .34 & .24 & .10 & .01 \\ .17 & .24 & .32 & .23 & .04 \\ .06 & .10 & .23 & .43 & .18 \\ .00 & .01 & .04 & .18 & .77 \end{pmatrix}}$$

$$\underset{\text{Clayton} \quad (d = .50)}{\begin{pmatrix} .73 & .19 & .05 & .02 & .01 \\ .19 & .39 & .24 & .12 & .06 \\ .05 & .24 & .30 & .24 & .17 \\ .02 & .12 & .24 & .31 & .31 \\ .01 & .06 & .17 & .31 & .45 \end{pmatrix}} \quad \underset{\text{FGM} \quad (d = .83)}{\begin{pmatrix} .33 & .26 & .20 & .14 & .07 \\ .26 & .23 & .20 & .17 & .14 \\ .20 & .20 & .20 & .20 & .20 \\ .14 & .17 & .20 & .23 & .26 \\ .07 & .14 & .20 & .26 & .33 \end{pmatrix}} \quad \underset{\text{Log-Copula} \quad (d = .17)}{\begin{pmatrix} .66 & .25 & .07 & .02 & .00 \\ .25 & .42 & .25 & .07 & .01 \\ .07 & .25 & .40 & .24 & .04 \\ .02 & .07 & .24 & .47 & .20 \\ .00 & .01 & .04 & .20 & .75 \end{pmatrix}}$$

Table 1.1: Fit of some parametric families of copulas to earnings data

In particular, the Gaussian copula, which is implied by an AR(1) model for log-earnings levels if their marginal distribution is normal,[4] does not provide as good a description of autocorrelations throughout the entire distribution as the Plackett copula. One will argue that this does not prove that the familiar linear time-series

[4] Let $r_t = F_t(y_t)$ be the rank of y_t in the marginal earnings distribution at time t. The Gaussian copula postulates the following relation between r_{t+1} and r_t:

$$\Phi^{-1}(r_{t+1}) = \tau \Phi^{-1}(r_t) + \sqrt{1 - \tau^2} \varepsilon_t,$$

with ε_t white noise. If $\ln y_t$ is normal $\mathcal{N}\left(0, \sigma_t^2\right)$, then $\Phi^{-1}(r_t) = \frac{1}{\sigma_t} \ln y_t$, but this is not true otherwise.

models of log-earnings are invalidated, in particular because innovations usually exhibit moving-average components, which are here ruled out by the Markov assumption. More work is needed to determine which approach provides the best trade-off between good fit and parametric parsimony. The aim of the current research is not to make the case that copulas are better than linear time-series models of earnings levels, but to take advantage of the capacity of simple parametric versions of copulas to fit the dynamics of ranks.

1.2.3 Heterogeneity

The first-order Markov assumption is rejected by earnings data. This assumption implies that the transition probability matrix of a given discrete mobility process between two non consecutive dates t and $t+2$ is the product of the $(t, t+1)$ and $(t+1, t+2)$ transition matrices. In practice, the product matrix thus obtained usually presents more mobility than the true transition matrix between t and $t+2$.[5]

We give an illustration using the French Labour Force Survey data of 1990-1992. The product matrix obtained from the observed mobility between 1990 and 1991 and the actual transition matrix between 1990 and 1992 are

$$P_{01}^2 = \begin{pmatrix} .57 & .26 & .11 & .05 & .01 \\ .27 & .38 & .22 & .10 & .03 \\ .11 & .24 & .37 & .22 & .06 \\ .04 & .10 & .24 & .42 & .20 \\ .01 & .02 & .06 & .21 & .70 \end{pmatrix}, \qquad P_{02} = \begin{pmatrix} .67 & .22 & .08 & .02 & .01 \\ .24 & .49 & .22 & .04 & .01 \\ .06 & .24 & .50 & .18 & .02 \\ .02 & .04 & .18 & .59 & .17 \\ .01 & .01 & .02 & .17 & .79 \end{pmatrix}.$$

As the other studies, we find more mobility between quantiles in the predicted matrix than in the actual 1990-1992 matrix of transition frequencies.[6] The discrepancy

[5]See for instance Blumen *et al.* (1955), Shorrocks (1976) and Singer and Spilerman (1976). Empirical evidence shows that earnings ranks are more persistent than what they should be if they were first-order Markov processes. For example, using PSID data, Gottschalk (1997) finds that being in the first quintile of the earnings distribution in 1970, there is more than 50% chances of being in the same quintile 20 years later. Guillotin and Bigard (1992), for France, obtain similar orders of magnitude.

[6]In this chapter, "more mobility" is to be understood in terms of the concordance ordering $\precsim_c$.

persists whatever control variable we use to condition the 1990-1991 transition probability matrix. Controlling for education, experience, etc., somewhat reduces the discrepancy but not all of it.

One way of explaining state dependence, without abandoning the first-order Markov assumption, is to postulate the existence of unobserved heterogeneity as in the mover-stayer model of Blumen, Kogan and McCarthy (1955). Allowing for unobserved heterogeneity should generate enough additional autocorrelation in individual earnings trajectories to fill the remaining gap. However, allowing for unobserved heterogeneity requires a model with a simple parametric form. This is another advantage of using parametric copulas. This parametric parsimony will even let us be nonparametric in the way we model unobserved heterogeneity. We shall thus assume a *discrete* mixture model, i.e. there exist at most K types of earnings processes differing in the values of the parameters, the number of groups, K, being estimated by penalizing the estimation criterion for the number of parameters (see Heckman and Singer, 1984).

1.3 Application to French individual earnings and employment dynamics, 1990-2002

We now turn to the study of earnings (and employment) dynamics in France, using data covering the period 1990-2002. We first introduce the data and present the empirical specification of our model. We then detail the estimation results and, lastly, we describe how the model fits the data.

1.3.1 The French Labour Force Survey, 1990-2002

The data we use come from the 1990-2002 French Labour Force Survey (LFS), conducted by INSEE, the French Statistical Institute. We use the LFS data instead of

See Appendix A.2 for details.

administrative earnings data (DADS), as in Alvarez, Browning and Erjnaes (2001), because the DADS data do not follow individuals across states which are not private jobs. The French LFS is a rotating panel, a third of the sampling units (dwellings) being replaced, every year, by an equivalent number of newly sampled units. Large samples of about 150,000 individuals aged 15 or more, in 75,000 households, can thus be interviewed three times, in March of three subsequent years, about various aspects of their employment histories. Note that a panel length of three years is intuitively barely enough to ensure identification: two years are necessary to characterize the first-order Markov dynamics, one additional year is required in case of unobserved heterogeneity. Yet the short panel length is compensated by the fact that the LFS provides large, nationally representative samples for every year.

As is usually the case for this sort of survey based on individuals' responses to interviews, hours worked are badly reported. For instance, we noticed that many individuals alternatively reported 39 hours worked in one week (the legal working time in the 1990's), and 40 hours in another. To limit the influence of hour measurement errors on our results, we chose to use monthly salaries, which is what the questionnaire asks for, rather than hourly wages. Monthly earnings were finally divided by the retail price index that is provided by INSEE.

We dropped all observations for students, retired persons and self-employed, and kept only male labour trajectories, removing female trajectories from the sample to limit the role of labour supply as a determinant of earnings dispersion. We also trimmed the data below the first and above the ninety ninth percentiles of the wage distribution.

Complete trajectories (y_t, y_{t+1}, y_{t+2}), where t is the date of the first recorded observation, account for 53% of the sample. Two-year trajectories (y_t, y_{t+1}) account for

25% ;and 22% of individuals drop out after the first year. There is thus a certain amount of attrition which we assume exogenous to the employment-earnings process.

1.3.2 Model specification

In this application, we introduce unemployment as a specific state. This is important for any reasonable description of labour market trajectories in the Euro area, where unemployment rates are between 8 and 10%. We also allow for observed and unobserved heterogeneity. The parametric specification is detailed in Appendix A.4. We here only summarize its main characteristics.

Heterogeneity. After trying various specifications, we decided to introduce two latent unobserved heterogeneity variables: one variable, $z_1 \in \{1, ..., K_1\}$, conditions every parameter of the model, marginal distributions of earnings and employment states as well as copulas, and a second variable, $z_2 \in \{1, ..., K_2\}$, specifically conditions the mobility process. Individual employment-earnings trajectories are also conditioned by a set of time-varying covariates, denoted as x_t. In our application, x_t only comprises labour market experience (age minus age at the end of school) and squared experience. Lastly, covariates (z_1, z_2) are determined by some vector z_0 of non time-varying individual attributes:

- Education (Ed), classified into five categories according to the highest degree obtained: "no degree," "junior high-school," "senior high-school," "less than three years of university" and "more than three years."

- The year of entry in the labour market (b), defining a "cohort." Given calendar time t, experience is equal to $t - b$ and x_t is thus deterministic conditional on b.

- Because a *baccalauréat* (high-school diploma in France) obtained in 1950 had more market value than a *baccalauréat* obtained in 1990,[7] we also allow for education and cohort interactions ($Ed * b$).

Therefore, education conditions the earnings process only *via* the latent heterogeneity variables z_1 and z_2, which may thus be understood as two different components of the human capital an individual possesses when entering the labour market. Latent variables z_1 and z_2 are like unobserved factors which are measured by education and cohort with error.

Stochastic parameterization. Let $e_t \in \{0, 1\}$ denote the employment state at calendar time t ($e_t = 1$ if employed, $= 0$ otherwise). Let y_t denote the logarithm of employees' wages. We write the joint density of individual data $(y_t, y_{t+1}, y_{t+2}, e_t, e_{t+1}, e_{t+2}, z_1, z_2)$ (with respect to the appropriate measure) conditional on z_0 as the following product:

$$\begin{aligned} f(y_t, y_{t+1}, y_{t+2}, e_t, e_{t+1}, e_{t+2}, z_1, z_2 | x_t, x_{t+1}, x_{t+2,} z_0) &= \Pr\{z_1|z_0\} \Pr\{z_2|z_1, z_0\} \\ &\times \Pr\{e_t|x_t, z_1\} \Pr\{e_{t+1}|e_t, x_t, z_1, z_2\} \Pr\{e_{t+2}|e_{t+1}, x_{t+1}, z_1, z_2\} \\ &\times f(y_t|x_t, z_1)^{e_t} f(y_{t+1}|x_{t+1}, z_1)^{e_{t+1}} f(y_{t+2}|x_{t+2}, z_1)^{e_{t+2}} \\ &\times c\left[F(y_t|x_t, z_1), F(y_{t+1}|x_{t+1}, z_1)|x_t, z_1, z_2\right]^{e_t e_{t+1}} \\ &\times c\left[F(y_{t+1}|x_{t+1}, z_1), F(y_{t+2}|x_{t+2}, z_1)|x_{t+1}, z_1, z_2\right]^{e_{t+1} e_{t+2}}, \end{aligned} \tag{1.3}$$

where each component is specified as follows:

- **Distribution of unobserved heterogeneity.** The probability distribution of the latent variable z_1 given z_0, $\Pr\{z_1|z_0\}$, and the conditional probability of z_2 given z_1 and z_0, $\Pr\{z_2|z_1, z_0\}$, are modelled as Multinomial LOGIT.

[7]In 1980-1985 only 35% of a student cohort would attain an education level assimilable to a high school diploma. Between 1985 and 1995 a voluntarist educational policy doubled that number: since 1995 about 70% of a student cohort reach grade 12 (*terminale*)–80% of them obtaining the *baccalauréat*.

- **Cross-sectional distributions**. The cross-section log-wage density, $f(y_t|x_t, z_1)$, is assumed Gaussian conditional on experience x_t and heterogeneity z_1. We let both the cross-section mean and the cross-section variance depend on x_t as in Moffit and Gottschalk (2002). We also allow the intercept and the returns to experience in both first and second order moments to be group-specific (z_1-dependent). The unconditional unemployment probability, $\Pr\{e_t = 0|x_t, z_1\}$, follows a PROBIT model. With unobserved heterogeneity, it is indeed important to specify this probability in order to compute the likelihood component corresponding to the first observation period of each individual trajectory (Heckman, 1981).

- **Mobility.** Employment/unemployment mobility and earnings mobility are both assumed first-order Markov. We model the conditional unemployment-unemployment probability, i.e. $\Pr\{e_{t+1} = 0|e_t = 0, x_t, z_1, z_2\}$, and the conditional employment-unemployment probability, i.e. $\Pr\{e_{t+1} = 0|e_t = 1, x_t, z_1, z_2\}$, using two PROBIT models.[8] Lastly, we use the Plackett copula to model the likelihood of earnings ranks, conditional on employment, $c(u, v|x_t, z_1, z_2)$. All parameters of the PROBIT and Plackett models specifically depend on unobserved heterogeneity (z_1, z_2).

1.3.3 Estimation methodology

Because of unobserved heterogeneity, the log of the expected likelihood (with respect to the latent variable z) will not change the products in (1.3) into sums and because products are numerically instable the EM algorithm is usually thought as being the best way to maximize the likelihood. To estimate the parameter vector we use the

[8]Alternatively modelling the latent variables of the Probit models as autoregressive processes would make the estimation much more complicated.

sequential EM algorithm that is described in details in Appendix A.3 and we bootstrap the estimation sample to compute standard errors.

We estimate the model assuming that every three-year panel identifies a specific set of cross-section and mobility parameters. However, we force the parameters of the distribution of unobserved heterogeneity (z_1, z_2) conditional on z_0 $(\Pr\{z_1|z_0\}\Pr\{z_2|z_1, z_0\})$ to remain the same across the different panels. We impose this constraint because we believe that there is too little information in three years of panel data to identify both the distribution of earnings processes given heterogeneity (z_1, z_2) and the distribution of unobserved heterogeneity (z_1, z_2) given observed heterogeneity z_0 (education and cohort). The way the parameters of marginal distributions and copulas vary across three-year panels (1990-1992 to 2000-2002) thus tells us how much of the changes in earnings distributions is not explained by observed composition effects (i.e. changes in the distribution of education and cohort) and by experience, assuming that there is no unobserved composition effect.

Note that, by proceeding this way, we allow three individuals from three different three-year panels to face different parameter sets at the same calendar year. This may seem undesirable but it is well known that time, cohort and experience effects are not separately identified. Alternatively, we could estimate parameters assuming they remained fixed over a certain period of time. For example, we could contrast the periods 1990-1993, 1994-1998 and 1999-2002. This is the strategy followed by Gottshalk and Moffit (1994) who estimate different models for the 1970's and the 1980's. Our approach avoids the arbitrary choice of different time intervals.

1.3.4 Estimation results

In this section, we present the estimates of the distribution of heterogeneity and the estimates of the cross-section and mobility parameters corresponding to the 1990-1992 three-year panel, as the results for the other three-year panels are qualitatively similar. The parameter estimates are displayed in the various tables in Appendix A.7. Because they are numerous, we tried as much as possible to represent these results in the more accessible form of figures and frequency or mean summaries.

Number of unobserved individual types. The number of heterogeneity groups is likely to influence the interpretation. We used two popular criteria based on penalized likelihood to select these numbers. We found $(K_1 = 2, K_2 = 2)$ optimal using the Bayesian Information Criterion (BIC) and $(K_1 = 3, K_2 = 2)$ using Akaike's (1969) criterion.[9] For expositional convenience, we shall present the parameters for the minimal number of groups, $(K_1 = 2, K_2 = 2)$. We shall then analyze the fit for $(K_1 = 3, K_2 = 2)$.

Cross-sectional earnings distribution and unemployment risk. The first group of cross-section heterogeneity $(z_1 = 1)$ collects about 25% of the whole population. It gathers individuals with, on average, higher (aggregate mean of 9.33 vs 8.75) and more dispersed (aggregate standard deviation of 0.39 vs 0.28) monthly earnings than those in the second group. Figure 2.1 brings a visual confirmation that one component of the earnings mixture is both to the left of the other one and less dispersed.[10] The probability of being unemployed is found significantly lower for Group

[9]Let $\mathcal{L}$ be the log-likelihood of a given sample of size N, and K be the number of parameters in a given model. BIC maximizes $\mathcal{L} - \frac{1}{2}K\ln(N)$ with respect to K. Akaike maximizes $\mathcal{L} - K$. Obviously, the optimal number of parameters is larger for Akaike than for BIC.

[10]In this figure, the area below each density component is equal to the group's relative frequency, so that the sum of all components is the aggregate density.

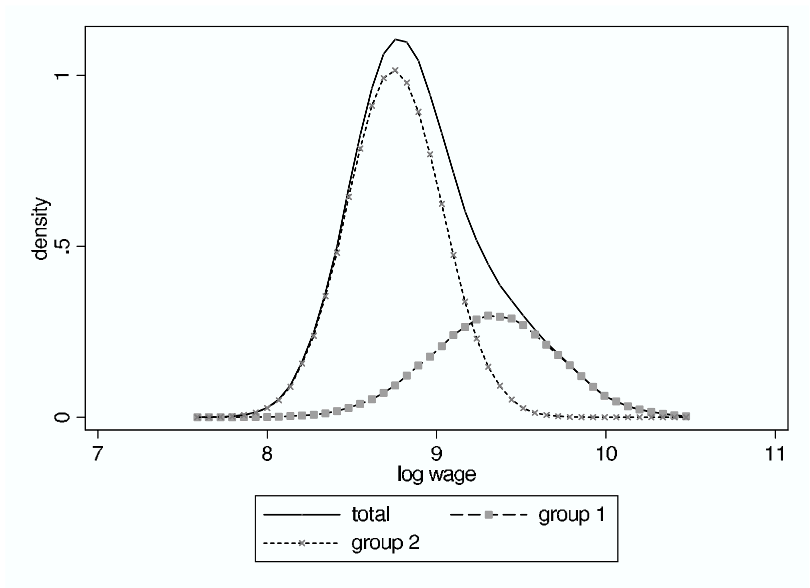


Figure 1.1: Components of the cross-sectional log-earnings density in 1990 ($K_1 = 2$)

1 than for Group 2 (aggregate probability of 3.1% vs 11.2%).

When a third group of heterogeneity is allowed for then a small third group of even poorer and more often unemployed workers is selected out of the second group (about 15% of them). The other two components of the earnings distribution mixture do not change much (Figure 1.2).

We also find that the experience-earnings profile is concave and the effect of experience on unemployment probabilities is U-shaped (Table F1 in Appendix A.7).

Cross-section heterogeneity. Table F2 in Appendix A.7 gives a full account of the estimates of the parameters of the (LOGIT) probability of $z_1 = 1$ (*versus* $z_1 = 2$) given the set z_0 of individual characteristics (education, cohort and interactions). To make the interaction of the cohort and education variables more easily interpretable, we compute in Table 1.2 the proportion of type-1 individuals by education and cohort (date of entry into the labour market).

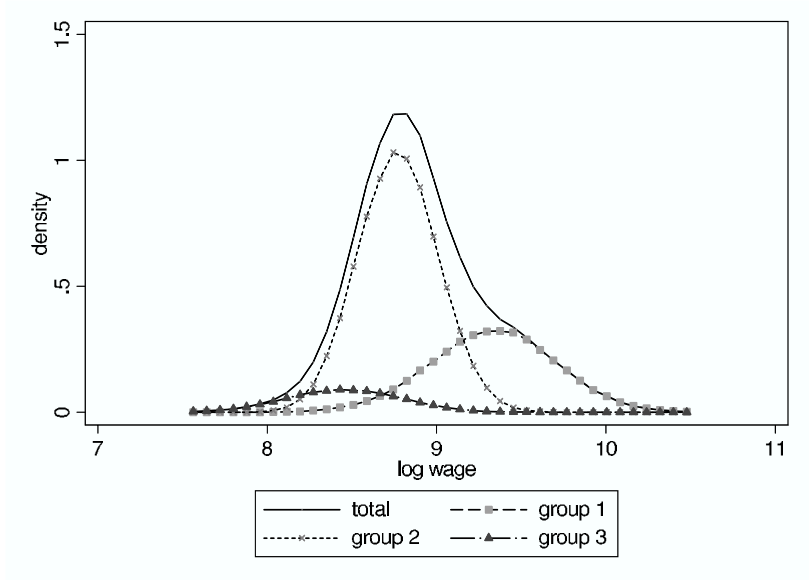


Figure 1.2: Components of the cross-sectional log-earnings density in 1990 ($K_1 = 3$)

First, one remarks that there is no direct cohort effect: all three cohorts have about the same probability of being in the first group (25%). By contrast, higher education unambiguously increases the probability of belonging to Group 1. But this effect is not uniform: only the top education group ("college +," i.e. more than three years of university) is guaranteed to belong to the top earnings group whatever the date of entry into the labour market. For the other education groups, older means better. This result is consistent with the accelerated democratization of the French education system between 1985 and 1995. The value of an education degree such as the *baccalauréat* (high school diploma) was much higher 20 or 30 years ago because *bacheliers* (high school graduates) were much fewer. This explains why 70% of the HS graduates of the oldest cohort belong to Group 1 when only 11% of the youngest HS graduates are classified in Group 1.

It is useful to know how the different individuals in the sample are reallocated

Cohort	None	Junior HS	Senior HS	Education Some college	College +	Overall
< 1970	.03	.11	.70	.90	.99	.24
1970 - 1985	.01	.05	.38	.86	.95	.25
> 1985	.01	.01	.11	.60	.97	.27
Overall	.01	.06	.30	.72	.97	.25

Table 1.2: Probability of being a high-wage earner ($z_1 = 1$) by education and cohort (model with 2 groups of cross-section heterogeneity)

when one allows for a third group of cross-section heterogeneity. For a given value of the postulated number of groups (K_1) and each individual in the sample, the posterior probability of belonging to one group given the individual's observations can be estimated. The group with the highest probability is the best predictor of individual types. Comparing the classification obtained for $K_1 = 2$ to the one obtained for $K_1 = 3$, ordering groups according to mean earnings, 83% of the individuals classified in Group 1 for $K_1 = 2$ remain classified in this group when a third group is allowed for and that about 16% move to Group 2. Only 1% move to Group 3. Moreover, 75% of type-2 individuals remain in Group 2 but 17% move to the third group and 8% to the first group.

Earnings mobility. Table F3 in Appendix A.7 displays the estimates of the group-specific parameters of the Plackett copulas used to model the transition process of marginal earnings ranks. The sign of the experience coefficient implies that wage mobility decreases with experience among low-wage earners ($z_1 = 2$) and increases with experience among high-wage earners ($z_1 = 1$). To give a better idea of the magnitude of wage mobility differentials across worker types, we display in Table 1.3 the predicted quintile-to-quintile probability transition matrices for the four combinations of (z_1, z_2) and three levels of experience: 5 years, 20 years and 35 years.

Experience	$z_1 = 1$, $z_2 = 1$					$z_1 = 1$, $z_2 = 2$				
5 years	.92	.07	.01	.00	.00	.87	.11	.01	.01	.00
	.07	.83	.09	.01	.00	.11	.74	.13	.01	.01
	.01	.09	.82	.09	.01	.01	.13	.72	.13	.01
	.00	.01	.09	.83	.07	.01	.01	.13	.74	.11
	.00	.00	.01	.07	.92	.00	.01	.01	.11	.87
20 years	.91	.08	.01	.00	.00	.82	.15	.02	.01	.00
	.08	.82	.09	.01	.00	.15	.65	.17	.02	.01
	.01	.09	.80	.09	.01	.02	.17	.62	.17	.02
	.00	.01	.09	.82	.08	.01	.02	.17	.65	.15
	.00	.00	.01	.08	.91	.00	.01	.02	.15	.82
35 years	.90	.09	.01	.00	.00	.75	.19	.04	.01	.01
	.09	.80	.10	.01	.00	.19	.56	.20	.04	.01
	.01	.10	.78	.10	.01	.04	.20	.52	.20	.04
	.00	.01	.10	.80	.09	.01	.04	.20	.56	.19
	.00	.00	.01	.09	.90	.01	.01	.04	.19	.75
	$z_1 = 2$, $z_2 = 1$					**$z_1 = 2$, $z_2 = 2$**				
5 years	.74	.20	.04	.02	.00	.41	.26	.16	.10	.07
	.20	.53	.21	.04	.02	.26	.27	.22	.15	.10
	.04	.21	.50	.21	.04	.16	.22	.24	.22	.16
	.02	.04	.21	.53	.20	.10	.15	.22	.27	.26
	.00	.02	.04	.20	.74	.07	.10	.16	.26	.41
20 years	.81	.16	.02	.01	.00	.51	.26	.12	.07	.04
	.16	.63	.18	.02	.01	.26	.32	.23	.12	.07
	.02	.18	.60	.18	.02	.12	.23	.30	.23	.12
	.01	.02	.18	.63	.16	.07	.12	.23	.32	.26
	.00	.01	.02	.16	.81	.04	.07	.12	.26	.51
35 years	.86	.12	.01	.01	.00	.60	.25	.09	.04	.02
	.12	.72	.14	.01	.01	.25	.39	.23	.09	.04
	.01	.14	.70	.14	.01	.09	.23	.36	.23	.09
	.01	.01	.14	.72	.12	.04	.09	.23	.39	.25
	.00	.01	.01	.12	.86	.02	.04	.09	.25	.60

Table 1.3: Predicted transition probability matrices (across quintiles) by experience and unobserved heterogeneity

The six featured matrices unambiguously designate individuals with $z_2 = 1$ as a group of "stayers" and those with $z_2 = 2$ as a group of "movers": the numbers in the first diagonal of the matrices corresponding to the second group are indeed smaller and the off-diagonal ones larger, irrespective of the cohort of origin. Note that there is less heterogeneity among high-wage individuals (i.e. those with $z_1 = 1$) than among low-wage individuals ($z_1 = 2$). Moreover, the effect of experience is more pronounced for low-wage workers than for high-wage workers.

Employment dynamics. Table F4 in Appendix A.7 presents the parameters of the PROBIT probabilities of unemployment at $t+1$ conditional on employment at t and of unemployment at $t+1$ conditional on unemployment at t, and Table 1.4 below shows the implied levels of these transition probabilities for the four combinations of (z_1, z_2) and the same three levels of experience as before. Compared to individuals with $z_2 = 2$ (movers), individuals with $z_2 = 1$ (stayers) clearly have very little chance of becoming unemployed. Yet, when they become unemployed they tend to remain unemployed longer. Being both a low-wage worker ($z_1 = 2$) and a mover ($z_2 = 2$) maximizes the chance of becoming unemployed. Older workers tend to remain unemployed longer while the risk of becoming unemployed decreases with experience.[11]

Mobility-specific heterogeneity. Table F5 in Appendix A.7 describes the probability of $z_2 = 1$ (*versus* $z_2 = 2$), conditional on cross-section heterogeneity z_1 and other covariates in z_0. As before, these parameters are more easily interpreted after computing the predicted probability of $z_2 = 1$ by education, cohort and cross-section heterogeneity (Table 1.5). There are slightly fewer stayers among high-wage earners ($z_1 = 1$) than among low-wage earners: about 50% and 60%, respectively. Moreover,

[11] For all but high-wage workers ($z_1 = 1$). But, high-wage workers have a very low risk of unemployment.

		$z_1 = 1$			
		$z_2 = 1$		$z_2 = 2$	
		year $t+1$			
Experience		empl.	unempl.	empl.	unempl.
	year t				
5 years	empl.	1.00	.00	.97	.03
	unempl.	.73	.27	.99	.01
20 years	empl.	1.00	.00	.97	.03
	unempl.	.02	.98	.87	.13
35 years	empl.	.99	.01	.97	.03
	unempl.	.00	1.00	.47	.53
		$z_1 = 2$			
		$z_2 = 1$		$z_2 = 2$	
		empl.	unempl.	empl.	unempl.
5 years	empl.	.89	.11	.91	.09
	unempl.	.36	.64	.91	.09
20 years	empl.	.99	.01	.93	.07
	unempl.	.19	.81	.68	.32
35 years	empl.	1.00	.00	.94	.06
	unempl.	.08	.92	.33	.67

Table 1.4: Transition probabilities across employment and unemployment by experience and unobserved heterogeneity

younger cohorts of high-wage earners are more mobile than older ones. The opposite conclusion applies to low-wage earners. Notice that the pseudo R^2 of the PROBIT regressions of unobserved heterogeneity variables z_1 and z_2 on observed covariates z_0 (50% vs less than 4%; see Table F2 and F5) indicate that education and cohort do not determine z_2 as much as they determine z_1.

	Cross-section heterogeneity: $z_1 = 1$					
	Education					
Cohort	None	Junior HS	Senior HS	Some college	College +	Overall
< 1970	n.s.	.58	.56	.73	.60	.61
1970 - 1985	n.s.	.40	.53	.58	.53	.54
> 1985	n.s.	.33	.31	.47	.45	.44
Overall	n.s.	.50	.50	.55	.50	.52
	Cross-section heterogeneity: $z_1 = 2$					
	Education					
Cohort	None	Junior HS	Senior HS	Some college	College +	Overall
< 1970	.44	.50	.60	.72	n.s.	.49
1970 - 1985	.60	.61	.65	.52	n.s.	.61
> 1985	.75	.73	.69	.65	n.s.	.71
Overall	.60	.61	.67	.63	n.s.	.62

Table 1.5: Conditional probability of being a stayer ($z_2 = 1$)

1.3.5 Model fit

In this subsection, we analyze the ability of the model to fit the data for the choice of the pair $(K_1, K_2) = (3, 2)$. Again, we focus on the years 1990-1992, because the results are very similar for the other periods.

Cross-Sections. Usual autoregressive specifications of earnings dynamics have troubles fitting marginal densities because their estimation interferes with the estimation of the dynamic parameters. For instance, the however rich model of Geweke and Keane (2000) underestimates the mode of the marginal distributions. This is clearly

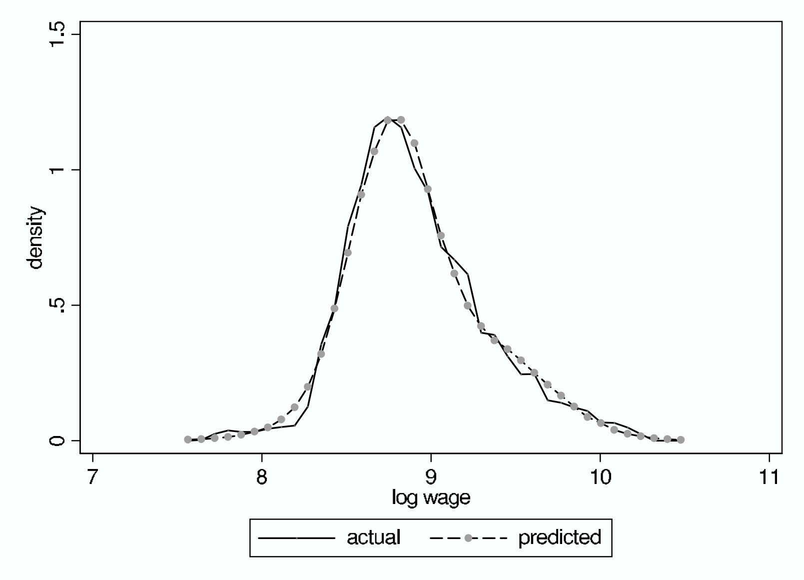


Figure 1.3: Cross-sectional earnings fit for $K_1 = 3$

not the case of our model. Figure 1.3 shows that a simple three-component mixture of normal distributions achieves a remarkable fit.

Earnings Mobility. We then analyze the capacity of the model to fit earnings mobility as characterized by the yearly dynamics of earnings quintiles. To this end, we use the estimated model to simulate individual states at dates $t+1$ and $t+2$ conditional on the observed state at t (=1990). Appendix A.5 details our simulation procedure. We then compare actual and predicted aggregate transition probability matrices.

The actual aggregate 1990-1991 transition probability matrix and the predicted matrix are as follows:

$$P_{01} = \begin{pmatrix} .72 & .19 & .06 & .02 & .01 \\ .21 & .54 & .19 & .05 & .01 \\ .06 & .21 & .53 & .18 & .02 \\ .01 & .05 & .20 & .61 & .13 \\ .00 & .01 & .02 & .14 & .83 \end{pmatrix}; \quad \widehat{P}_{01} = \begin{pmatrix} .70 & .20 & .07 & .02 & .01 \\ .21 & .54 & .20 & .04 & .01 \\ .06 & .21 & .53 & .18 & .02 \\ .02 & .04 & .18 & .62 & .14 \\ .01 & .01 & .02 & .14 & .82 \end{pmatrix}.$$

These two matrices are very close indeed, confirming the ability of the Plackett family to fit empirical earnings data.

Over two years, unobserved heterogeneity is also found to account properly for most of the non first-order-Markov state dependence. The actual 1990-1992 transition probability matrix and the predicted one are, respectively:

$$P_{02} = \begin{pmatrix} .67 & .22 & .08 & .02 & .01 \\ .24 & .49 & .22 & .04 & .01 \\ .06 & .24 & .50 & .18 & .02 \\ .02 & .04 & .18 & .59 & .17 \\ .01 & .01 & .02 & .17 & .79 \end{pmatrix}; \quad \widehat{P}_{02} = \begin{pmatrix} .63 & .24 & .09 & .03 & .01 \\ .24 & .43 & .22 & .09 & .02 \\ .08 & .24 & .43 & .21 & .04 \\ .04 & .08 & .22 & .50 & .16 \\ .01 & .01 & .04 & .17 & .77 \end{pmatrix}.$$

The fit is slightly worse than before but considerably better than with the homogeneous model (i.e. when predicting P_{02} by P_{01}^2; see subsection 1.2.3).

Three-year averages of individual earnings. In this paragraph, we test the ability of the model to fit the distribution of individual log-earnings averaged over the three panel dates (we here only use the balanced panel of individuals with non missing earnings at all three dates). This is a way of analyzing the capacity of the model to fit the dynamics of earnings levels instead of earnings ranks.

We simulate individual trajectories over years $t+1$ and $t+2$ conditional on year-t earnings. The actual distribution of three-year log-earnings means and its prediction using simulated data are displayed in Figure 1.4. The model fit is only slightly worse than for one single cross-section (Figure 1.3).

Moreover, to show the importance of taking mobility into account when simulating earnings sequences, we also provide in this figure the distribution of three-year earnings means simulated under the assumption of independence.[12] Omitting earnings dynamics has dramatic effects on the distribution of three-year averages. With a

[12]This corresponds to the use of the independent copula: $C^{\perp}(u,v) = uv$, which is a particular case of the Plackett copula with parameter η approaching zero. See Appendix A.2.

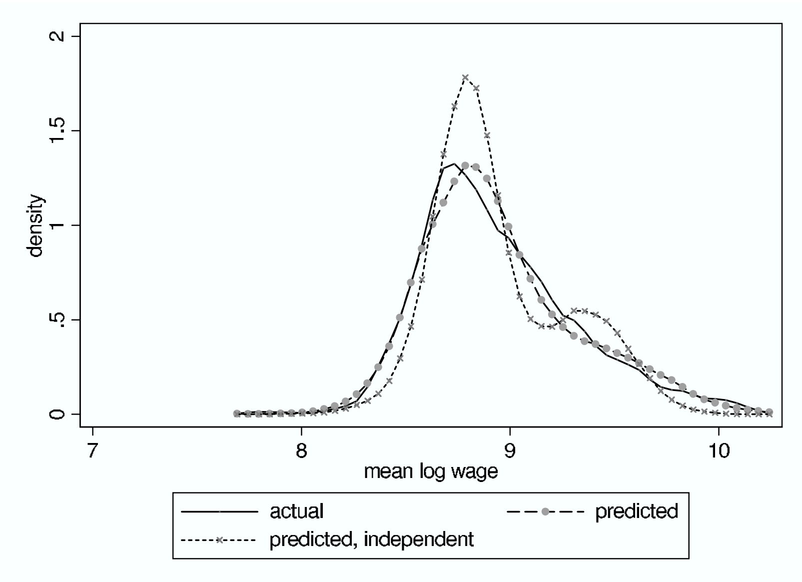


Figure 1.4: Effect of mobility on the distribution of 3-year log-earnings means ($K_1 = 3$)

single cross-section the different components of the earnings mixture are not separate enough to create multiple modes. When one follows individuals over time, however, low-wage and high-wage earners accumulate in different intervals with sufficient densities for different modes to appear.

1.4 Changes and composition of earnings inequality and mobility in France over 1990-2002

Having shown that our model fits individual employment trajectories well, we now use it to decompose inequality and mobility indices into various components of interest. We shall compare two alternative specifications: The first specification conditions earnings and employment dynamics on experience (x_t), education and cohort (z_0) but assumes away any other source of individual heterogeneity. The second specification conditions individual dynamics on experience and unobserved heterogeneity (z_1 and

z_2). Comparing both specifications will tell us how useful it is to allow for unobserved heterogeneity.

1.4.1 Cross-sectional inequality

Let Y be a random variable and let X, Z be two vectors of conditioning variables. The variance of Y can be decomposed as follows:

$$\begin{aligned} \mathbb{V}(Y) &= \mathbb{V}(\mathbb{E}(Y|X)) + \mathbb{E}(\mathbb{V}(Y|X)) \\ &= \mathbb{V}(\mathbb{E}(Y|X)) + \mathbb{E}[\mathbb{V}(\mathbb{E}(Y|X,Z)|X) + \mathbb{E}(\mathbb{V}(Y|X,Z)|X)] \\ &= \mathbb{V}(\mathbb{E}(Y|X)) + \mathbb{E}[\mathbb{V}(\mathbb{E}(Y|X,Z)|X)] + \mathbb{E}(\mathbb{V}(Y|X,Z)). \end{aligned} \tag{1.4}$$

Equation (1.4) generically decomposes $\mathbb{V}(Y)$ into three components: $\mathbb{V}(\mathbb{E}(Y|X))$ is the variance of Y *between* the individual groups defined by the values of X, $\mathbb{E}[\mathbb{V}(\mathbb{E}(Y|X,Z)|X)]$ is the variance of Y *within* individual groups with the same value of X and *between* groups defined by Z, and $\mathbb{E}(\mathbb{V}(Y|X,Z))$ is the residual variance of Y *within* both X and Z groups. In the application below, X is experience and Z is individual heterogeneity. Notice that if $X = 1$ and $Z = z_1$, the variance decomposition degenerates into the one computed by Gottschalk and Moffit (1994) who interpret the between term as the "permanent" component of earnings shocks and the within term as reflecting the contribution of "transitory" shocks.

Labour earnings. The evolution of cross-sectional earnings inequality over 1990-2000 is hump-shaped, with a peak around 1995 (the solid line in Figure 1.5; the interpretation of the dotted curve will be given below). The log-earnings variance and the ninety-ten percentile ratio both follow the same evolution. Between 1990 and 1995 the variance of employees' log-earnings increases from .166 to .183 in 1995, that is a 10% rise.

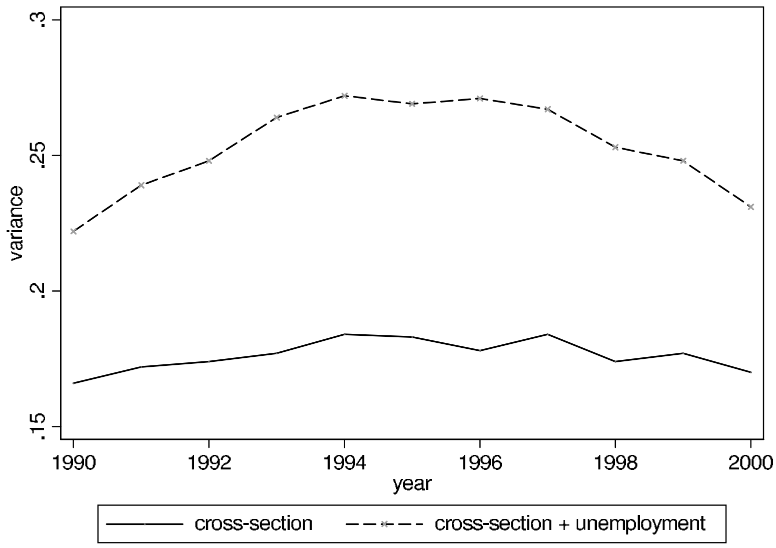


Figure 1.5: Evolution of log-earnings cross-section variance ($K_1 = 3$; replacement ratio of 60%)

Table 1.6 provides the log-earnings variance analyses of five estimation samples: 1990-92, 1993-95, 1995-97, 1997-99 and 2000-2002,[13] and Figure 1.6 plots the evolution of the shares of various between components. First, education explains significantly less aggregate variance than unobserved heterogeneity: about a quarter of the variance is explained by education whereas three groups of unobserved cross-section heterogeneity explain 50%. Experience alone explains about 7% and 43% remain unexplained.[14] This result clearly shows that allowing for unobserved heterogeneity can strongly affect the study of earnings inequality, since more heterogeneity implies that a larger share of cross-section earnings inequality will be transmitted to life-cycle

[13] D9/D1 is the ninety-ten percentile ratio.

[14] Notice that thanks to the quasi-independence of z_1 and x_t and to the assumption that the distribution of log-earnings, y_t, is Gaussian conditional on x_t and z_1, the order of the variance decomposition (X first, Z second or *vice versa*) does not matter. The log-earnings variance is therefore approximately equal to the sum of the variance "explained" by experience plus the variance explained by heterogeneity plus the residual variance.

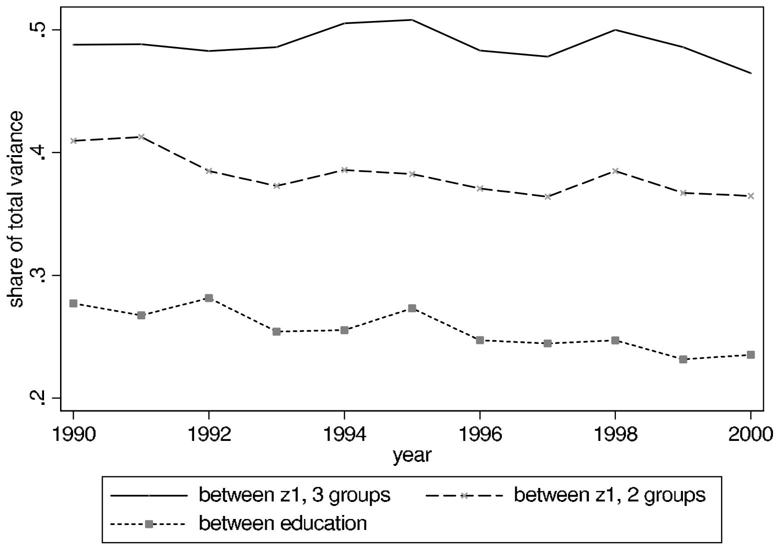


Figure 1.6: Cross-sectional log-earnings variance decomposition ($K_1 = 3$)

income inequality. Second, there is a clear decreasing trend in the share explained by education whereas this declining pattern is less marked for unobserved heterogeneity (z_1). Lastly, business cycle fluctuations does not affect these shares much.

Unemployment propensity. For unemployment risk we repeat the variance decomposition exercise on the latent variable implicit in the PROBIT model of unemployment probability, forcing the residual variance to be equal to one.[15] The results are displayed in Table 1.7.

The peak of inequality is attained in 1993, which corresponds to the bottom of

[15] We write $\Pr\{e_t = 0|x_t, z_1\} = \Pr\{y_t^* > 0|x_t, z_1\}$ with $y_t^* = x_t'\delta_{z_1} + u_t$ and $u_t \sim \mathcal{N}(0,1)$. We then compute the different variance components as follows:

$$\begin{aligned} \mathbb{V}(\mathbb{E}(y_t^*|x_t)) &= \mathbb{V}(x_t'\delta_{z_1}), \\ \mathbb{E}[\mathbb{V}(\mathbb{E}(y_t^*|x_t, z_1)|x_t)] &= \mathbb{E}[\mathbb{V}(x_t'\delta_{z_1}|x_t)) \\ \mathbb{E}(\mathbb{V}(y_t^*|x_t, z_1)) &= \mathbb{V}(u_t) = 1. \end{aligned}$$

	1990	1993	1995	1997	2000
Mean	8.91	8.93	8.91	8.92	8.93
D9/D1	2.85	2.95	3.02	3.02	2.88
Variance	.166	.177	.183	.184	.170
Observed heterogeneity model					
Between experience	.010	.010	.012	.014	.011
Within experience/ Between z_0	.053	.054	.060	.058	.054
Between z_0	.046	.045	.050	.045	.040
Within z_0/Between experience	.017	.019	.022	.027	.025
Within (experience, z_0)	.103	.113	.111	.112	.105
Unobserved heterogeneity model, $K_1 = 2$					
Between experience	.009	.010	.013	.014	.012
Within experience/ Between z_1	.068	.066	.070	.070	.064
Between z_1	.068	.066	.070	.067	.062
Within z_1/Between experience	.009	.010	.013	.017	.014
Within (experience, z_1)	.089	.101	.100	.100	.094
Unobserved heterogeneity model, $K_1 = 3$					
Between experience	.008	.009	.011	.013	.011
Within experience/ Between z_1	.081	.086	.094	.089	.080
Between z_1	.081	.086	.093	.088	.079
Within z_1/Between experience	.009	.008	.012	.014	.012
Within (experience, z_1)	.077	.082	.079	.082	.078

Notes: z_0 = education, cohorts, interactions education/cohorts;
z_1 = unobserved heterogeneity in cross-sections;
z_2 = unobserved heterogeneity in copulas.

Table 1.6: Variance analysis of log-earnings cross-sections (employees only)

the French business cycle bust of the 1990's. Moreover, there is little difference between the observed heterogeneity model, and the model with two cross-section groups ($K_1 = 2$): for both specifications, the model explains around ten percent of the aggregate variance of unemployment risk.

Allowing for a third cross-section group ($K_1 = 3$) improves the prediction significantly: the model now explains almost one third of the unemployment propensity variance. Furthermore, three groups of unobserved heterogeneity (z_1) explain about 20% of the total variance. This is considerably more than the 4-5% two groups can explain, thereby showing that the third group of cross-section heterogeneity is composed of workers who have a high propensity of being unemployed.

Labour earnings and unemployment compensation. In the final step of our analysis of cross-section earnings inequality we intend to measure the potential impact of unemployment on inequality. To this end, we simulate cross-sections of individual earnings and unemployment compensation income (with three groups of unobserved cross-section heterogeneity). For given parameter values (e.g. estimated on the 1993-1995 panel) we assign to each individual in the panel his most probable unobserved type (z_1, z_2) given his observations, using the posterior probabilities estimated by the EM algorithm. Then, we draw an initial employment state and an initial wage given state variables (experience, education, cohort and unobserved heterogeneity). If unemployed, we impute a compensation income equal to a fraction of his previous wage (observed or drawn from the estimated distribution if unobserved).

The dotted line in Figure 1.5 (above) shows the level and the evolution of the variance of all workers' log-earnings using a replacement ratio of 60% for unemployment compensation income, a value close to the ratio computed in OECD publications for France (Martin, 1996). The solid line plots the series of log-earnings variances for

	1990	1993	1995	1997	2000
Observed heterogeneity model					
Total Variance	1.125	1.138	1.124	1.143	1.113
Between experience	.053	.070	.071	.072	.040
Within experience/ Between z_0	.075	.071	.052	.058	.072
Between z_0	.072	.059	.043	.061	.056
Within z_0/Between experience	.056	.078	.079	.082	.056
Within (experience, z_0)	1	1	1	1	1
Unobserved heterogeneity model, $K_1 = 2$					
Total Variance	1.154	1.125	1.137	1.111	1.094
Between experience	.048	.070	.075	.074	.046
Within experience/ Between z_1	.106	.055	.062	.038	.047
Between z_1	.092	.058	.050	.053	.046
Within z_1/Between experience	.062	.066	.087	.058	.047
Within (experience, z_1)	1	1	1	1	1
Unobserved heterogeneity model, $K_1 = 3$					
Total Variance	1.352	1.468	1.395	1.359	1.334
Between experience	.067	.119	.114	.114	.083
Within experience/ Between z_1	.284	.349	.245	.32	.251
Between z_1	.246	.311	.249	.228	.233
Within z_1/Between experience	.105	.157	.146	.131	.101
Within (experience, z_1)	1	1	1	1	1

Table 1.7: Cross-sectional variance analysis of the latent unemployment propensity variable

employees. As expected, unemployment risk increases inequality a lot. The evolution of this (static) measure of employed and unemployed workers' earnings inequality is also humped-shaped and shows no trend over the 1990-2002 period. The rise in inequality between 1990 and 1995 is however more pronounced: with three groups of unobserved heterogeneity z_1 and a replacement ratio of 60% the rise in earnings inequality between 1990 and 1995 is around 23%. Unemployment risk increases in business cycle busts and therefore contributes more to earnings inequality in busts than in booms.[16]

1.4.2 Relative earnings mobility

Among the many possible mobility indicators proposed in the literature,[17] we selected Spearman's rho for being both natural in the current framework and simple to compute. Let X and Y be two random variables with marginal cdf's F^X and F^Y, then the Spearman rho between X and Y is the correlation coefficient between ranks $U = F^X(X)$ and $V = F^Y(Y)$, i.e.

$$\rho_S(X,Y) = \rho(U,V) = \frac{Cov(U,V)}{\sqrt{Var(U)Var(V)}}.$$

We compute a Spearman rho for different values of the heterogeneity vector $z = (z_0, z_1, z_2)$ using the procedure described in Appendix A.6. Table 1.8 shows that the dispersion of ρ_S essentially originates from differences in unobserved heterogeneity (z_1, z_2). Education generates very little mobility differences. Experience generally has a increasing effect on ρ_S of moderate size when compared to unobserved heterogeneity, but of significant size when compared to education. Besides, Table 1.8 confirms the analysis of Section 2.5. Higher z_2 is unambiguously associated with being a mover, and thus with more mobility. The effect of z_1 is less clear as there are more movers among high-wage individuals ($z_1 = 1$) but these individuals are less mobile than low-wage movers ($z_1 = 2$).

The evolution of Spearman rho's over the nineties displays the same hump-shaped

[16]To evaluate the sensitivity of this effect to the replacement ratio, we computed the contribution of unemployment risk to total earnings inequality, for three different replacement ratios: 40%, 60% and 80%. This contribution is computed as the cross-section variance of unemployed workers' replacement income in proportion to the variance of employees' earnings and unemployed workers' replacement income (in log). Unemployment risk explains about 40-50% of the overall log-earnings variance for a replacement ratio of 40%, about 30% for a ratio of 60% and less than 20% for a ratio of 80%. This emphasizes how misleading international comparisons of earnings inequality can be if they do not properly account for unemployment risk.

[17]See Fields and Ok (1999) for a comprehensive survey on mobility indices.

	1990	1993	1995	1997	2000
Observed heterogeneity model					
Overall	.861	.903	.915	.914	.890
Education = 1	.857	.906	.911	.909	.890
Education = 2	.854	.904	.919	.919	.885
Education = 3	.879	.895	.915	.913	.898
Education = 4	.870	.893	.900	.913	.895
Education = 5	.867	.915	.918	.902	.887
Experience = 5	.826	.875	.886	.880	.850
Experience = 20	.854	.904	.912	.913	.885
Experience = 35	.889	.913	.934	.929	.917
Unobserved heterogeneity model, $K_1 = 2$					
Overall	.798	.861	.866	.875	.847
$z_1 = 1$	.827	.866	.863	.848	.835
$z_1 = 2$	.707	.859	.868	.887	.851
$z_2 = 1$	.919	.907	.960	.911	.886
$z_2 = 2$	.631	.794	.717	.823	.790
Experience = 5	.822	.844	.836	.852	.819
Experience = 20	.804	.876	.889	.888	.866
Experience = 35	.846	.866	.879	.882	.894
Unobserved heterogeneity model, $K_1 = 3$					
Overall	.759	.828	.829	.844	.812
$z_1 = 1$	.803	.851	.860	.839	.822
$z_1 = 2$	.757	.825	.832	.863	.822
$z_1 = 3$	.554	.700	.567	.627	.630
$z_2 = 1$	.900	.885	.950	.894	.859
$z_2 = 2$	.589	.752	.662	.772	.742
Experience = 5	.771	.779	.716	.753	.746
Experience = 20	.774	.837	.823	.847	.828
Experience = 35	.820	.823	.863	.855	.838

Note: Education code is 1 = "no degree", 2 = "junior HS", 3 = "senior HS", 4 = "some college", 5 = "college graduate"

Table 1.8: Relative mobility measured by Spearman rho

form as before. In the context of the French recession of 1993, higher wage inequality appears positively associated with more inequality in unemployment risk and less relative wage mobility: there is more cross-section inequality in busts than in booms and less mobility. This results does not depend on which specification of the empirical model we use.

1.5 Longitudinal inequality

The aim of this section is to complete the static picture of earnings inequality displayed in the preceding section by incorporating into the analysis unemployment risk, relative mobility and unobserved heterogeneity. To this end, we simulate individual trajectories over a fixed horizon and compute an intertemporal income utility of the form

$$V = \frac{\sum_{t=1}^{T} (1+r)^{-t} U(R_t)}{\sum_{t=1}^{T} (1+r)^{-t}},$$

where r is the discount rate, arbitrarily set equal to 10%, R_t is year-t labour earnings level or unemployment compensation income, and U is an increasing function (we use a logarithmic utility in the sequel). The distribution of intertemporal utility then yields new inequality indices.

A detailed presentation of our simulation strategy is given in Appendix A.5. For a given set of parameter estimates and for each individual in a three-year panel, we draw a couple of heterogeneity types (z_1, z_2), an initial employment state and an initial replacement income for unemployed workers, as indicated in 1.4.1. Then, we draw subsequent states and earnings using the estimated employment-unemployment transition probability matrices and copula parameters.

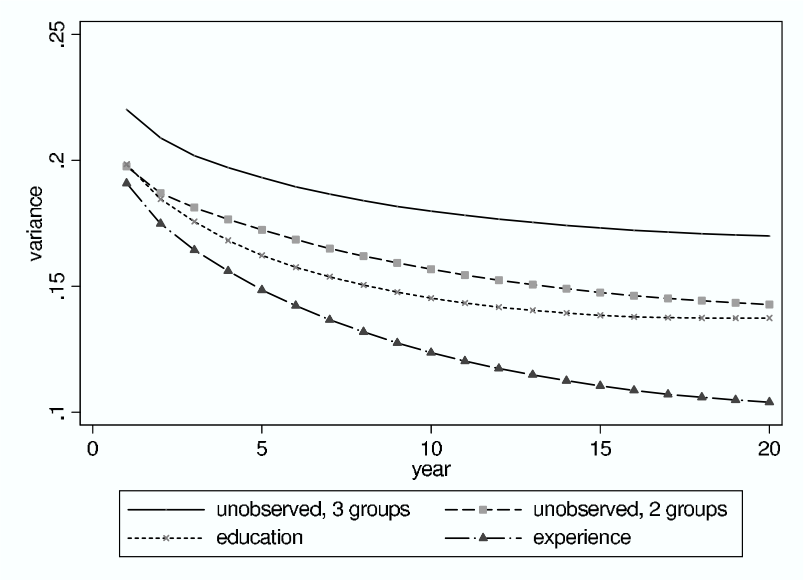


Figure 1.7: The equalizing effect of income mobility

1.5.1 Evolution

In Figure 1.7, we compare and contrast the evolution of the intertemporal log-earnings variance for four different specifications of the empirical model: a first version in which the only variable conditioning the earnings/employment process is experience and the only dispersion factor is therefore the individual date of entry into the labour market (the "homogeneous" model); a second version adds education to cohort in the list of conditioning variables (the "observed heterogeneity" model); lastly, we consider both versions of the unobserved heterogeneity model with $K_1 = 2$ or $K_1 = 3$ and with $K_2 = 2$. We use the parameters estimated from the 1990-1992 panel as they are representative of all other panels' estimates.

The first observation we draw from Figure 1.7 is that the more heterogeneity we allow for the more intertemporal inequality there is. The model with 3 groups of cross-section unobserved heterogeneity ($K_1 = 3$) generates far more inequality than the models with observed heterogeneity or the model with only two groups of cross-section unobserved heterogeneity ($K_1 = 2$). Interestingly, we again observe that setting $z_1 = z_0$ is approximatively the same as constraining the support of z_1 to two mass points. This clearly illustrates the importance of allowing for unobserved heterogeneity in a flexible way.

The dispersion of intertemporal log-earnings decreases with the horizon over which we average future log-earnings. This is the well-known equalizing effect of mobility. However, even over very long horizons, there remains a substantial amount of inequality. For instance, the variance of the discounted sum of log-earnings over a horizon of twenty years is 70% of the variance of one-year log-earnings when we condition individual trajectories on experience and education only ("observed heterogeneity") and it is still 80% of the one-year variance for the unobserved-heterogeneity model with $K_1 = 3$ and $K_2 = 2$ (see Figure 1.8). In comparison, Gottschalk and Moffit (1994), who construct permanent income measures by averaging earnings over ten years, find that the "permanent" earnings variance is equal to two-third of the total log-earnings variance. The numbers we obtain are robust to changes in the discount rate and the replacement ratio for unemployed workers' earnings. Setting the discount rate to zero, varying the replacement ratio had little effect on the size of the variance reduction for various horizons. Therefore, mobility might have a weaker equalizing role in France than in the US.

Lastly, the equalizing effect of mobility does not seem to vary much over the period. Intertemporal log-earnings inequality changes over time in a similar way as in cross-section, once unemployment is taken into account (see Figure 1.8). In particular, intertemporal earnings inequality also shows a peak around 1995. This was to be expected as we already saw that earnings inequality, unemployment risk

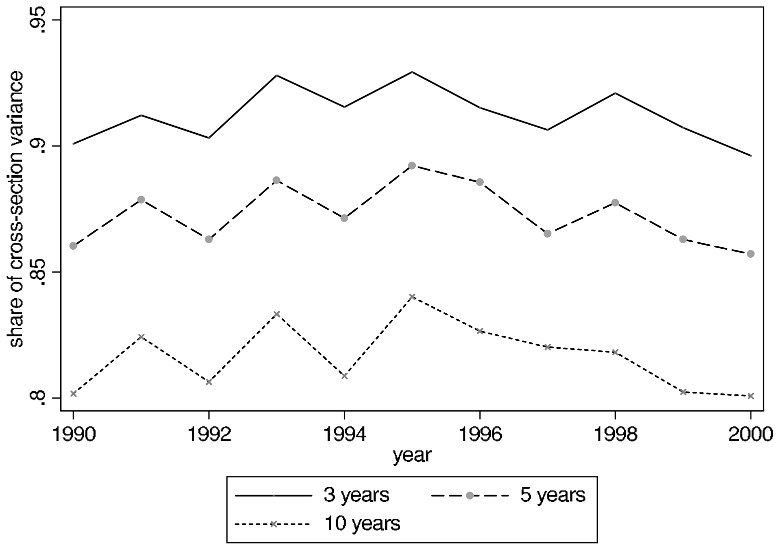


Figure 1.8: Intertemporal log-earnings variance evolution in proportion to one-year log- earnings variance computed for various time horizons (unobserved-heterogeneity model with $K_1 = 3, K_2 = 2$). The variance of one-year log-earnings is obtained by simulation, imputing a compensation income equal to 60% of previous earnings to the unemployed.

and immobility increase in business cycle busts. Note that more heterogeneity changes the level of inequality but does not affect time changes: for example, for a ten-year horizon, all models find about 20-25% rise in variance and about 10% rise in the 90-10 percentile ratio between 1990 and 1995.

1.5.2 Variance analysis

We now address the issue of the respective importance of each dispersion factor (z_0, z_1, z_2) in determining the variance of intertemporal income utility. Let Z denote a dispersion factor. One can write:

$$\left(\sum_{t=1}^{T}\frac{1}{(1+r)^t}\right)^2 \mathbb{V}(V) \equiv \mathbb{V}\left(\sum_{t=1}^{T}\frac{U(R_t)}{(1+r)^t}\right) \tag{1.5}$$

$$= \underbrace{\mathbb{V}\left(\mathbb{E}\left(\sum_{t=1}^{T}\frac{U(R_t)}{(1+r)^t}|Z\right)\right)}_{\text{between-}Z} + \underbrace{\mathbb{E}\left(\mathbb{V}\left(\sum_{t=1}^{T}\frac{U(R_t)}{(1+r)^t}|Z\right)\right)}_{\text{within-}Z}$$

$$= \underbrace{\mathbb{V}\left(\sum_{t=1}^{T}\frac{\mathbb{E}[U(R_t)|Z]}{(1+r)^t}\right)}_{\text{between-}Z}$$

$$+\underbrace{\mathbb{E}\left(\sum_{t=1}^{T}\frac{\mathbb{V}[U(R_t)|Z]}{(1+r)^{2t}}\right)}_{\text{within-}Z\text{, cross-section}} + \underbrace{\mathbb{E}\left(\sum_{t\neq t'}\frac{Cov\left[U(R_t),U(R_{t'})|Z\right]}{(1+r)^{t+t'}}\right)}_{\text{within-}Z\text{, dynamics}}.$$

Equation (1.5) shows that the intertemporal log-earnings variance $\mathbb{V}(V)$ is the sum of three components. The first component (between) reflects *permanent* inequality determined by the dispersion of individual heterogeneity (Z). The residual variance (within) is itself the sum of two terms. One term is the contribution of cross-section inequality, or the part of the variance within homogeneous groups which would subsist, were there be no autocorrelation in individual ranks. This term does not vanish for large T because of the non-zero discount rate.[18] Yet, for reasonable values of discount rates and sufficiently large horizon values, this effect should be small. We call it *residual transitory.* The second term is the pure contribution of earnings dynamics. We call it *persistent.*

Notice that, if Z comprises all dispersion factors, i.e. $Z = (x_{t,}z_0, z_1, z_2)$, then

[18]We use a 10% discount rate. The Law of Large Numbers does not apply and the limiting intertemporal utility is a non degenerate random variable. Indeed, let $(x_t)_{t=1}^{T}$ be an i.i.d. sequence of r.v.'s, then $\frac{\sum_{t=1}^{T}\rho^t x_t}{\sum_{t=1}^{T}\rho^t}$ converges a.s. to a r.v. with mean $\mathbb{E}x_t$ and variance

$$\lim_{T\to\infty}\mathbb{V}\left(\frac{\sum_{t=1}^{T}\rho^t x_t}{\sum_{t=1}^{T}\rho^t}\right) = \frac{1-\rho}{1+\rho}\mathbb{V}x_t.$$

$\mathbb{E}[U(R_t)|Z] = \mathbb{E}[U(R_t)|x_{t,}z_1]$ and $\mathbb{V}[U(R_t)|Z] = \mathbb{V}[U(R_t)|x_t, z_1]$. Hence, pure mobility heterogeneity (z_2) does not contribute to permanent and residual transitory effects. The influence of z_2 on intertemporal inequality is thus embodied in the persistent term of the above decomposition.

Results. Table 1.9 displays the results of the preceding intertemporal log-earnings variance analysis. First, one remarks that the between-cohort component of the variance of the discounted sum of log-earnings over ten years accounts for less than five percent of the aggregate variance. This term is the only source of "permanent" inequality in the homogeneous model (experience is the only covariate). In this specification, the main part (around 75%) of intertemporal log-earnings inequality is therefore persistent. As before, the model with observed heterogeneity and the model with unobserved heterogeneity and $K_1 = K_2 = 2$ give similar results. For these two models, heterogeneity explains about 40% of the intertemporal variance, the rest being divided into persistent (around 45%) and residual transitory inequality (around 10%). If a third group of cross-section heterogeneity is allowed for, 60% of the variance is now permanent (explained by unobserved heterogeneity), showing again the importance of not constraining too much the support of the distribution of z_1.

However, despite a long horizon and a relatively low discount rate, heterogeneity still explains a relatively small fraction of the intertemporal log-earnings variance (60% *versus* 50% in the case of a cross-section of log-earnings). Without being non-stationary, the estimated process thus shows a high level of persistence. The "residual transitory" component rapidly wears out, as illustrated by Figure 1.9. However, since we allow for a non-zero discount rate, this effect never vanishes completely and eventually levels off to 7%. When increasing the time horizon, the "persistent" component of the intertemporal variance rapidly substitutes for this "residual transitory" component of inequality. After ten years, the "persistent" component (the last term of equation (1.5)) amounts to one third of the variance.

	1990	1993	1995	1997	2000
Homogeneous model (only experience matters)					
Variance	.125	.142	.150	.145	.130
Between cohort	.004	.004	.005	.004	.004
Within cohort/Residual transitory	.021	.022	.023	.022	.020
Within cohort/Persistent	.100	.116	.122	.119	.106
Observed heterogeneity model					
Variance	.148	.166	.175	.172	.155
Between cohort	.006	.006	.006	.006	.004
Within cohort/Between education	.059	.063	.066	.062	.058
Within z_0/Residual transitory	.016	.017	.017	.017	.015
Within z_0/Persistent	.067	.080	.086	.087	.078
Unobserved heterogeneity model, $K_1 = 2$					
Variance	.156	.173	.182	.174	.156
Between cohort	.006	.006	.007	.007	.005
Within cohort/ Between z_1	.068	.066	.072	.072	.066
Within cohort/Within z_1/Residual transitory	.013	.016	.016	.015	.014
Within cohort/ Within z_1/Persistent	.069	.085	.087	.080	.071
Unobserved heterogeneity model, $K_1 = 3$					
Variance	.178	.220	.226	.219	.185
Between cohort	.007	.006	.008	.008	.005
Within cohort/Between z_1	.101	.136	.135	.133	.111
Within cohort/Within z_1/Residual transitory	.012	.012	.014	.013	.013
Within cohort/ Within z_1/Persistent	.058	.067	.069	.065	.056

Table 1.9: Variance decomposition of intertemporal log-earnings over a 10-year horizon

1.5.3 Counterfactual analysis

To complete our understanding of the influence of unobserved heterogeneity on longitudinal inequality, we simulate individual trajectories under various heterogeneity

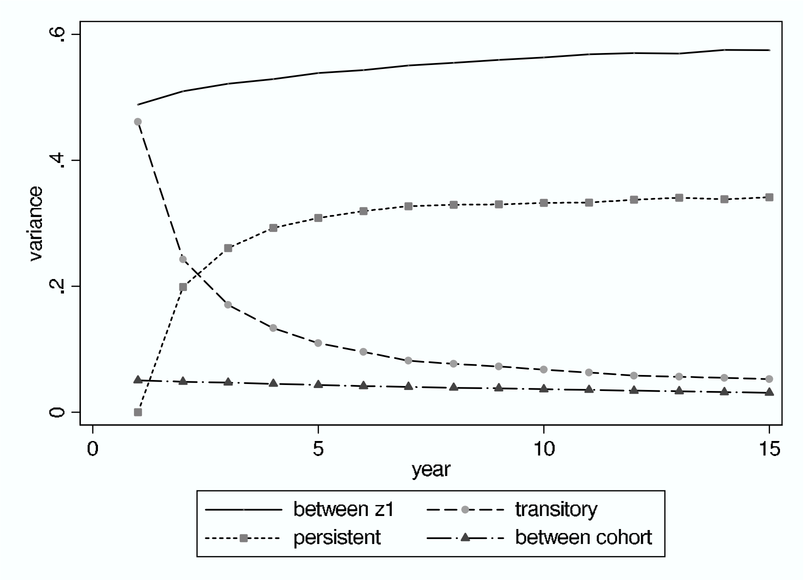


Figure 1.9: Intertemporal log-earnings variance decomposition (unobserved heterogeneity, $K_1 = 3$)

assumptions. The different panels in Table 1.10 correspond to different horizons (from one to fifteen years). In each panel, the first row displays the intertemporal log-earnings variances obtained for the homogeneous model (the only source of heterogeneity is experience). The three other rows display the variances resulting from three counterfactual simulations: first, we allow for unobserved heterogeneity (with $K_1 = 3, K_2 = 2$) in the employment-unemployment process, then, in the marginal distribution of employees' earnings and, lastly, in relative earnings mobility.

We find that unobserved heterogeneity increases earnings inequality essentially through marginal earnings distributions. This is especially true for short horizons. For example, over one year, using the parameters estimated from the 1995-97 panel instead, we find that the difference between the variance obtained from the homogeneous model and the one obtained in the model with unobserved heterogeneity (20%) is almost entirely due to heterogeneous marginal earnings distributions. Over

15 years, the overall variance increase due to unobserved heterogeneity is of 60%. This overall effect is decomposed as follows. Allowing for unobserved heterogeneity in the employment-unemployment process increases longitudinal inequality by 13%. On the top of that, adding unobserved heterogeneity in marginal earnings distributions increases inequality by an additional 35% (60% of overall), and relative mobility yields an increment of 12%.

	1990	1993	1995	1997	2000
			1 year		
Homogeneous	.194	.214	.223	.222	.199
+Unemployment	.196	.214	.224	.224	.200
+Cross-section	.222	.264	.269	.267	.231
+Mobility	.222	.264	.269	.267	.231
			5 years		
Homogeneous	.150	.170	.176	.171	.155
+Unemployment	.153	.173	.182	.179	.161
+Cross-section	.188	.224	.230	.227	.195
+Mobility	.191	.234	.240	.231	.198
			10 years		
Homogeneous	.125	.142	.150	.145	.130
+Unemployment	.132	.149	.163	.156	.138
+Cross-section	.174	.207	.211	.212	.180
+Mobility	.178	.220	.226	.219	.185
			15 years		
Homogeneous	.111	.126	.136	.131	.116
+Unemployment	.123	.136	.153	.144	.126
+Cross-section	.166	.197	.200	.206	.172
+Mobility	.172	.212	.216	.217	.180

Table 1.10: Counterfactual effect of heterogeneity on longitudinal inequality

Lastly, in order to determine which component of the various parameters is instrumental in determining inequality changes over time, we perform a series of dynamic counterfactuals, simulating the model with unobserved heterogeneity ($K_1 = 3, K_2 = 2$) for different dates, imposing some of the parameters to be equal to their 1990-92 value.

In Table 1.11, each panel corresponds to a different horizon. In each panel, the first row shows the variance of intertemporal log-earnings when individual trajectories are simulated using the actual distribution of z_0 for each year and with all other parameters fixed at their 1990-92 value. Remember that the distribution of unobserved heterogeneity (z_1, z_2) conditional on z_0 was constrained, in the estimation, to be the same throughout the decade. The only composition effects are thus those embodied in changes in the distribution of education and age in the population. The second row uses the correct parameters of the employment-unemployment process. The third row adds the correct parameters of cross-section earnings distributions. The fourth row corresponds to a simulation using all estimated parameters in every year.

For all horizons, we detect very little composition effects. The other effects are of small size and are thus difficult to measure precisely. Yet, marginal earnings distributions seem to be driving the main part of time changes. For example, over one year, progressively substituting the parameters estimated with the 1995-97 panel for those estimated with the 1990-92 panel, we find that changing the employment-unemployment process increases longitudinal log-earnings inequality from .226 (after accounting for composition effects) to .238, i.e. by 5%. On the top of that, changing marginal earnings distributions increases inequality by an additional 14%, and relative mobility contributes to virtually nothing. Overall, the 1995-97 parameters augment the variance by 19%. Over 15 years, we find, respectively, 5%, 8% and 4% and an overall effect of 17%.

	1990	1993	1995	1997	2000
	1 year				
Composition	.222	.224	.226	.230	.228
+Unemployment	.222	.236	.238	.245	.228
+Cross-section	.222	.264	.269	.267	.231
+Mobility	.222	.264	.269	.267	.231
	5 years				
Composition	.191	.195	.191	.198	.193
+Unemployment	.191	.202	.196	.206	.194
+Cross-section	.191	.221	.237	.231	.201
+Mobility	.191	.234	.240	.231	.198
	10 years				
Composition	.178	.179	.177	.183	.177
+Unemployment	.178	.184	.186	.192	.178
+Cross-section	.178	.211	.218	.215	.187
+Mobility	.178	.220	.226	.219	.185
	15 years				
Composition	.172	.172	.170	.176	.171
+Unemployment	.172	.174	.178	.185	.171
+Cross-section	.172	.204	.209	.209	.182
+Mobility	.172	.212	.216	.217	.180

Table 1.11: Counterfactual effect of structural change in the model parameters - Model with unobserved heterogeneity ($K_1 = 3$)

1.6 Conclusion

In this chapter, we develop a new methodology to analyze earnings inequality in a dynamic context. We construct a model of earnings dynamics with unobserved heterogeneity which is consistent with the literature on earnings mobility as we model the dynamics of individuals' positions or ranks within cross-section distributions instead of the dynamics of earnings levels. To make this approach tractable when only short panel data are available to the researcher and in the presence of unobserved heterogeneity, we use the statistical tool of copula. Since we model unobserved heterogeneity in a discrete way, the EM algorithm becomes a natural device for estimation. We estimate the model using this technique on data drawn from the French Labour Force Survey. The model is found to fit the data very well.

We use our model to study earnings and employment inequality in France over the 1990-2002 period. We find rather small changes in earnings inequality over that period. Earnings inequality, unemployment risk and immobility seem to increase in business cycle busts and to decrease in booms. These three risks together generate a 25% rise in the variance of intertemporal log-earnings between 1990 and 1995, compared to a 10% rise in the cross-section variance.

We also show how to decompose earnings inequality into separate permanent, transitory and persistent components. The variance of intertemporal log-earnings remains a very significant share of the variance of one-year log-earnings, whatever the horizon over which they are computed (80% in the limit). Moreover, about 60% of the variance of intertemporal log-earnings is due to individual heterogeneity (permanent inequality) and 33% is persistent, leaving a mere 7% to residual transitory variance. Counterfactual analysis allows to dissect the role of heterogeneity in determining longitudinal log-earnings inequality. By comparing a virtual economy with no other individual heterogeneity but age differences to different economies with various degrees of heterogeneity, we find that heterogeneity in cross-sectional log-earnings distributions accounts for 60% of the computed rise in longitudinal inequality and that the remaining 40% are evenly split between unemployment risk and relative earnings mobility.

Copulas thus seem to play well the role we assigned them, i.e. reducing the dimension of the vector of parameters necessary to fit earnings dynamics. It remains to evaluate if they outperform or not the classical linear ARIMA models. The data we use do not permit the comparison and we therefore leave this evaluation to further study. But it would be useful to know, in particular, whether the MA(1) error that is usually detected in autoregressive models of log-earnings levels does not come from misspecifying the nonlinear transformation of earnings (namely, the logarithm) or whether it really is in the data. In the absence of clear empirical evidence for linearity, copula models offer an appealing alternative.

Chapter 2

The Pervasive Absence of Compensating Differentials

2.1 Introduction

In theory, non wage job characteristics (*e.g.* type of work, working conditions, job security) are potential determinants of wage dispersion and labor market turnover (see Rosen, 1986). However, these insights lack empirical support as their confrontation with the data has not led to clear-cut conclusions. A prominent example is the mixed and sometimes conflicting estimates of Marginal Willingness to Pay (MWP hereafter) for these amenities. The first purpose of this chapter is to estimate workers' MWP for job attributes by an original method based on job-to-job transitions. Our second goal is to provide an explanation for the conflicting results found in the literature on compensating wage differentials.[1]

In a perfectly competitive labor market, there must exist positive wage differentials for disamenities (Smith, 1976). The literature on hedonic models, initiated by Rosen (1974, 1986), provides a relevant theoretical framework for the analysis of these compensating differentials. In these models, perfect competition implies that workers' preferences for job attributes translate one-to-one into wage differences. Thus, MWP can be estimated by cross-sectional hedonic wage regressions, as Rosen (1974) proposes.[2] However, this method has not yielded strong empirical evidence of compensating differentials. Typical estimates in this literature, starting with Thaler and Rosen (1975), are of small order of magnitude, often less than five percent of the

[1] We should here mention that these issues are not specific to labor economics. For instance, estimating the willingness-to pay for environmental amenities (such as air quality) has also motivated many articles in urban economics (*e.g.* Roback, 1982, or more recently Chay and Greenstone, 2005). In the conclusion, we shortly discuss the parallel between some recent developments in this literature (Bayer et al., 2005) and our approach.

[2] Rosen's complete proposal involves a two-step method, in which workers' preferences and firms' technology are estimated in a second step. Full identification of general equilibrium hedonic models under perfect competition faces many obstacles (Brown and Rosen, 1982). This task has been recently tackled by Ekeland et al. (2002, 2004) and Heckman et al. (2005). In this chapter, we do not pursue full identification of the model as we are mainly interested in workers' MWP for amenities.

wage, if not insignificantly different from zero or wrong-signed. Correcting for endogeneity and heterogeneity biases using cross-sectional (Goddeeris, 1988, Kostiuk, 1990, Daniel and Sofer, 1998) or panel data (Brown, 1980), several studies have found compensating differentials in some specific cases.[3] Still, the general picture of the literature based on hedonic regressions is not one of systematic evidence of wage/amenity compensation.

Labor market frictions can provide an explanation for the rather inconclusive findings of hedonic wage regressions. If searching for job offers is costly and subject to incomplete information, hedonic prices and workers' MWP need not coincide. Therefore, low wage/amenity correlations must not be interpreted as reflecting weak preferences for job attributes. Hwang *et al.* (1998) emphasize this insight in the context of an on-the-job search model with heterogeneous firms. Lang and Majumdar (2004) reach the same conclusion within a non sequential search framework where firms and workers are homogeneous. This concern has triggered a new empirical approach to estimate workers' MWP, which has led to strikingly different results. In an innovative study, Gronberg and Reed (1994) derive a simple relation between workers' MWP and job hazard rates. Using job duration data, they estimate high and significant MWP for two non wage attributes– measuring several aspects of working conditions –out of four.[4] Subsequently, Van Ommeren et al. (2000) and Dale-Olsen (2005) obtain MWP estimates ranging around one third of the wage, for commuting and safety respectively. Moreover, hedonic regression estimates based on the same data yield a much smaller wage/amenity correlation, in line with previous evidence.

[3]For instance Goddeeris (1988) looks at lawyers' choice to work in the public or private sector, Kostiuk (1990) studies workers' compensation for working on a night shift, and Daniel and Sofer (1998) focus on several job attributes for unionized and non unionized workers.

[4]An early contribution by Herzog and Schlottmann (1990) makes use of workers' mobility between industries to estimate workers' MWP for safety.

Although search frictions are a likely explanation for the contrast between cross-section and job duration results, there has been no attempt to analyze these differences in a single framework. We here intend to reconcile the pervasive absence of compensating differentials in cross section with high MWP for amenities, as estimated from workers' job mobility decisions.

To do so, we write and estimate a dynamic model of wages, amenities and individual labor market transitions. Job-to-job transitions and their wage and amenity outcomes play a key role in our analysis. The main ingredient is the modelling of workers' voluntary decisions to move to a new job or stay in their current job. Individuals value jobs according to their wages and non wage characteristics. Consequently, when taking their job change decisions, they trade off wage and amenity offers according to their MWP for the various amenities. This behavior tends to create a negative wage/amenity correlation posterior to job change.

However, workers' mobility decisions are subject to partly stochastic transition costs, that we model as a random term in the job change equation. This term aims at capturing differences in "true" mobility costs among workers as well as heterogeneity in opportunities to change job and search behavior. Therefore in the model workers cannot move "freely" across jobs. We show that this feature implies that workers' MWP for amenities do not necessarily translate into compensating wage differences posterior to job change. The link between workers' preferences and the observed wage/amenity correlation is made explicit in a structural relation which emphasizes the role of heterogeneity in mobility costs. This relation also involves the correlation between wage and amenity offers, generated at the firm's level. In the model, this demand-side effect is exogenous. Estimating a general equilibrium model is out of the scope of this chapter.

The model thus rationalizes the evidence found in the literature. Even if workers value job characteristics significantly (if MWP are high), there can be no compensating wage differences between jobs with distinct amenities if mobility costs are highly heterogeneous. Then, if the demand-side effect reflecting wage/amenity correlation in job offers is not sufficiently negative, regressing wages on amenities for a sample of job changers will yield a correlation very far from the true MWP.

The identification and estimation of the model's parameters is challenging. First, we need mobility decisions to be as little constrained as possible, so that they reveal individual preferences. Second, as rejected job offers are not observed, we have to deal with selection issues. We use different types of labor market transitions to treat these two problems. Our data allow us to isolate transitions to a "better or more suitable job", which we assume to be voluntary and to reveal individual preferences.[5] Then, to identify the wage/amenity offer distribution, we use one or several types of constrained transitions. Doing so, we adopt a method inspired from the treatment effects literature, where the endogeneity of a "treatment" (here voluntary mobility) is corrected for by using suitable control groups. Our approach depends on the exogeneity of constrained mobility. As our data do not permit to precisely discriminate among constrained transitions, we address the concern that the exogeneity assumption might be violated. In the absence of a convincing instrument for job mobility, we proceed to several changes in the definition of constrained mobility and check the robustness of the results with respect to these variations.

We complete the picture by incorporating two sources of unobserved heterogene-

[5] In a recent contribution, Villanueva (2005) uses a similar definition of voluntary mobility to derive bounds on the "market price" of non wage attributes, correcting for the endogeneity of voluntary transitions. The main difference with our work lies in the considering of workers' selection into jobs. While Villanueva treats selection as a source of bias, we claim workers' mobility choices to be informative on their preferences.

ity. Past literature has argued that if workers are heterogeneous with respect to their productive characteristics, then hedonic regression coefficients can be strongly biased (Hwang *et al.*, 1992). We make use of the panel dimension of our data and extend Brown's (1980) approach in order to control for job, rather than individual, fixed effects. We model the wage process as depending on unobserved ("productive") job specific characteristics. We also incorporate a job-specific effect common to all amenities and independent of the wage. This second heterogeneity is motivated by the nature of the amenity variables we use, which consist of self-reported measures of satisfaction with several dimensions of the job. In line with Duncan and Holmlund (1983), we think that such indicators may suffer from substantial biases, of a more "subjective" nature.

We then estimate the model on European data from the European Community Household Panel (ECHP) for the years 1994-2001. We study nine countries (Austria, Denmark, Spain, Finland, France, Ireland, Italy, the Netherlands and Portugal) and allow for five amenities simultaneously. The empirical results neatly illustrate our discussion of the relationships between wages, amenities and job mobility. We find positive MWP for most job attributes in all countries. Our estimates are systematically significant for at least two amenities: type of work and job security. The MWP for these characteristics even range around one third of the wage. However, both the wage and amenities account for a small share of the variance of mobility decisions. We interpret this result as evidence of heterogeneity in mobility costs. When combined with the MWP, we find very small wage differentials posterior to job change. These findings are robust to changes in the model's assumptions.

The outline of the chapter is as follows: we first present our data and compute several descriptive statistics in Section 2.2. In Section 2.3, we present the model and emphasize the link between MWP and hedonic prices. Identification and estimation are discussed in Section 3.2 while Section 2.5 is devoted to estimation results and robustness checks. Lastly, Section 2.6 concludes.

2.2 Job mobility, wages and amenities: First empirical evidence

In this section, we conduct a simple descriptive analysis of a multi-country sample of individual transitions on the labor market. This allows us to emphasize a number of salient facts about workers' mobility, wages and non wage job characteristics that will motivate our study. First, we present the data and describe the specific variables we will use for the analysis of job mobility and amenities.

2.2.1 The ECHP

We use the European Community Household Panel (ECHP). The ECHP is a panel of ex-ante homogenized individual data covering 15 countries from 1994 to 2001. Each household is interviewed once a year and every individual present in the initial sample is followed over the eight waves. Each observation consists of a rich set of individual characteristics, such as age and gender, together with standard information on the present job: wage, date of start, etc... Two groups of variables are especially relevant to our analysis: the nature of job-to-job transitions, and satisfaction variables with various non wage characteristics.

Classifying job mobility: Our approach uses job-to-job transitions to identify workers' preferences. As emphasized in the introduction, constrained job change may imperfectly reveal these preferences. To discriminate between various degrees of mobility constraints, we use a variable which presents the reason why the individual has stopped working in her previous job. The twelve possible answers are the following:

1	**obtained better/ more suitable job**	7	looking after old, sick, disabled persons
2	obliged to stop by employer	8	partner's job required move to another place
3	end of contract/ temporary job	9	study, national service
4	sale/ closure of own or family business	10	own illness or disability
5	marriage	11	wanted to retire or live off private means
6	child birth/ need to look after children	12	other

Every answer, except 2, 3 and 4, could be thought of as a voluntary quit since the worker has not been laid off. However we consider answers 5 to 12 (when job-to-job mobility is caused *e.g.* by a marriage or the birth of a child) as a sort of constrained mobility which may not reveal the individual's preferences over jobs. In this chapter, we define voluntary mobility as the transitions from one job to a "better or more suitable" one (answer 1). All the other transitions (answers 2 to 12) are constrained.

It is apparent from the twelve answers that a constrained transition can correspond to very different economic events. We define displacements to be transitions corresponding to answers 2 to 4, and cluster all the other answers (5 to 12) into the category of partially constrained mobility. Athough imperfect, the disaggregation of constrained transitions into these two subcategories will allow us to test the robustness of our results.

Amenities: Among the numerous job characteristics available in the ECHP is a set of job amenities. These variables give the subjective valuation of the worker with a given aspect of her job. The typical question is:

How satisfied are you with your present job in terms of (amenity)*?*

and individuals use a scale from 1 (*"not satisfied at all"*) to 6 (*"fully satisfied"*) to indicate their degree of satisfaction. The question remains the same for the following job characteristics:

- TY	:	type of work
- CD	:	working conditions
- WT	:	working times
- DI	:	distance to job/ commuting
- SE	:	job security

For the analysis to be clearer and the estimation to be more tractable, we will cluster the answers into two levels of satisfaction: an amenity equal to 1 (answers 5 and 6) will mean that individuals are actually satisfied and 0 (answers 1 to 4) that they are either unsatisfied or neutral. This clustering is consistent with the literature following Rosen (1986) where amenities take two values: zero for "bad" jobs, one for "good" jobs.[6]

The interpretation of these subjective variables calls for prudence. In the next section, we model each satisfaction variable as a noisy measure of the true amenity on the job. More precisely, we assume that a worker claims to be satisfied with an amenity if the true amenity exceeds a job-specific threshold. We let this threshold depend on observable and unobservable characteristics in order to account for the subjective nature of the data. This approach is close to Clark and Oswald (1996), where workers' satisfaction with their job results from the comparison of their wage with a reference, or "comparison", level. However, contrary to Clark and Oswald (1996), our goal is not to model overall job satisfaction, but satisfaction with specific dimensions of the job.

Even though the ECHP is an ex-ante harmonized panel, some variables (especially amenities) may not be available in every wave and/or country. In particular, the survey only lasts three years in Germany and the United Kingdom. Therefore we

[6]It is common practice in the analysis of subjective data to estimate ordered models, such as ordered PROBIT (see Senick, 2003, and the references therein). Still these methods often involve the arbitrary clustering of some categories (typically the lowest levels of satisfaction). We also estimated our model for "good" amenities corresponding to levels 4, 5 and 6. The results remained qualitatively similar.

restrict our analysis to countries where amenities are available and rarely missing (the non-response rate is less than 1%). In this version of the chapter, we focus on Austria, (AUS), Denmark (DNK), Spain (ESP), Finland (FIN), France (FRA), Ireland (IRL), Italy (ITA), the Netherlands (NLD) and Portugal (PRT). These nine countries cover the scope of the different mobility patterns one can encounter in Europe, from rather static labor markets in some Latin countries (France, Italy and Portugal) to markets with a high turnover (Denmark).

Individual Characteristics: We lastly present the individual characteristics we use in the subsequent analysis: "age" and "age^2" are continuous variables; "male" is a gender dummy, which equals 1 for men; "married" indicates whether the individual is married (= 1) or not; and "kid" equals 1 if the individual has children under 12. Finally, "education" is a variable taking three values, from 1 (less than second stage of secondary education) to 3 (third level education).

2.2.2 Sample description

We merge every two consecutive waves of the ECHP and append the seven resulting tables in order to have a sample containing an *ex-ante* and an *ex-post* situation (respectively denoted as t and $t+1$) for every individual/year in the survey. Thus, a worker present in the eight waves is associated with seven observations, each observation containing her job status (employment, wage,[7] amenities, etc...) and individual characteristics (age, marital status, etc...) both at date t and $t+1$. Therefore an individual appears up to seven times in the data and for each observation she can experience one of the following transitions:

[7] We use the logarithm of monthly wages detrended on year dummies.

- stay employed in the same job
- stay non employed
- make an non employment-to-job transition
- make a job-to-non employment transition
- make a constrained job-to-job transition
- make a voluntary job-to-job transition

Since we do not focus on labor participation, we cluster unemployment and inactivity. Moreover we define employment as paid jobs that last more than 15 hours per week.[8] We explain in Appendix B.1 how we constructed the samples (one per country).

Table 1 shows the main descriptive statistics on our samples. The first two rows present the number of individuals and the number of actual *ex-ante/ex-post* observations. The next six rows give the proportions of each type of transition (except the ones from non employment to non employment). We note that individuals tend to stay in their job. The corresponding probability ranges around two thirds in all countries, which leads to an average job duration of three years. Yet, there is more dispersion across countries in the probability of making a job-to-job transition, which ranges between four and ten percent of total transitions. In particular, voluntary job-to-job mobility is significantly more frequent in Denmark (4.2%) than in Italy (1.3%). In all cases, though, these amount to a small proportion of transitions on the labor market.

The last three rows of Table 1 are important motivations for our analysis. We can see that most voluntary job changes are associated with a wage gain whereas job stayers and constrained job movers more frequently experience a wage cut. This suggests that the wage influences job change decisions. Yet, the proportion of wage increases ranges from only 60% (in France) to 73% (in the Netherlands) of *voluntary*

[8]Self-employed people are likely to differ from other workers in many ways. In particular, lower risk aversion can cause much different career profiles. In this chapter we assume away this issue, and drop the self-employed from our samples.

job-to-job transitions. Up to 40% of voluntary job movers experience a wage cut even if the new job is said to be "better or more suitable" than the previous one. If at least part of these transitions with wage cuts are not spuriously generated by measurement error, these statistics suggest that the wage is not the only characteristic workers value, and we should look at other job characteristics to explain voluntary mobility.[9,10]

Table 1: Sample description

	AUS	DNK	ESP	FIN	FRA	IRL	ITA	NLD	PRT
individuals	4 100	4 010	7 531	4 430	7 513	3 760	7 799	6 492	6 124
observations	18 455	20 025	37 683	16 786	35 571	16 127	42 527	31 892	32 877
Transitions : in % of all obs.									
- non emp.-to-job	959	1 306	3 798	1 511	2 464	1 565	2 713	2 053	2 211
	5.2%	6.5%	10.1%	9.0%	6.9%	9.7%	6.4%	6.4%	6.7%
- job-to-non emp.	1 002	1 133	2 685	1 083	2 214	959	2 289	1 321	1 560
	5.4%	5.7%	7.1%	6.5%	6.2%	5.9%	5.4%	4.1%	4.7%
- stay in same job	13 382	13 070	20 349	11 028	24 233	9 683	27 889	21 534	22 624
	72.5%	65.3%	54.0%	65.7%	68.1%	60.0%	65.6%	67.5%	68.8%
- job-to-job	933	2 058	2 677	962	1 429	1 140	1 558	2 296	1 646
	5.1%	10.3%	7.1%	5.7%	4.0%	7.1%	3.7%	7.2%	5.0%
- voluntary j-t-j	436	849	754	419	603	544	541	988	672
	2.4%	4.2%	2.0%	2.5%	1.7%	3.4%	1.3%	3.1%	2.0%
- constrained j-t-j	497	1209	1 923	543	826	596	1 017	1308	974
	2.7%	6.0%	5.1%	3.2%	2.3%	3.7%	2.4%	4.1%	3.0%
% of wage increases among:									
- vol. j-t-j	63.3%	61.2%	64.5%	69.7%	60.5%	68.8%	60.4%	73.3%	66.1%
- constr. j-t-j	54.2%	51.4%	54.6%	54.3%	53.2%	53.9%	47.1%	64.8%	51.6%
- stay in same job	52.1%	47.2%	53.5%	52.2%	54.5%	52.1%	48.2%	59.7%	48.1%

To investigate further this issue, we report in Table 2 the conditional transition probabilities for the various amenities and types of transitions (voluntary, constrained or within-job). To save space, we show the probabilities for Denmark only, the results being qualitatively similar in all countries. For every amenity and transition type,

[9] Changes in hours worked provide a possible explanation for voluntary quits associated with wage cuts. However, less than 10% of these transitions correspond to changes from full-time (defined as more than 30 hours per week) to part-time work in the three countries we consider.

[10] This chapter focuses on non wage characteristics. Wage growth expectations are an alternative explanation for job transitions associated with wage cuts (*e.g.* Postel-Vinay and Robin, 2002).

the number on the left ($\mathbb{P}(1|0)$) is the probability that a transition starting with a low level of satisfaction with the amenity is associated with an increase in satisfaction. Conversely, the number on the right ($\mathbb{P}(0|1)$) is the probability that a transition starting with a high level of satisaction goes with a fall in satisfaction. As seen in Table 2, $\mathbb{P}(1|0)$ is higher for voluntary job changers than for constrained ones while the converse is true for $\mathbb{P}(0|1)$ (although the difference between the two probabilities is smaller). This suggests that voluntary job change is associated with greater average gains in satisfaction than constrained mobility. Job stayers show a marked contrast with job changers, as they present both fewer increases and fewer decreases in satisfaction with the various amenities.

Table 2: Conditional probabilities $\left(\ \mathbb{P}(1|0) \quad \mathbb{P}(0|1)\ \right)$ in Denmark

Amenities	TY	CD	WT	DI	SE
vol j-t-j	(.68 .15)	(.61 .27)	(.63 .24)	(.50 .32)	(.54 .25)
constr. j-t-j	(.57 .29)	(.51 .29)	(.57 .24)	(.50 .32)	(.43 .32)
stay in job	(.37 .18)	(.39 .21)	(.42 .13)	(.25 .11)	(.44 .15)

These few descriptive statistics tend to confirm the idea that the wage is not the only determinant of workers' voluntary mobility and that non wage characteristics are likely to enter job valuation. We now proceed to a formal test of this intuition.

2.3 A model of wages, amenities, and job mobility

This section is divided into two parts. First, we develop a model of individual transitions on the labor market where workers' Marginal Willingness to Pay for job characteristics can influence their job change decisions. Then we focus on the part of the model which links workers' MWP to the correlation between wages and amenities posterior to job change, emphasizing the key role played by mobility costs.

2.3.1 The model

Voluntary job mobility is the keystone of the model. Still, we also allow for transitions into and out of non employment, together with constrained job changes. In the next section, we shall argue that these transitions can be informative to deal with the endogeneity of voluntary job mobility. We here present the model with a single amenity, for clarity. The extension to multiple amenities, which we use for estimation, is given in the Appendix. Also, for reading convenience, we summarize in Figure 1 the sequence of shocks faced by workers between two consecutive periods, which we now present in detail.

In the model, every match between a worker and a job is described by a pair $\theta = (\theta_1, \theta_2)$, where θ_1 is a productive characteristic and θ_2 is a non productive aspect of the match, standing for the worker's subjective satisfaction with her job. Vector x represents individual characteristics.

Wages and amenities within jobs: Let a worker be employed at date $t = 0, 1, ...$, and let θ be the characteristics of the worker/job match. We assume that the match has already lasted for at least one period. The special case of starting jobs will be addressed at the end of this subsection together with the realization of new matches' characteristics. At the beginning of period t, the wage y_t and the binary non wage characteristics $a_t \in \{0, 1\}$ of the job are drawn, given θ and x. The amenity is drawn according to:

$$a_t = \mathbf{1}\{\alpha_a x + \beta_{1a}\theta_1 + \beta_{2a}\theta_2 + u_{at} > 0\}. \tag{2.1}$$

Simultaneously, the (logarithm of the) wage y_t follows from the hedonic equation:

$$y_t = \rho a_t + \alpha_y x + \beta_y \theta_1 + u_{yt}. \tag{2.2}$$

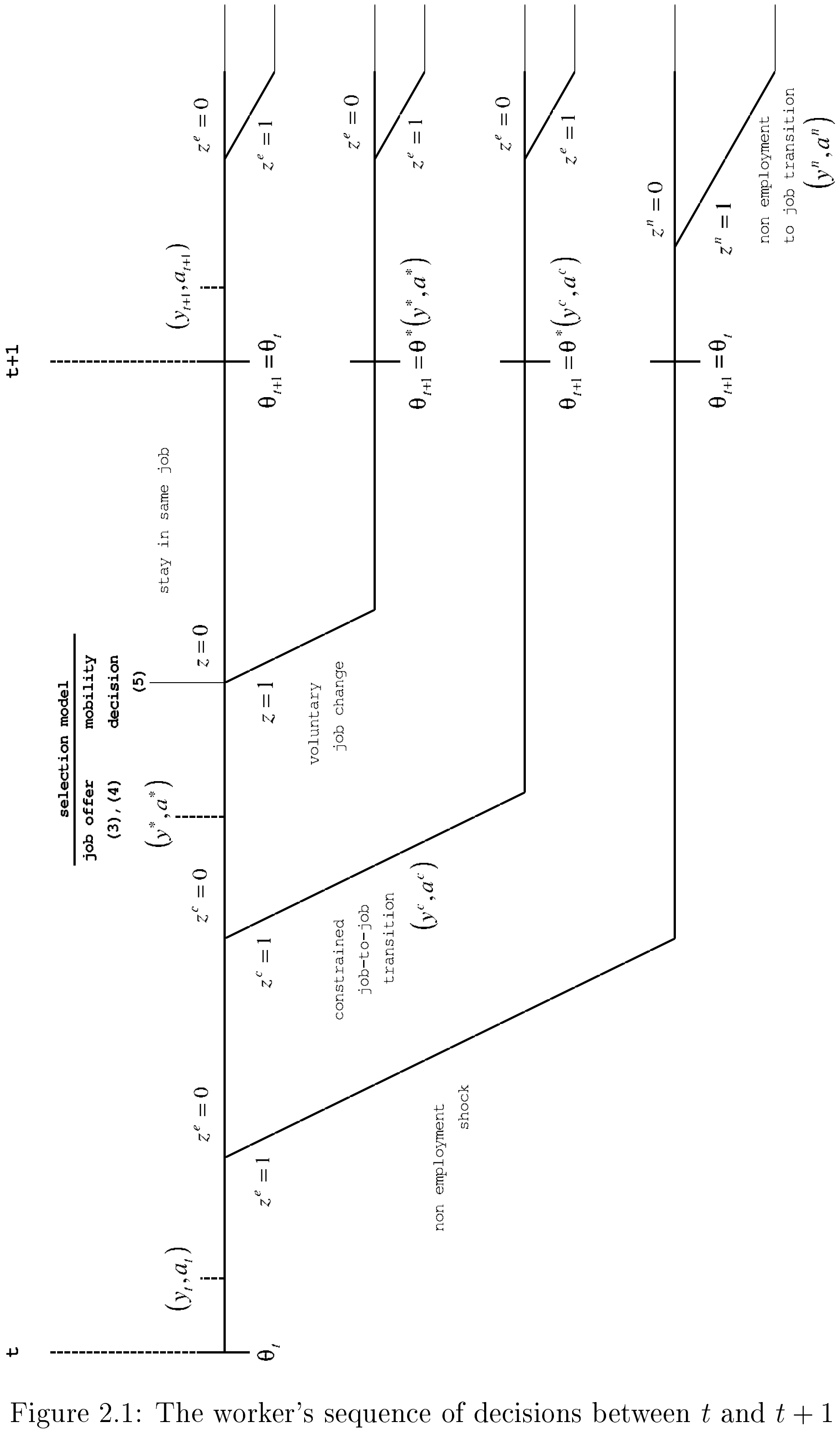


Figure 2.1: The worker's sequence of decisions between t and $t+1$

It is convenient to think of equation (2.1) as a reduced form. The "objective" attributes of a given job are supposed to depend on common characteristics (θ_1, x), and transitory shocks (u_{yt}, u_{at}). Then, amenities are evaluated by the worker in terms of satisfaction, by comparison with a subjective threshold above (resp. under) which $a_t = 1$ (resp. $a_t = 0$). This threshold depends on x, θ_1, and on the second (non productive) component of the quality of the worker/job match θ_2. Then, equation (2.1) mixes the "objective" amenity and the "subjective" interpretation of the amenity in terms of satisfaction. Moreover, as shocks u_{at} are transitory, amenities are i.i.d. within jobs.

Then, parameter ρ in the hedonic equation (2.2) is the implicit price of the amenity. Equations (2.1)-(2.2) can follow from the negotiation of the employment contract between the worker and the firm, given the characteristics of the match. We do not model this process and thus treat the parameters in (2.1)-(2.2) as exogenous.

Once drawn, the wage and amenity are valued by the worker, according to the instantaneous indirect utility function $v(y_t, a_t, x)$. Then she can experience several types of transitions.

Adverse shocks: First, the worker can be constrained to move to another job, or forced back to non employment. Both events happen with exogenous probabilities depending on x and θ.

If the worker has experienced a constrained transition, her new wage/amenity pair (y^c, a^c) is drawn from a specific distribution given θ and x, similar to (2.1)-(2.2) but with different parameters indexed by the (^{c}) superscript. These parameters result from the new firm's technological choices, rather than from contract negotiation with the worker. In particular, the compensating differential for the amenity posterior to constrained job change (ρ^c) is *a priori* different from the within-job compensating differential (ρ).

Job offers: If the worker has experienced no adverse shock, then she gets an offer from an outside firm that she can either accept or turn down.[11] Amenity offers a^* are drawn from a specific distribution:

$$a^* = \mathbf{1}\{\alpha_a^* x + \beta_{1a}^* \theta_1 + \beta_{2a}^* \theta_2 + u_a^* > 0\}, \tag{2.3}$$

and wages are given by another hedonic equation:

$$y^* = \rho^* a^* + \alpha_y^* x + \beta_y^* \theta_1 + u_y^*. \tag{2.4}$$

Note that wage and amenity offers depend on the characteristics of the *current* match, as firms are assumed to direct their search on groups of workers with given individual (x) and match (θ) characteristics.

Equation (2.4) reflects the trade-off between wages and (possibly costly) amenities at the firm's level, and ρ^* is the compensating differential for the amenity in job offers. We allow the three hedonic equations– corresponding to continuing jobs, constrained job change and outside offers –not to be the same. Different wage and amenity offers can arise if firms can condition their job offer on the labor market status of the applicant. Then, an employer might not offer the same amenities, say working conditions, to a displaced worker whose alternative is non employment than to a worker who will remain in her job if she rejects the offer. This would be the case in a model where the current employer could make a counter offer to prevent the worker from going to the poaching firm, as in Postel-Vinay and Robin (2002). Our identification strategy will be based on the similarity between outside job offers and offers received

[11] The assumption of a systematic arrival of outside job offers is balanced by our modeling of stochastic mobility costs. We discuss this in more details in the next subsection, after having introduced mobility costs.

by some constrained workers. Letting the offer parameters to differ between types of constrained transitions will provide some flexibility to test the robustness of the results.

Voluntary mobility rule: Let $V(y_t, a_t, \theta, x)$ be the present value, at the beginning of period t, of a job with characteristics θ and (instantaneous) wage/amenity values (y_t, a_t). Future periods are discounted at a constant rate. Note that $V(y_t, a_t, \theta, x)$ takes into account possible future spells of non employment, associated with a constant instantaneous utility.

The worker bases her decision whether or not to change job on the comparison of the job offer she has just received to the expected present value of her current job. The latter is given by: $V_{stay}(\theta, x) = \mathbb{E}_{y_{t+1}, a_{t+1}}[V(y_{t+1}, a_{t+1}, \theta, x)|\theta, x]$. Note that, as shocks u_{at} and u_{yt} in (2.1)-(2.2) are transitory, this expression does not depend on past wage/amenity values $(y_t, a_t, y_{t-1}, a_{t-1}, ...)$.

When the worker takes her decision, the characteristics of the new match are not yet realized. The decision is thus based on the expected value of the proposed job, given wage and amenity offers: $V_{mov}(y^*, a^*, \theta, x) = \mathbb{E}_{\theta^*}[V(y^*, a^*, \theta^*, x)|y^*, a^*, \theta, x]$, where θ^* stands for the characteristics of the offered worker/job match, unknown to the worker at time t.

Let z_t be the dummy variable indicating if the individual has changed job voluntarily between periods t and $t+1$. The mobility decision reads:

$$z_t = \mathbf{1}\{V_{move}(y^*, a^*, \theta, x) > V_{stay}(\theta, x) + c(\theta, x)\}, \tag{2.5}$$

where $c(\theta, x)$ are transition costs.[12] We adopt a linear specification for this relation-

[12]Examples of mobility costs correlated with θ could be housing and children's education expenses. Van den Berg (1992) estimates an on-the-job search model where mobility costs depend on the current wage, constant within job.

ship, and assume:

$$z_t = \mathbf{1}\left\{y^* + \delta^* a^* > \alpha_z x + \beta_{1z}\theta_1 + \beta_{2z}\theta_2 + u_z\right\}. \qquad (2.6)$$

The parameter δ^* in this equation reflects the trade off between wage and amenity in the mobility decision. Precisely, δ^* is the wage differential that has to be paid to make a job offer associated with a low degree of satisfaction with the amenity equivalent to a job offer giving high satisfaction. As, in the model, the value of an offer incorporates all future outcomes, we shall refer to δ^* as the dynamic Marginal Willingness to Pay for the amenity. Note in contrast that, since x and θ intervene in both V_{stay} and V_{move}, the interpretation of their associated parameters in (2.6) calls for prudence.

The Right-Hand Side inside the index in equation (2.6):

$$\tau(x, \theta) = \alpha_z x + \beta_{1z}\theta_1 + \beta_{2z}\theta_2 + u_z, \qquad (2.7)$$

is a combination of the value of staying in the present job and transition costs. In this chapter, we do not intend to disentangle these two dimensions. Moreover, the random shock u_z renders $\tau(\theta; x)$ stochastic conditional on observed and unobserved characteristics. Although many models of transitions assume deterministic reservation wages (see *e.g.* Flinn and Heckman, 1982), our results will demonstrate that, in the context of job-to-job mobility, it is important to allow for variability in τ.

Then, the voluntary job change decision has two potential outcomes. If the worker accepts the offer, she gets $(y_{t+1}, a_{t+1}) = (y^*, a^*)$ at the beginning of the next period $t+1$. If not, she draws a new pair (y_{t+1}, a_{t+1}) in (2.1)-(2.2).

Match characteristics: If the worker has remained in the same job between t and $t+1$, then the characteristics of the match stay the same. Match characteristics are thus constant within-job. Moreover, non employed individuals are assumed to “keep” the θ of their previous match, using θ as a signal (for instance in their vita) when applying for a new job.

If the worker has started to work at a new job at the next period $t+1$ with starting wage and amenities (y_{t+1}, a_{t+1}), then the productive and non productive characteristics of the match are realized. We assume that the new match characteristics θ^* are drawn from a distribution depending on the starting wage/amenity values. Doing so, we intend to capture both a worker and a firm effect on the formation of the match. Indeed, the starting wage and amenities at a job depend on the characteristics of the previous match (θ) and individual characteristics (x), as well as non worker-specific components, represented by the residuals in the starting wage/amenity equations as (2.3)-(2.4). Decomposing further between the worker and the firm effect on the match characteristics would require matched employer/employee data.

So far, the worker has been assumed to be employed in a continuing job at time t. We now focus on the two cases when she is either non employed at time t, or starting to work in a new job.

Non employment-to-job transitions: If the worker is non employed at the beginning of period t, then we assume that she can find a job with exogenous probability depending on the θ of her previous job.[13] Then, at the beginning of the next period, she draws a wage/amenity pair (y^n, a^n) from a distribution similar to (2.1)-(2.2) but with different parameters, in particular another ρ^n (as we discussed in the paragraph on job offers). Subsequently, the characteristics of the new match are realized. Hence non employed individuals draw from different wage/amenity distributions than em-

[13] In the estimation, if an individual is non employed in the first wave (1994), we wait until she gets a job to start modelling her transitions. Hence every observation will be associated to a θ.

ployed ones when finding a job. This can reflect a loss of human capital associated with the period of non employment. However, as individuals keep the θ of their previous jobs, the loss of human capital is supposed not to depend on the length of the non employment spell.

Starting jobs: Lastly, consider a worker employed in a new job at the beginning of period t. Her starting wage and amenity are either equal to (y^*, a^*) (if she has changed job voluntarily at the end of period $t-1$), to (y^c, a^c) (if she has been constrained to change job), or to (y^n, a^n) (if she has was previously non employed). The new match characteristics are then drawn given these starting values, and condition the arrival rates of adverse shocks and the offers she might receive from an outside firm.

Additional specifications: In the econometric model, all variables are indexed by individual i and time period t. Individual characteristics x_{it} can be time-varying (*e.g.* age), or not (sex, education). All residuals in the model are *i.i.d*, independent of covariates, independent of one another and normally distributed with zero means.

All types of transitions– but voluntary ones –are modeled according to a common pattern. For instance, the occurence of constrained job change is specified as:

$$z^c = \mathbf{1}\left\{\alpha_z^c x + \beta_{1z}^c \theta_1 + \beta_{2z}^c \theta_2 + u_z^c > 0\right\}. \tag{2.8}$$

The variance of all residuals in the transition equations, such as u_z^c, are normalized to one. We adopt the same normalization for the residuals in the amenity equations. In contrast, the variance of u_z in the voluntary mobility equation (2.6) can be identified, as the coefficient of wage offers in equation (2.6) is set to one by assumption. The variance of u_z, which we denote as σ_z^2, will play a key role in our analysis.

In the econometric model, characteristics θ_1 and θ_2 are unobserved match-specific effects. Their identification comes from within-job repetitions (see 2.4.1). We model θ_1 and θ_2 as discrete random variables, following the approach of Heckman and Singer (1984).

After a transition out of non employment, a constrained or a voluntary job-to-job transition, new match characteristics are drawn. We model the conditional distributions of these characteristics, given starting wage and amenities at the new job, by Ordered PROBIT specifications. In addition, we assume that the conditional distribution of θ^* given starting wage and amenities is identical for every individual and independent of the type of the last transition. In other words, we assume that starting wage/amenity values, which do depend on the worker's previous labor market status, are sufficient statistics for the match characteristics, and we neglect state dependence beyond these starting conditions.

We adopt a similar approach to model initial conditions. Namely, we assume that all individuals are initially employed, and condition the worker's likelihood on her starting wage and amenity. The match characteristics in the initial job are modeled by Ordered PROBIT specifications conditional on the initial wage and amenities (see Appendix B.2 for details). This choice is motivated by the complex dynamics of the model. As match characteristics are job-specific, the characteristics corresponding to the first observations in the sample will be correlated to individual covariates such as age. Computing the conditional distribution requires integrating over past wages, amenities and match characteristics. We found this approach computationally intractable. Our solution involves conditioning on initial wages and amenities. We also tried an alternative specification, modelling directly the distribution of initial match characteristics on individual covariates such as age and education, with little influence on the results.

2.3.2 Hedonic wage regressions and job mobility

In this subsection, we derive some of the model's implications. In particular we emphasize the key role of mobility costs in the relation between workers' MWP and cross-sectional wage differences between amenity levels.

The wage/amenity offer equations (2.3)-(2.4) and the rule of voluntary mobility (2.6) form the core of the model. The two parameters ρ^* and δ^* represent specific trade-offs between wages and amenities: at the firm's level, ρ^* reflects substitution between two costs while at the worker's level, δ^* represents the trade-off between two goods. When workers take their mobility decisions, both effects are at play. Posterior to job change, these "demand" and "supply" effects can *a priori* strengthen or weaken the correlation between wages and amenities.

One of the main insights of the theory of compensating differentials is that the two trade-offs, at the firm's and the worker's levels, tend to create a negative correlation. In the model, this correlation depends on parameters ρ^* and δ^*, as well as on a third parameter measuring the extent of transition costs. We now derive the structural relation linking these different parameters.

Let us define:

$$\Delta_z = \mathbb{E}(y^*|a^* = 1, z_t = 1, \theta, x) - \mathbb{E}(y^*|a^* = 0, z_t = 1, \theta, x).$$

Δ_z is the wage differential between the two amenity levels, for voluntary job changers. It could be estimated by a hedonic wage regression on the subsample of job changers, controlling for unobserved heterogeneity.

Using the specification of the hedonic curve (2.4), the differential can be written as:

$$\Delta_z = \underbrace{\rho^*}_{\equiv \Delta_z^{(d)}} + \underbrace{\mathbb{E}(u_y^*|a^* = 1, z_t = 1, \theta, x) - \mathbb{E}(u_y^*|a^* = 0, z_t = 1, \theta, x)}_{\equiv \Delta_z^{(s)}}. \qquad (2.9)$$

In equation (2.9), Δ_z has been written as the combination of a "demand" and a "supply" effect. The first effect, ρ^*, is taken as given in the model. Then, the "supply" effect is likely to be negative, consistently with the intuition that workers can accept lower (higher) wage in exchange of better (worse) amenity when moving to another job. To see why, note that the mobility rule (2.6) implies that

$$\Delta_z^{(s)} = \mathbb{E}(u_y^* | u_z - u_y^* < -\mu + \rho^* + \delta^*) - \mathbb{E}(u_y^* | u_z - u_y^* < -\mu),$$

where we have defined for compactness:

$$\mu = (\alpha_z - \alpha_y^*)x - (\beta_y^* - \beta_{1z})\theta_1 + \beta_{2z}\theta_2.$$

As u_y^* and u_z are normally distributed and uncorrelated by assumption, we can write:

$$\Delta_z^{(s)} = \frac{(\sigma_y^*)^2}{\sigma}\left[\frac{\phi}{\Phi}\left(-\frac{\mu - \delta^* - \rho^*}{\sigma}\right) - \frac{\phi}{\Phi}\left(-\frac{\mu}{\sigma}\right)\right], \tag{2.10}$$

where ϕ (respectively Φ) denotes the standard normal pdf (resp. cdf), and $(\sigma_y^*)^2$, σ_z^2 and $\sigma^2 = \left(\sigma_y^*\right)^2 + \sigma_z^2$ are the variances of u_y^*, u_z and $u_y^* - u_z$, respectively.

Then it follows from (2.10) that $\Delta_z^{(s)}$ has the same sign as $-(\delta^* + \rho^*)$.[14] In the benchmark case when there is no correlation in wage/amenity offers ($\rho^* = 0$), the worker's trade-off between wage and amenity leads to a negative wage/amenity correlation posterior to job change. Even in the absence of negative correlation on the demand-side of the market, this supply effect is enough to create "compensating" wage differences between jobs with distinct amenities.

Order of magnitude: In the model, workers are constrained on their mobility, as we assume job-to-job transitions to be costly. This feature can have a strong effect on the wage differential posterior to job change Δ_z, as we now illustrate.

[14]This is because the inverse Mills ratio $\frac{\phi}{\Phi}$ is strictly decreasing on the real line. The normality of u_z and u_y^* is not essential for this result to hold, however.

To proceed, let us suppose that the probability of job change is small, conditional on "good" or "bad" amenity offers, and conditional on individual and job characteristics. This assumption is consistent with the descriptive evidence in 2.2.2, where the aggregate probability of voluntary job change was found to be less than five per cent in the three countries we study. Precisely, for all x, θ and a^*, we suppose:

$$\Phi\left(-\frac{\mu-(\delta^*+\rho^*)a^*}{\sigma}\right) << 1.$$

This assumption permits to approximate $\Delta_z^{(s)}$ as:[15]

$$\Delta_z^{(s)} \approx \frac{(\sigma_y^*)^2}{\sigma}\left[\left(\frac{\mu-\delta^*-\rho^*}{\sigma}\right)-\left(\frac{\mu}{\sigma}\right)\right] = -\left(\frac{\sigma_y^*}{\sigma}\right)^2(\delta^*+\rho^*)\,.$$

We then define:

$$R_{yz} \equiv \left(\frac{\sigma_y^*}{\sigma}\right)^2. \tag{2.11}$$

Hence $\Delta_z^{(s)}$ can be approximated as:

$$\Delta_z^{(s)} \approx -R_{yz}\left(\delta^*+\rho^*\right). \tag{2.12}$$

The key parameter R_{yz} measures the weight of wage offers in mobility decisions. We interpret this ratio as a measure of the heterogeneity in mobility costs. If costs are homogeneous, then the variance of wage offers accounts for a large share of σ^2 and R_{yz} is close to 1. On the contrary, if mobility decisions involve many factors other than the wage, R_{yz} is small. These mobility costs represent restrictions that labor market imperfections can impose on individual mobility. They can consist of three types of costs: true transition costs of moving to another job, costs of searching for jobs, as well as costs of having access to outside offers.

[15]The inverse Mills ratio has the property that $\frac{\phi}{\Phi}(x) = -x + o(x)$ when $x \to -\infty$. Hence, for probability p close to zero, $\frac{\phi}{\Phi}(\Phi^{-1}(p))$ is close to $-\Phi^{-1}(p)$. Note that, unlike in the previous paragraph, the assumption of normally distributed residuals is here critical.

We do not model the probability of receiving an outside offer, since we assume that every worker who has not faced an adverse shock gets an alternative offer. However, the stochastic mobility costs we have introduced make this assumption immaterial. In the model, not having access to an alternative job (*i.e.* receiving no job offer) is interpreted as facing a high mobility cost, *i.e.* drawing a large u_z. It is thus clear that R_{yz} represents also the heterogeneity in opportunities to change job.

Equation (2.12) shows that, the larger the MWP for the amenity, the more negative the correlation posterior to job change. However, in the case where mobility costs are heterogeneous, large MWP for job attributes (*i.e.* large and positive δ^*) can translate into very weak wage/amenity correlation (if R_{yz} is low).

Now if we go back to the wage differential for job changers and still assume a low probability of job change, we can write, combining (2.9) and (2.12):

$$\Delta_z \approx (1 - R_{yz})\,\rho^* + R_{yz}\,(-\delta^*)\,. \tag{2.13}$$

According to (2.13), the wage/amenity correlation posterior to job change combines the two parameters ρ^* and δ^*, in a proportion determined by the heterogeneity in mobility costs R_{yz}. In the limit, when R_{yz} is close to zero, then the worker's trade-off between wages and amenities has no impact on the correlation posterior to job change. Equation (2.13) involves the three key parameters of the model: wage/amenity substitution on the demand (ρ^*) and supply (δ^*) sides, together with an indicator of mobility costs (R_{yz}). Identifying and estimating these parameters is one of our main purposes.

2.4 Identification and estimation issues

In this section, we address the identification and estimation of the model's parameters. We first focus on the identification problem arising from workers' selection into jobs.

We then discuss the identification of unobserved match characteristics. Lastly, we outline the estimation method.

2.4.1 Identification of the key parameters

We here assume that match characteristics are observed by the econometrician. Next subsection will deal with their identification.

If job offers were observed for all workers, not only actual job changers, then the model's key parameters would be identified without further assumption. Indeed ρ^* follows directly from the hedonic curve (2.4). Moreover, all parameters ruling job offer equations would be identified in that case, including σ_y^*, the standard deviation of u_y^*. As for δ^*, it follows from equation (2.6) that:

$$\Phi^{-1}\left[\mathbb{P}\left(z_t = 1 | y^*, a^*, \theta, x\right)\right] = \frac{1}{\sigma_z} y^* + \frac{\delta^*}{\sigma_z} a^* - \frac{\alpha_z}{\sigma_z} x - \frac{\beta_{1z}}{\sigma_z} \theta_1 - \frac{\beta_{2z}}{\sigma_z} \theta_2. \qquad (2.14)$$

Equation (2.14) shows that, along with data on job turnover, data on wage offers would permit to identify all parameters in (2.6), including the standard deviation of u_z, σ_z.

The selection problem: The absence of data on job offers in the ECHP complicates the researcher's task. As wage and amenity offers are not observed for job stayers, we face a selection problem.

Identification of selection models is often achieved by the use of exclusion restrictions. In the case of job-to-job mobility, however, it seems difficult to find a covariate that both significantly influences the job change probability and is uncorrelated with job offers. Potential candidates in the ECHP are the indicators of being married and having young children. However, their effect on job change propensity turns out to be small.

Therefore we here take a different route. Our approach builds on the remark that observing the realizations of job offers (y^*, a^*) is not strictly necessary for the parameters to be identified. For this purpose, knowing their distribution is sufficient. This is straightforward for the parameters appearing in (2.3)-(2.4). For mobility parameters in (2.6), the argument makes use of Bayes' rule. Namely, if the wage/amenity offer distribution is known (given characteristics), with density ℓ^*, then:

$$\mathbb{P}\left(z_t = 1|y^*, a^*, \theta, x\right) = \frac{\ell^z(y^*, a^*|z_t = 1, \theta, x)}{\ell^*(y^*, a^*|\theta, x)} \mathbb{P}\left(z_t = 1|\theta, x\right), \tag{2.15}$$

where ℓ^z denotes the density of wages and amenities accepted by voluntary job changers. As ℓ^z and $\mathbb{P}\left(z_t = 1|\theta, x\right)$ involve observed quantities, the Left-Hand Side in (2.15) is identified as soon as ℓ^* is known. Then, using (2.14), the parameters appearing in the mobility equation can be recovered.

Finding proxies for the distribution of wage/amenity offers: For the model's parameters to be identified, a sufficient condition is thus that the distribution of job offers be itself identified. We propose to augment the model by the following identity:

$$\ell^* = \ell^c, \tag{2.16}$$

where ℓ^c is the density of wages and amenities drawn by constrained job changers.

For (2.16) to be satisfied, two conditions need to hold: First, constrained job change has to be exogenous from the worker's perspective. If this is the case, then ℓ^c can also be interpreted as the density of "offers" received by constrained job changers– offers which cannot be turned down by the worker. Second, constrained and voluntary job changers need to draw from the same distribution of job offers.

The first condition, exogeneity, could be violated for two reasons: workers can self-select with respect to their unobserved characteristics (ability bias) or with respect to the characteristics of the current and offered jobs (selection). We may argue that we control for the first source of endogeneity, as the probability of job change is conditional on θ in equation (2.8). However, we do not control for the endogeneity possibly arising from selection with respect to job offers. As for the second condition, equality of job offers, it could be violated if, say, experiencing a constrained transition were seen by employers as a signal of low productivity. This argument could well hold in the case of lay-offs, as suggested in Gibbons and Katz (1991).

To guarantee the exogeneity of constrained job change, Gibbons and Katz (1992) consider displaced workers consecutive to plant closure. They claim that such displacements approximate the natural experiment of exogenous job loss, and use this insight to correct for ability bias in inter-industry wage differentials. A recent empirical analysis of returns to tenure by Dustmann and Meghir (2005) builds on the same idea. In our case, however, this approach is not directly applicable because the ECHP data are not precise enough to identify "true" displacements, exogenous from workers' perspective. We are not aware of a data set providing information on amenities and voluntary mobility (as in the ECHP) together with "true" displacements (as in firm level data). Still, the ECHP data allow to distinguish between different types of constrained job transitions. This feature permits to develop informal robustness checks.

An informal test of robustness. To address the concern that the above conditions might be violated, we test the robustness of our results to variations in assumption (2.16). Our approach consists in substituting three other wage/amenity distributions for the original distribution of reference (constrained transitions). First, we decompose constrained transitions into displacements and partially-constrained transitions, as explained in 2.2.1. We take the wage/amenity distributions of either of the two types as a reference. For instance, replacing identity (2.16) by: $\ell^* = \ell^d$, where ℓ^d denotes the density of wages and amenities drawn by *displaced* workers, a different set of parameters ruling voluntary job mobility can be derived. We proceed similarly in the case of partially constrained transitions.

Lastly, we also try a third specification, taking wages and amenities posterior to non employment as the reference for job offers. In this case, the identity becomes: $\ell^* = \ell^n$. This restriction has been widely used in the job search literature, as an assumption permitting the job offer distribution to be identified. In these models, a subpopulation of workers (usually unemployed) draw their jobs in the offer distribution and are forced to accept the offer because their alternative is not preferable. Hence the observed distribution of jobs drawn by these workers is the same as the actual offer distribution (see *e.g.* Christensen et al., 2005).

These different types of constrained transitions provide imperfect proxies for job offers. By comparing the results obtained using these different proxies, we expect empirical regularities to emerge. We will see in section 5.3 that the MWP and correlation estimates turn out to be qualitatively similar between specifications.

2.4.2 Identification of match characteristics

In practice, match characteristics are unobserved by the econometrician. However, since θ_1 and θ_2 are constant within-job, they are identified by wage and amenity repetitions provided that jobs last more than two periods with positive probability. For θ_1, the argument comes *e.g.* from a theorem by Kotlarski (1967). In that case, the density of θ_1 is identified nonparametrically. As for θ_2, one also needs that the regressors vary sufficiently over time. For instance, if one of the regressors has large support in all its dimensions as in Manski (1988) then the latent distributions are identified and Kotlarski's result applies.

Properly speaking, we do not dispose of such a regressor in amenity equation (2.1), but unobserved heterogeneity distributions are modeled as discrete random variables, allowing for a parsimonious number of groups. For this specification, we found no evidence of identification problems.

2.4.3 Estimation: EM with a Sequential M-step (ESM)

We here briefly present the estimation of the model's parameters. The details of the procedure are given in Appendix B.2. We model θ_1 and θ_2 as follows. Let N denote the number of individuals in the sample. We assume that there exist two integers K_1 and K_2, a mapping:

$$\begin{aligned} \{1...N\} &\rightarrow \{1...K_1\} \times \{1...K_2\} \\ i &\mapsto (k_{1i}, k_{2i}), \end{aligned}$$

and parameters $(\vartheta_{11}, ...\vartheta_{1K_1})$, $(\vartheta_{21}, ...\vartheta_{2K_2})$ such that $(\theta_{1i}, \theta_{2i}) = (\vartheta_{1k_{1i}}, \vartheta_{2k_{2i}})$.

We use the EM algorithm of Dempster *et al.* (1977) to estimate the model's parameters. This amounts to treating k_{1i} and k_{2i} as random variables. Starting with initial guesses for the parameters, one computes, in the expectation (E) step, the posterior probabilities that $(k_{1i}, k_{2i}) = (k_1, k_2)$ given the data, for all k_j in $\{1, ..., K_j\}$, $j = 1, 2$ and for all individuals. Then in the maximization (M) step one maximizes the likelihood of the observations, weighted by the posterior probabilities.

As for the choice of K_1 and K_2 there is a trade-off between the accuracy of the description of the unobserved heterogeneity distributions and the tractability of the estimation due to the small number of voluntary job-to-job transitions. We found $K_1 = 4$ and $K_2 = 2$ to be a convenient choice for the countries we study. In the empirical analysis, we shall test the robustness of our conclusions to variations in K_1 and K_2.

The estimation of the model takes the form of simple steps. In the M-stage of the algorithm, all parameters– except the ones ruling voluntary job-to-job mobility, given by equations (2.3)-(2.4) and (2.6) –are estimated either by PROBIT, Ordered PROBIT or OLS– weighted by the posterior probabilities. Then, equations (2.3)-(2.4) and (2.6) form a censored regression model with endogenous threshold. Theoretically, one could estimate the model under the constraint (2.16). However, in the empirical analysis, we allow for five amenities and a bivariate distribution of unobserved heterogeneity, so the joint estimation turns out to be impractical.

Instead, within each M-step of the algorithm, we proceed in two steps: First, we estimate the wage/amenity distribution posterior to constrained job change (ℓ^c). Second, we estimate the parameters ruling the voluntary mobility decision setting $\ell^* = \ell^c$. Appendix B.2.2 details the mathematical expression of the second-step likelihood. The resulting algorithm follows the pattern introduced in Arcidiacono and Jones (2003) of EM with a sequential M-step (ESM). Therefore, our method provides consistent estimates of the parameters. However, it is not as efficient as FIML.

To compute asymptotic standard errors, we write the First-Order Conditions of the algorithm as population moment conditions. Arcidiacono and Jones (2003) propose to compute the asymptotic variance-covariance matrix by the usual GMM formula. In our case, we found that second derivatives of the type-conditional likelihoods could be very long to compute. In Appendix B.2.3, we propose a generalization of the information matrix identity that allows to significantly reduce computing time, by up to a factor 100 in our case.

2.5 Estimation results

In this Section, we first present the parameter estimates. We then contrast the estimated MWP with the wage/amenity correlations in cross-section, interpreting the results in light of section 2.3.2. Lastly, we check the robustness of our findings to several changes in the model's specification.

2.5.1 Parameter estimates

As the model presented in 2.3.1 contains many parameters, we here give a partial account of the results, focusing on the parameters of interest. Additional estimates are available from the authors upon request.

Wage and amenity equations: Table 3 presents the parameter estimates of wage/amenity equations (2.1)-(2.2) for Denmark. The results for other countries are qualitatively similar. Recall that all five amenities have been simultaneously included in the estimation. Parameters ρ_1, ..., ρ_5, which represent the within-job compensating differentials for each of the five amenities, are reported separately in Table 4.

Table 3 shows that human capital determinants and unobserved match characteritics have very different effects on the wage and the amenities. The wage is concave in age, and is higher for males and educated workers. In constast, all amenities but distance to job are convex in age, yet no clear pattern arises from the effects of gender and education. In addition, θ_2 has a significant effect on all the amenities, whereas it is independent of the wage by construction (See Appendix B.2 for details). Interestingly, this effect has the same sign and roughly the same magnitude for the five different amenities. This finding could suggest that workers differ in their understanding of the 1 to 6 ranking given in the ECHP. High-θ_2 and low-θ_2 workers could attach different meanings to words such as "very satisfied" or "satisfied", consistently with a "subjective" interpretation of θ_2.

Table 3: Wage and amenities within job (Denmark)

	Wage	TY	CD	WT	DI	SE
	Observed heterogeneity x					
age	.0272 (.00081)	−.00512 (.010)	−.0244 (.0096)	−.0256 (.0098)	.0369 (.011)	−.0726 (.010)
age^2	−.000316 (.00001)	.000116 (.00012)	.000333 (.00011)	.000454 (.00012)	−.000188 (.00013)	.000851 (.00012)
male	.120 (.0028)	.0707 (.027)	.0871 (.026)	.00445 (.027)	−.0958 (.030)	−.0929 (.027)
edu= 2^{nd} level	.07160 (.0035)	.0135 (.037)	.0264 (.036)	.000929 (.036)	−.177 (.041)	.0747 (.036)
edu$\geq 3^{rd}$ level	.146 (.0038)	.0362 (.040)	.0192 (.039)	−.0419 (.039)	−.223 (.044)	.193 (.039)
constant	8.989 (.017)	.348 (.21)	.798 (.20)	1.0379 (.21)	−1.427 (.23)	1.820 (.21)
	Unobserved heterogeneity (θ_1, θ_2)					
$\theta_1 = 1$	−.986 (.0052)	−.319 (.083)	−.522 (.079)	−.651 (.082)	.632 (.086)	−.372 (.078)
$\theta_1 = 2$	−.568 (.0038)	−.603 (.053)	−.823 (.052)	−.613 (.056)	.563 (.055)	−.389 (.051)
$\theta_1 = 3$	−.326 (.0036)	−.263 (.050)	−.456 (.049)	−.359 (.053)	.356 (.050)	−.0517 (.048)
$\theta_2 = 1$	-	1.0190 (.029)	.981 (.028)	.684 (.029)	1.390 (.033)	.825 (.029)

We also proceeded to a variance decomposition in Denmark (not shown here). Observed covariates account for 33% of the wage variance,[16] and unobserved heterogeneity θ_1 accounts for 44%. In contrast, both regressors have little explanatory power in amenity equations. For instance, for the amenity "type of work", both x and θ_1 account for 3% of the variance, while θ_2 accounts for 26%. Similar orders of magnitude are obtained for the other amenities. Therefore, while observed and unob-

[16]To compute this variance decomposition, we first regressed the wage on x, then on x and θ_1. This last regression was weighted by the groups' posterior probabilities computed at the parameter estimates (see Appendix B.2). We proceeded similarly for the latent variables of amenities.

served characteristics account for a large part of the wage variance, the determinants of amenity variables seem essentially unobserved.

The patterns found in Table 3 are qualitatively similar for the wage/amenity equations posterior to constrained job change and posterior to non employment (not shown here). Moreover, the estimates of the parameters of the Ordered PROBIT linking the characteristics of a new match to the starting wage/amenity values in a job are consistent with the patterns of Table 3: higher wages are associated with high θ_1. Then, a high starting satisfaction with any amenity yields a higher θ_2.

Table 4: Within-job compensating differentials, ρ

	AUS	DNK	ESP	FIN	FRA	IRL	ITA	NLD	PRT
TY	.0142	−.00328	.0197	.000724	.0188	.00561	.00242	−.00319	.0154
	(.0046)	(.0033)	(.0035)	(.0038)	(.0031)	(.0062)	(.0027)	(.0030)	(.0044)
CD	−.00162	−.00976	.00672	−.0176	−.00643	−.0102	−.00422	−.0110	−.0101
	(.0046)	(.0032)	(.0035)	(.0037)	(.0029)	(.0063)	(.0027)	(.0027)	(.0043)
WT	−.00278	−.0383	−.00544	−.0246	−.0167	.000235	−.0149	−.0123	−.0219
	(.0044)	(.0033)	(.0034)	(.0036)	(.0027)	(.0062)	(.0026)	(.0029)	(.0045)
DI	.0456	.0158	.00931	.0556	.00896	−.00615	.00197	.0171	−.000334
	(.0038)	(.0031)	(.0032)	(.0037)	(.0027)	(.0059)	(.0023)	(.0029)	(.0037)
SE	.0203	−.0102	.0161	−.00584	.00922	.0497	.00705	.00933	.0190
	(.0040)	(.0030)	(.0035)	(.0038)	(.0027)	(.0062)	(.0026)	(.0029)	(.0044)

We now turn to parameters ρ_1, ..., ρ_5 that give the within job wage/amenity correlation for various amenities. As shown by Table 4, the estimates are very close to zero and, in some cases, insignificant at conventional levels. Moreover, there is no clear pattern in their sign. For instance, in the Netherlands "working conditions" and "working times" are associated to negative wage differentials, while the correlation is positive for "distance to job" and "job security". In all four cases the estimates are significant at the five percent level, yet their order of magnitude is of less than 2%. We also computed ρ estimates without unobserved heterogeneity (the results are not shown here). Comparing these estimates with those displayed in Table 4, we found that controlling for the ability bias, through θ_1, tends to create or increase compensating differentials for disamenities. Even with this correction, though, the ρ point estimates remain low.

Compensating differentials in job offers ρ^* are especially relevant to our analysis. Recall that in this section, restriction (2.16) is imposed, so that job offer parameters are those corresponding to constrained job change. We shall try other hypotheses at the end of this section. The ρ^* estimates displayed in Table 5 are close to zero, roughly of the same order of magnitude as the ρ estimates reported in Table 4. Moreover, as there are fewer constrained job changers than job stayers, standard errors are higher, resulting in mostly insignificant estimates. Interpreting these results in the light of 2.3.2 suggests that, for most amenities, the correlation on the demand side might not be sufficiently negative to create large "compensating" wage differences posterior to job change.

Table 5: Wage/amenity correlation in job offers, ρ^*

	AUS	DNK	ESP	FIN	FRA	IRL	ITA	NLD	PRT
TY	.0134	.0261	.0205	.0365	−.000838	.0503	.0446	.0317	.1056
	(.045)	(.018)	(.020)	(.036)	(.027)	(.042)	(.028)	(.019)	(.033)
CD	−.0217	−.00337	−.00561	.00150	−.0222	−.0106	−.0433	−.0332	.0756
	(.044)	(.019)	(.020)	(.035)	(.027)	(.051)	(.031)	(.018)	(.029)
WT	−.0786	−.00139	−.0110	−.0278	.0105	−.123	.0114	−.00351	−.0815
	(.043)	(.018)	(.019)	(.037)	(.024)	(.0452)	(.029)	(.019)	(.034)
DI	−.0208	−.0215	−.0315	−.0151	.00314	−.00109	.0228	−.0389	−.0216
	(.037)	(.016)	(.019)	(.034)	(.022)	(.045)	(.024)	(.020)	(.027)
SE	.0228	−.00726	.0556	−.0172	−.0142	.0352	−.000783	.0268	.00520
	(.034)	(.017)	(.022)	(.035)	(.023)	(.039)	(.031)	(.018)	(.041)

Voluntary mobility: We here focus on the determinants of voluntary mobility. In the model, job change decisions are based on the comparison of value functions and mobility costs. Therefore, if costs depend on job characteristics, the parameter estimates featured in equation (2.6) will be a mixture of these two elements. In this chapter we make no attempt to separate the value of a job from true transition costs. Instead, we report in Table 6 the results of an OLS regression of $\mathbb{E}[\tau(\theta,x)]$, where $\tau(\theta,x)$ is defined by (2.7), on individual covariates and the last wage/amenity values in the current job.[17] We interpret these coefficients as the weights of different factors in voluntary mobility decisions.

Table 6: Weight of several covariates in $\mathbb{E}[\tau(\theta,x)]$

	AUS	DNK	ESP	FIN	FRA	IRL	ITA	NLD	PRT
age	−.0125 (.018)	−.0513 (.022)	−.0205 (.023)	−.0326 (.058)	.0956 (.073)	.0389 (.039)	−.137 (.077)	−.0423 (.024)	.0986 (.050)
age^2	.000611 (.00026)	.00119 (.00031)	.000619 (.00032)	.00165 (.00087)	.00019 (.00089)	.000168 (.00050)	.00292 (.0013)	.00117 (.00035)	−.0000791 (.00054)
male	−.0260 (.041)	−.119 (.050)	−.157 (.060)	−1.211 (.31)	−1.075 (.38)	−.257 (.10)	−.784 (.27)	−.0916 (.050)	−.924 (.22)
edu= 2^{nd} level	−.258 (.054)	−.139 (.069)	.00277 (.085)	.274 (.19)	.770 (.30)	−.195 (.12)	.0893 (.14)	.550 (.086)	.224 (.17)
edu≥ 3^{rd} level	−.151 (.089)	−.163 (.077)	−.195 (.098)	.0639 (.17)	−.388 (.22)	−.519 (.16)	.149 (.27)	.4005 (.079)	.199 (.22)
married	.0383 (.047)	.00702 (.051)	.0387 (.057)	.2694 (.18)	.287 (.18)	−.0974 (.13)	.475 (.22)	.0737 (.053)	−.233 (.14)
kid	−.0564 (.041)	.0792 (.050)	.123 (.061)	−.186 (.14)	−.188 (.17)	.197 (.11)	.154 (.17)	.0323 (.052)	.124 (.11)
constant	4.843 (.31)	6.244 (.51)	4.218 (.52)	−8.324 (2.219)	3.056 (2.19)	−1.323 (.83)	2.258 (3.35)	3.815 (.51)	−7.228 (1.41)
Wage	.656 (.00053)	.560 (.0066)	.725 (.0063)	2.230 (.039	1.168 (.029)	1.488 (.0019)	1.881 (.057)	.754 (.0037)	1.90 (.026)
TY	.166 (.016)	.102 (.016)	.120 (.021)	.375 (.091)	.372 (.12)	.286 (.045)	.307 (.044)	.112 (.015)	.271 (.040)
CD	.152 (.017)	.0905 (.013)	.120 (.020)	.200 (.077))	.410 (.15)	.217 (.046)	.128 (.054)	.0887 (.013)	.244 (.054)
WT	.0811 (.010)	.0842 (.0082)	.0987 (.014)	.236 (.065)	.303 (.10)	.101 (.034)	.166 (.061)	.0968 (.014)	.263 (.039)
DI	.00959 (.010)	.135 (.031)	−.0178 (.020)	.267 (.075)	.231 (.093)	.158 (.043)	.0852 (.047)	.154 (.030)	.238 (.063)
SE	.140 (.015)	.0988 (.014)	.0913 (.0092)	.374 (.075)	.486 (.15)	.172 (.030)	.376 (.046)	.0652 (.011)	.315 (.057)

The signs and significance of the estimates are rather intuitive. In particular, Table 6 shows that the total effect of age and/or age^2 on $\mathbb{E}[\tau(\theta,x)]$ is positive. Age thus

[17] Again, OLS regressions were weighted by the groups' posterior probabilities. Standard errors are conditional on $x, y_t, \mathbf{a}_t$. Yet, they do account for the variability of the parameters entering $\mathbb{E}[\tau(\theta,x)]$.

reduces significantly the probability of job change. Being a woman is also associated with a lower propensity to change job. These effects have been already noted in the literature (*e.g.* Groot and Verbene, 1997, and Xenogiani, 2003). Having children and being married generally have a similar, though weaker, influence. The effect of education on voluntary mobility seems non-monotonic. Then, the higher the current wage, the lower the probability to quit voluntarily (the estimates range between .56 in Denmark and 2.2 in Finland). Lastly, the coefficients of amenities on the current job are also positive and significant. These findings, common to all countries and amenities, suggest that being satisfied either with one's wage or non wage characteristics deters one from quitting. This result is consistent with the literature starting with Freeman (1978) which studies the effects of job satisfaction on the quit probability.

Then, we divide the coefficients of current amenities by the coefficient of the wage. We interpret the estimates reported in Table 7 as the relative weight of each amenity in the decision to change job. Unsurprinsingly, we find positive and significant estimates for virtually all amenities in every country, ranging around .20. The smallest coefficients are obtained for "distance to job", for which the estimates are insignificant at the 95% level in Austria, Finland and Italy. Note that the estimates are on average both higher and less precisely estimated in France than in the other countries.

Table 7: Weight of current amenities in the mobility decision, relative to the current wage

	AUS	DNK	ESP	FIN	FRA	IRL	ITA	NLD	PRT
TY	.252	.182	.143	.165	.318	.192	.163	.148	.143
	(.024)	(.032)	(.023)	(.030)	(.11)	(.030)	(.028)	(.021)	(.023)
CD	.231	.162	.128	.166	.351	.146	.0683	.118	.128
	(.026)	(.026)	(.030)	(.029)	(.13)	(.031)	(.031)	(.017)	(.030)
WT	.124	.150	.138	.136	.260	.0678	.0885	.128	.138
	(.015)	(.016)	(.023)	(.021)	(.093)	(.023)	(.035)	(.020)	(.023)
DI	.0146	.241	.125	−.025	.198	.106	.0453	.205	.125
	(.016)	(.058)	(.035)	(.028)	(.085)	(.029)	(.026)	(.041)	(.035)
SE	.214	.176	.166	.126	.416	.116	.20	.0865	.166
	(.024)	(.026)	(032)	(.014)	(.14)	(.020)	(.031)	(.015)	(.032)

The results in Table 7 are close in spirit to the methodology introduced by Gronberg and Reed (1994), who estimate the "Marginal Willingness to Pay" for an amenity as the ratio of the elasticities of the hazard rate of job duration with respect to the amenity and the wage, respectively. We obtain comparable results: Gronberg and Reed find that two amenities out of four– measuring several dimensions of "objective" working conditions –have a positive and significant effect on job duration. In the case where they are significant, the relative weights are close to one third of the wage. Van Ommeren *et al.* (2000) obtain similar orders of magnitude in their analysis of commuting.

Heterogeneity in mobility costs. Table 8 reports the estimates of the standard deviation of the stochastic shocks on mobility costs u_z, together with estimates of the weight of wage offers in voluntary job change R_{yz}, defined by (2.11).

Table 8: Heterogeneity in mobility costs

	AUS	DNK	ESP	FIN	FRA	IRL	ITA	NLD	PRT
σ_z	.689 (.052)	1.122 (.13)	3.22 (.83)	.898 (.12)	3.597 (1.19)	1.81 (.32)	3.22 (1.01)	1.245 (.15)	2.52 (.61)
R_{yz}	.122 (.018)	.0422 (.0097)	.0058 (.0030)	.0688 (.018)	.00455 (.0030)	.0296 (.011)	.0058 (.0036)	.0326 (.0078)	.00930 (.0045)

We note that the estimates of the standard deviation σ_z range between .69 in Austria and 3.6 in France. The second row in Table 8 illustrates the magnitude of these standard errors by reporting the estimates of R_{yz}. As explained in section 2.3.2, we interpret this quantity as a measure of heterogeneity in mobility costs. Estimates of R_{yz} are strikingly low in the nine countries, suggesting that many other factors than wage offers might influence the decision to quit. Moreover, voluntary mobility seems much more heterogeneous in the Latin countries (Spain, France, Italy and Portugal) with a ratio R_{yz} of less than one percent and insignificant from zero at the 95% level. This could indicate that individual mobility is highly constrained in these countries, which would be in accordance with the descriptive statistics featured in Table 1.

The explanatory power of both the wage and the amenity variables in voluntary mobility decisions thus appears to be weak, suggesting that mobility costs are highly heterogeneous. Such a low explanatory power is one of our main findings, and has strong implications on the order of magnitude of wage/amenity correlation. At the end of this section we shall investigate the robustness of this result to variations in the model's specifications.

MWP for amenities: We lastly turn to the estimates of the key parameters, the MWP δ^* in job offers. Most estimates in Table 9 are positive, and several are significant. In particular, the type of work and job security are associated to large MWP, around .30 in Denmark and the Netherlands.

The point estimates for δ^* in Latin countries deserve a few specific comments. In France for instance, the MWP for job security is very high. However, although significant, this MWP is not precisely estimated (1.1 with a standard error of .4). This remark carries out to all other amenities in the Latin countries (Spain, France, Italy and Portugal), for which standard errors are much higher than in the other countries. Note that, for these countries, imprecise estimates of MWP are associated with low R_{yz} estimates (see Table 8). Hence, for these four countries only, we should consider estimation results as qualitative findings rather than focusing on the point estimates. At the end of this section, we shall isolate two factors explaining the different results between Latin and non Latin countries: constrained transitions seem more heterogeneous in Latin countries, and men and women have more contrasted mobility behaviors. We shall see that taking these factors into account yields more comparable results between the two groups of countries.

The most notable exception to the general pattern drawn from Table 9 is "working times" in all countries, insignificantly different from zero in most cases. This non intuitive result could be due to the fact that wage rates and hours worked are aggregated in the model (into monthly wages), resulting in a crude modeling of hours worked in workers' preferences. Other exceptions are "working conditions" in the Netherlands and "distance to job " in Finland and the Netherlands, the latter amenity being associated to a negative and significant MWP.

Table 9: MWP for amenities in job offers, δ^*

	AUS	DNK	ESP	FIN	FRA	IRL	ITA	NLD	PRT
TY	.191	.275	.512	.191	.792	.248	.877	.271	.893
	(.058)	(.067)	(.21)	(.062)	(.32)	(.13)	(.34)	(.067)	(.27)
CD	.225	.0798	.203	.0834	.172	.0313	.318	.00712	.319
	(.059)	(.053)	(.18)	(.057)	(.18)	(.13)	(.20)	(.049)	(.16)
WT	.162	.0225	−.196	.0654	−.0395	−.180	−.0371	.00882	.0988
	(.052)	(.053)	(.18)	(.056)	(.15)	(.11)	(.19)	(.052)	(.16)
DI	.0553	.0614	.126	−.0833	.0886	.134	.317	−.190	.336
	(.041)	(.048)	(.16)	(.052)	(.15)	(.11)	(.17)	(.055)	(.15)
SE	.254	.273	.720	.348	1.133	.426	.994	.265	.868
	(.046)	(.056)	(.26)	(.070)	(.40)	(.14)	(.35)	(.057)	(.26)

Nevertheless, the general impression that emerges from Table 9 is one where MWP in job offers can be large. This result can be contrasted with Tables 4 and 5, showing compensating differentials for job stayers and in job offers, respectively. Thus, both the wage and non wage characteristics seem to influence voluntary mobility decisions, suggesting that non wage characteristics do enter individual preferences. In terms of variance, though, the influence appears quite weak, as both the wage and amenities have a low explanatory power in job change decisions. In the next section, we intend to quantify the impact of these two findings on the presence/absence of "compensating" wage differences in cross-section.

2.5.2 MWP and wage differentials

The analysis in 2.3.2 shows that the actual wage differential posterior to job change– between two jobs of different levels of amenities – combines the model's three key parameters: the correlation in wage/amenity offers ρ^*, the MWP for the amenity δ^* and the heterogeneity of mobility costs R_{yz}. We here report the estimates of the various wage differentials, and of their decomposition in terms of the demand and supply effects introduced in 2.3.2.

For a given amenity, the demand effect $\Delta_z^{(d)}$ is equal to the compensating differential ρ^* in job offers corresponding to this amenity. Estimates of this effect can be found in Table 5. The supply term, $\Delta_z^{(s)}$, arises from workers' trade-offs between wages and amenities when deciding to change job. In the upper half of Table 10, we report the estimates of $\Delta_z^{(s)}$ for all countries and amenities.[18]

Estimates of $\Delta_z^{(s)}$ are mostly negative, consistently with wage/amenity compensation on the supply side, as MWP δ^* are mostly positive (see Table 9). However, heterogenity in mobility costs significantly reduces the magnitude of this effect. Comparing Table 10 with Table 9 shows that MWP of .30 translate into correlations of less than .02.[19] Then, the lower half of Table 10 shows the sum of the demand and supply effects $\Delta_z = \Delta_z^{(d)} + \Delta_z^{(s)}$. It is clearly apparent from the table that, when heterogeneous mobility costs and non zero correlation in job offers are taken into ac-

[18]To estimate the latter, we computed the RHS in (2.10) for every individual in the sample, and averaged over x, weighted by the groups' posterior probabilities. In theory, the delta-method is not sufficient to compute standard errors in this case, as one has to account also for the estimation of the expectation in (2.10) by a sample mean. As this latter source of variation turned out to be negligible relative to the variation in the model's parameters, however, the delta-method was used as yielding a very good approximation of true standard errors.

[19]Note that the approximation of the wage differential $\Delta_z^{(s)}$ given in (2.13) works here very well. Indeed, combining the results in Tables 8, 9 and 10 it is easy to check that compensating differentials are roughly the product of MWP δ_z^* (net of ρ^*, which is close to zero in most cases) and heterogeneity in mobility costs R_{yz}.

count, the resulting wage/amenity correlation posterior to job change does not reflect workers' underlying preferences for non wage attributes.

Table 10: Wage differentials for voluntary job changers

	AUS	DNK	ESP	FIN	FRA	IRL	ITA	NLD	PRT
					$\Delta_z^{(s)}$				
TY	−.0231	−.0112	−.00279	−.0141	−.00327	−.00781	−.00483	−.00871	−.00832
	(.0084)	(.0027)	(.0012)	(.0048)	(.00129)	(.0035)	(.0017)	(.0020)	(.0024)
CD	−.0230	−.00286	−.00104	−.00528	−.000621	−.000544	−.00145	.000758	−.00331
	(.0086)	(.0021)	(.00093)	(.0041)	(.000738)	(.0036)	(.0010)	(.0015)	(.0015)
WT	−.00943	−.000791	.00109	−.00233	.000120	.00801	.000135	−.000154	−.000146
	(.0076)	(.0021)	(.00089)	(.0042)	(.000651)	(.0038)	(.00099)	(.0016)	(.0014)
DI	−.00392	−.00150	−.000497	.00614	−.000381	−.00349	−.00179	.00670	−.00264
	(.0063)	(.0019)	(.00086)	(.0038)	(.000626)	(.0033)	(.0010)	(.0017)	(.0013)
SE	−.0312	−.00990	−.00405	−.0204	−.00461	−.0121	−.00520	−.00840	−.00728
	(.0069)	(.0023)	(.0013|)	(.0049)	(.00166)	(.0037)	(.0019)	(.0019)	(.0023)
					$\Delta_z = \Delta_z^{(d)} + \Delta_z^{(s)}$				
TY	−.00970	.0149	.0177	.0224	−.00411	.0425	.0398	.0230	.0973
	(.040)	(.018)	(.020)	(.034)	(.027)	(.041)	(.028)	(.018)	(.033)
CD	−.0447	−.00624	−.00665	−.00377	−.0228	−.0112	−.0447	−.0325	.0722
	(.039)	(.019)	(.020)	(.033)	(.027)	(.049)	(.031)	(.018)	(.029)
WT	−.0880	−.00218	−.00987	−.0302	.0106	−.115	.0115	−.00366	−.0817
	(.039)	(.017)	(.019)	(.035)	(.024)	(.044)	(.029)	(.019)	(.033)
DI	−.0247	−.0230	−.0320	−.00893	.00276	−.00458	.0210	−.0322	−.0242
	(.033)	(.016)	(.019)	(.032)	(.022)	(.044)	(.024)	(.020)	(.027)
SE	−.00845	−.0171	.0516	−.0377	−.0188	.0231	−.00598	.0184	−.00208
	(.031)	(.017)	(.022)	(.033)	(.023)	(.038)	(.031)	(.017)	(.041)

Thus, evidence of "compensating" wage differences is rather weak in cross-section, although workers seem to value non wage characteristics. Workers' trade-offs between wages and amenities translate into a very small, possibly still negative, correlation. This section has emphasized two key elements in this mechanism: the low explanatory power of the wage and amenities in job mobility decisions (low R_{yz}), and the often insignificant correlation in job offers (low $|\rho^*|$). Our results thus shed light on the difficulty of finding compensating differentials in cross-section, even conditional on unobserved heterogeneity, and even if individuals value non wage characteristics significantly relatively to the wage.

2.5.3 Robustness checks

We here check the robustness of the parameter estimates reported in 2.5.1 to changes in the model's specification. We start by checking whether changes in the distribution of reference for job offers leads to empirical regularities supporting the results displayed in section 2.5.1. Then we see how results might be affected if mobility behaviors differ greatly with respect to gender. Lastly, we proceed to additional checks by changing the number of groups of heterogeneity and introducing measurement error in wages.

Changing the distribution of reference. First, we disaggregate constrained transitions into partially constrained transitions and displacements, and model each process separately. We do so to address the concern that constrained transitions, as defined in this chapter, may recover different phenomena. For instance, family-related job mobility is an example of partially constrained transition. However, employer-related job changes (firm closure, layoff...), that we call displacements, could be more exogenous from the worker's perspective.

We model partially constrained transitions and displacements as in equation (2.8), with different sets of parameters. Likewise, wage and amenity equations follow the pattern of (2.1)-(2.2), again with different parameters. The sequence of events is the following: between t and $t+1$ employed individuals can experience a job to non employment transition. If they do not, they can still lose their job and get a new one before the next interview. Then, if they are not displaced, they can quit their present job for personal reasons and make a partially constrained transition to a new job. All the probabilities and corresponding wage/amenity distributions are conditional on x

and θ. Lastly, if the worker has experienced none of these shocks, she receives a job offer that she can accept or turn down, as in 2.3.1.

Table 11: MWP (δ^*) and R_{yz} estimates for various transitions of reference

	AUS	DNK	ESP	FIN	FRA	IRL	ITA	NLD	PRT
					Displacements				
TY	.263	.193	.659	.177	.725	.143	1.569	.208	.437
	(.051)	(.055)	(.31)	(.065)	(.29)	(.083)	(1.11)	(.038)	(.10)
CD	.162	.0366	.223	.0676	.0951	.0779	.645	−.0169	.130
	(.050)	(.045)	(.24)	(.059)	(.16)	(.088)	(.55)	(.030)	(.081)
WT	.178	−.0570	−.348	.0114	−.111	−.206	−.470	−.00579	.0526
	(.045)	(.047)	(.25)	(.059)	(.15)	(.082)	(.50)	(.032)	(.089)
DI	.0482	.116	.0735	−.102	.0245	.0146	1.012	−.122	.221
	(.036)	(.042)	(.21)	(.055)	(.14)	(.080)	(.68)	(.032)	(.079)
SE	.378	.418	1.033	.370	1.323	.401	3.020	.3296	.758
	(.046)	(.057)	(.43)	(.076)	(.43)	(.095)	(1.96)	(.036)	(.13)
R_{yz}	.117	.0489	.00338	.0446	.00555	.0517	.00150	.0806	.0289
	(.023)	(.010)	(.0023)	(.019)	(.0035)	(.014)	(.0019)	(.012)	(.0080)
					Partially constrained transitions				
TY	.157	.423	.171	.165	.411	.227	.434	.301	.725
	(.077)	(.092)	(.053)	(.041)	(.077)	(.099)	(.11)	(.075)	(.20)
CD	.428	.156	.146	−.0374	.117	−.0490	.308	.0451	.183
	(.082)	(.067)	(.054)	(.041)	(.064)	(.10)	(.093)	(.052)	(.14)
WT	.134	.130	.0181	.10	.185	−.0185	.0418	.0402	−.00932
	(.069)	(.065)	(.055)	(.040)	(.060)	(.086)	(.086)	(.056)	(.15)
DI	.0899	.00902	.152	−.0243	.143	.261	.0624	−.206	.0953
	(.053)	(.058)	(.060)	(.035)	(.056)	(.092)	(.067)	(.060)	(.12)
SE	.222	.131	−.0409	.305	.0895	.237	.235	.156	.467
	(.059)	(.063)	(.050)	(.047)	(.056)	(.095)	(.083)	(.056)	(.16)
R_{yz}	.0772	.031	.0591	.174	.0242	.046	.0217	.0241	.0108
	(.017)	(.0088)	(.015)	(.046)	(.0068)	(.014)	(.0068)	(.0063)	(.0049)
					Out-of-non employment				
TY	.902	.154	−.0155	.209	.454	.405	.292	.401	.125
	(.25)	(.070)	(.059)	(.10)	(.11)	(.14)	(.088)	(.065)	(.048)
CD	.759	.127	.0445	.175	.518	−.179	−.0536	.0241	.202
	(.22)	(.061)	(.060)	(.099)	(.11)	(.15)	(.080)	(.045)	(.053)
WT	−.046	.173	−.0469	−.0525	.119	−.369	−.00877	.109	−.0088
	(.17)	(.062)	(.058)	(.094)	(.073)	(.14)	(.082)	(.049)	(.053)
DI	.228	−.0918	.0723	−.310	.204	.332	−.0537	−.0598	.0944
	(.13)	(.057)	(.055)	(.11)	(.074)	(.13)	(.066)	(.045)	(.047)
SE	.597	.317	.0931	.580	.354	.0402	.363	.227	.0845
	(.17)	(.066)	(.058)	(.14)	(.088)	(.12)	(.086)	(.048)	(.052)
R_{yz}	.0104	.0281	.0516	.0104	.0106	.00949	.00581	.0635	.0531
	(.0044)	(.0071)	(.0082)	(.0044)	(.0032)	(.0035)	(.0017)	(.012)	(.0092)

Within this framework, we try three different specifications for job offers. Subsequently, job offer parameters are assumed equal to the ones in displacements, partially constrained transitions and transitions out of non employment.[20] Table 11 presents the estimation results for MWP in job offers δ^* and heterogeneity in mobility costs R_{yz}.

Table 11 reinforces the two main qualitative findings of 2.5.1. First, MWP for amenities are mostly positive, and can be large for some amenities. Thus, the type of work and job security are associated with positive and significant MWP for almost all countries and every choice of transition of reference. Moreover, many MWP amount to a large share of the wage, up to 40% in non Latin countries. Second, every specification shows large heterogeneity in mobility costs. For instance, in the Netherlands wage offers account for less than 8% of the variation in voluntary mobility, irrespective of the type of transitions chosen as a reference.

The main results of 2.5.1 thus appear robust to changes in the reference distribution. Still, a closer look at Table 11 reveals several interesting features. First, we note that the choice of the distribution of reference rarely influences the sign or significance of the MWP estimates. However, it can alter their ranking in a given country. For instance, the type of work and job security are associated with the highest MWP in any country when using either constrained transitions or displacements as the reference. In contrast, when partially constrained job change is used, the MWP for job security strongly decreases. For instance, in Denmark the MWP estimate for this amenity is .42 when using displacements as the reference, and .13 (yet still significant at the 95% level) when using partially constrained transitions. An intuitive explanation could be

[20]In several cases (5 out of 27), there was not enough information in the data to identify $K_1 = 4$ groups of heterogeneity θ_1. We have thus retained $(K_1, K_2) = (3, 2)$ for the following crossings: displacements and out-of-non employment in Ireland and partially constrained transitions in Spain, Ireland and Portugal.

that such transitions are experienced by workers who change jobs for personal reasons (marriage, geographic mobility...) but whose alternative is not non employment. In this interpretation, MWP estimates for job security based on the use of displacements are higher, because they incorporate the risk of non employment– possibly correlated with the aversion to job insecurity.

A second interesting feature of Table 11 concerns the grouping of countries which emerged from Tables 8 and 9. In the Latin countries, indeed, in the two cases where displacements are not part of the transitions of reference, heterogeneity in mobility costs is reduced. In Spain, France, Italy and Portugal, the R_{yz} estimates for these two specifications are higher, and significantly different from zero. Simultaneously, MWP estimates are more in line with the results for non Latin countries, ranging around one third of the wage for the type of work and job security. Moreover, when restricting constrained transitions to partially constrained or out-of non employment transitions, standard errors of MWP estimates are lower, indicating that MWP are better estimated. These findings suggest that displacements (as defined in this chapter) are not a satisfying control group in Latin countries.

Testing for the robustness of the estimates with respect to changes in the distribution of reference was essential to our approach and we view these empirical regularities as supportive of our results and interpretation. Still, one can question other features of the model than the identification assumption discussed in 2.4.1. We thus proceeded to a series of alternative robustness checks.[21]

Separate analyses with respect to gender: In the model, voluntary mobility depends on individual covariates in a parametric way, as shown by (2.6). However, it

[21]In the rest of this section, constrained transitions are taken as reference and restriction (2.16) is imposed, as in 2.5.1.

could be that women and men, have very different mobility behaviors, which could not be well captured by a parametric specification. Moreover, this problem could affect some countries more than others. For instance, in Table 6, we see that, compared to other covariates, being a man has a large positive effect on the probability of changing job voluntarily in France, Italy or Portugal. To address this issue, we proceeded to the estimation of the model on the subsamples of men. We found that the heterogeneity in mobility costs remains practically similar in non Latin countries, except in the Netherlands where σ_z increases to 1.6. In Latin countries, σ_z decreases but stays at rather high levels. For instance, the σ_z estimate goes down to 2.8 in Spain, to 2.4 in France and to 2.6 in Italy (the σ_z estimates are 3.2, 3.6 and 3.2, respectively, when pooling men and women, see Table 8). Moreover, if the MWP for amenities are not qualitatively affected, their order of magnitude is somewhat closer to the one obtained for non Latin countries. For instance, in Italy the MWP for "type of work" is reduced from .90 (men and women) to .58 (men only), the MWP for "job security" from .99 to .63. The highest estimate reported in Table 9, the MWP for "job security" in France, goes down from 1.13 to a still high– yet more reasonable– .76. To save space, we do not report the corresponding results in the present version but they are available upon request.

Additional checks: Lastly, we varied the number of groups for θ_1 or θ_2 and tried the following (K_1, K_2) pairs: $(1,1)$, $(2,2)$, $(3,2)$, $(5,2)$ and $(4,3)$. The results, as far as MWP and R_{yz} estimates are concerned, are qualitatively similar when allowing for more than $K_1 = 2$ and $K_2 = 2$ groups. In some countries such as Denmark and Ireland, this is also the case when no heterogeneity is allowed for. In the Netherlands or Portugal, however, results differ greatly in the homogeneous case, with σ_z estimates of 3.9 and 11.0, respectively, and large and badly estimated MWP. Also, we modified the wage observations by adding an i.i.d. perturbation, normally distributed with standard error equal to 10% of that of the observed wage. We found the estimates strongly robust to this kind of measurement error.

To summarize the main results of this section, we find large and significant MWP in many cases, especially for two amenities: the type of work and job security. We also find, under every specification, high heterogeneity in mobility costs. An intriguing finding is the contrast between Latin and non Latin countries, the former being associated with even higher measures of heterogeneity in mobility costs, together with imprecise MWP estimates. In the end of this section, we have emphasized two possible explanations: the greater heterogeneity in mobility behavior between men and women in Latin countries, and the greater heterogeneity in types of constrained transitions (displacements *versus* partially constrained).

2.6 Conclusion

The theory of compensating differentials builds on Adam Smith's original statement:[22] *"The whole of the advantages and disadvantages of the different employments of labour and stock must, in the same neighborhood, be either perfectly equal or continually tending to equality."*

On the labor market, this implies that "bad" non-monetary characteristics of one's job must be compensated by higher wages. However, hedonic wage regressions lead to weak or even wrong-signed wage/amenity correlations. In this chapter, we show that these results must not be interpreted as reflecting individual preferences for non wage amenities. Smith had indeed pointed out the conditions under which the *"equality of*

[22] *An Inquiry Into the Nature and Causes of the Wealth of Nations*, Book 1, Chapter 10, Introduction.

advantages and disadvantages" was to be expected:

"This at least would be the case in a society where things were left to follow their natural course, where there was perfect liberty, and where every man was perfectly free both to choose what occupation he thought proper, and to change it as often as he thought proper."

In modern European economies, very low rates of voluntary mobility suggest that workers are far from being *"perfectly free"* to change jobs. Consequently, the predictions of the theory of compensating differentials are unlikely to hold.

Our estimation results show significant valuation of several non wage characteristics, mostly the type of work and job security, in spite of low wage/amenity correlations. However, the low explanatory power of both the wage and amenities in job mobility, and the small correlation in job offers, imply that workers' preferences do not translate into significant negative correlation.

The method advocated in this chapter makes use of the difference in the degree of constraints in mobility decisions to reveal individual preferences: constrained transitions allow to estimate the available alternatives, then voluntary ones permit to measure the true effects of individual choice. This approach could be applied to other fields where hedonic methods are widely used. An example is the estimation of MWP for air quality in environmental economics. In a recent paper, Bayer et al. (2005) estimate a model of residential sorting allowing for mobility costs. Their MWP estimates are larger than usual hedonic regression estimates. The method and results of the present chapter suggest that, in such a field, distinguishing between the reasons to migrate could prove fruitful to deal with endogeneity problems.

Lastly, on the labor market, our results shed light on the empirical content of non wage job characteristics, showing that they are part of workers' preferences. This evidence could be seen as a motivation for labor economists to incorporate other job attributes than the wage into their models. We note that there have recently been several attempts at broadening the analysis of earnings inequality to that of monetary compensation inequality (Pierce, 2001) or, more generally, inequality in the returns to work (Hamermesh, 1999). In a different field, dynamic structural models of the labor market are just starting to take amenities into account. For instance, Dey and Flinn (2005) write and estimate an equilibrium job search model where firms can also provide health insurance to their employees. We view our results and their implications as supportive of these multi dimensional analyses of the labor market, for which the availability of more informative data sets now seems to be the main obstacle.

Chapter 3

Using High-Order Moments to Estimate Linear Independent Factor Models

3.1 Introduction

Linear factor models are routinely used in social sciences. Spearman's (1904) "g" factor is one of the earliest applications in psychology. Principal component analysis (PCA) is a leading technique in sociology to construct social indices and to uncover hidden causes of individual actions. Econometric applications include measurement error models, error component models for panel data, structural VAR models in macroeconomics, and multifactor asset pricing models in empirical finance. Linear factor models have also been used in nonlinear empirical microeconomic models. For example, Carneiro, Hansen and Heckman's (2003) Roy model of educational choice is a successful application of factor models for estimating treatment effects and other policy parameters using microdata.[1]

Despite these empirical successes, it is usually thought that the interest of linear multifactor models for structural applications is severely hampered by a fundamental lack of identification. Suppose that a vector of L observed measurements, Y, be related to a vector of K unobserved factors, X, by a noisy linear relationship: $Y = \Lambda X + U$, where Λ is a matrix of parameters (factor loadings) and U is a vector of errors. In ordinary Factor Analysis, the identification of factor loadings rests on covariance restrictions, and it is well known that matrix Λ is identified only up to a multiplicative orthogonal matrix (Anderson and Rubin, 1956). Parametric restrictions, often in the form of exclusion restrictions, are usually added for identification. In VAR models, for example, the identification of structural shocks is achieved by assuming a particular triangular form for Λ. In the same spirit, Carneiro, Hansen and Heckman (2003)

[1] Continuous instruments with large supports allow to identify the distribution of latent variables and a linear factor structure is used to model the effect of unobserved heterogeneity on latent variables. See Cunha *et. al* (2005) and Heckman and Navarro (2005) for other applications of this idea.

assume that there is at least two specific measurements for each factor.

In this chapter, we show that these exclusion restrictions are unnecessary if two key conditions are satisfied: First, factors and errors are *independent*, not just uncorrelated. Second, the third and/or fourth-order moments of the vector of observed measurements are informative, which implies that factors are *not Gaussian*. If $K \leq L$, we show that the matrix of factor loadings Λ is generically identified from second, third and fourth-order moments of the data. Then, if $K < L$, we show that Λ is identified from second and third-order moments only. In both cases, identification is unambiguously defined up to multiplication of each column by ± 1 and column permutations.

The importance of the assumptions of independence and non normality for the identification of one-factor models is well known in the measurement-error literature. Since the seminal contributions of Geary (1942) and Reiersol (1950) a long series of papers have proposed different ways of using third and fourth-order moments to correct estimators for measurement errors in the regressors.[2] The class of estimators introduced in this chapter can be seen as a generalization of this approach to multifactor structures.

In a different branch of statistics, signal processing, linear factor models are commonly used to separate the components of linear mixtures of signals. Since its introduction at the beginning of the 1990's, Independent Component Analysis (ICA) has rapidly become a leading technique for *blind signal separation*.[3] In this vast lit-

[2]Relevant contributions include Madanski (1959), Pal (1980), Dagenais and Dagenais (1997), Cragg (1997), Lewbel (1997), and Erickson and Whitted (2002). Less directly related to our work are the papers of Spiegelman (1979) and Van Montfort *et al.* (1989), using more of the information contained in the characteristic function of measurements than the value at zero of its first few derivatives. Lastly, Lewbel (2004) and Doz and Renault (2005) use heteroskedasticity as a source of identification.

[3]The designation "Independent Component Analysis" was first proposed par Comon (1994). See Hyvärinen *et al.* (2001) and Cardoso (1999) for surveys.

erature, one of the most popular methods is Cardoso and Souloumiac's (1993) JADE algorithm. This is a joint diagonalization algorithm of a set of well chosen matrices of fourth-order cumulants of measurements. In the past ten years, the ICA problem has also become an important topic in the neural networks literature and Hyvärinen's (1999) FastICA algorithm has become another very popular algorithm.[4]

One serious drawback of ICA, at least for econometric applications, is that it rules out measurement errors. The estimated model is $Y = \Lambda X$, with $K = L$, not $Y = \Lambda X + U$. Neglecting noise can be a source of severe biases, as we shall show. All existing extensions of ICA allowing for noise make parametric assumptions on the distributions of errors (usually Gaussian) and factors (usually Gaussian mixtures).[5] As far as we know, this chapter is the first attempt, out of a long list of contributions, at proposing a semiparametric statistical procedure for consistently estimating a $K \times K$ matrix of factor loadings from data moments in a linear factor model with error distributions of unknown form.

We develop an algebraic procedure that builds on Cardoso and Souloumiac's JADE algorithm. Our *quasi-JADE* algorithm proceeds in two stages: First, we estimate the second, third and fourth-order error moments, which we use to "remove" the noise component from the second, third and fourth-order moments of the data ("whitening" stage). Then, we straightforwardly apply Cardoso and Souloumiac's joint diagonalization algorithm to the "whitened" data. Notice that we therefore do not need to assume full independence between factor and error components, only that they are orthogonal up to third or fourth order.

[4]See Xu, 2003, for a survey of Bayesian learning applications to ICA.

[5]For example, Moulines *et al.* (1997) and Attias (1999) use a ML approach and the EM algorithm. Xu (2000, 2001) allows for non-Gaussian errors and uses Bayesian learning algorithms. Ikeda and Toyama (2000) adopt a two-stage method which combines PCA and JADE to reduce the size of the noise. As such, the estimator they propose is still inconsistent in the presence of noise.

The outline of the paper is as follows. In Section 3.2, we study the semiparametric identification of factor loadings. Section 3.3 deals with estimation issues. We first discuss the estimation of the number of common factors using Robin and Smith's (2000) rank test. We then present Cardoso and Souloumiac's (1993) JADE algorithm and study its asymptotic properties. Lastly, we introduce the quasi-JADE algorithm. In Section 4.6, we investigate the finite-sample properties of quasi-JADE by means of Monte-Carlo simulations.

In Section 3.5, we apply our methodology to estimating returns to schooling in France, using microdata from the French Labor Force Survey. Our method allows to identify two factors of individual wages and education. Interestingly, while the first factor has a positive effect on wages, the second factor is positively related to education, yet negatively to wages. This is evidence that there exist individual characteristics which are valued by the education institution but not by the labor market. Moreover, the exhibited factor structure is consistent with the standard model of education returns if one allows measurement errors on the education measure and unobserved heterogeneity.

Lastly, Section 3.6 concludes.

3.2 Identification of linear independent factor models

Let $Y = (Y_1, ..., Y_L)^T$ be a vector of $L \geq 2$ zero-mean, real-valued random variables (measurements). Let $X = (X_1, ..., X_K)^T$ be a random vector of $K \geq 1$ zero-mean, real valued, non degenerate random variables (factors). Let $U = (U_1, ..., U_L)^T$ be a vector of L zero-mean, real-valued random variables (errors). Both factors and errors are unobserved.

Assumption A1 ***(Linearity)*** *There exists a $L \times K$ matrix of scalar parameters (factor loadings), Λ, such that $Y = \Lambda X + U$.*

The difference between factors and errors is a matter of definition. A given covariate is called a factor if it enters at least two measurement equations (i.e. every column $\boldsymbol{\lambda}_k$, $k = 1, ..., K$, of Λ has at least two non-zero entries). Otherwise, it is called an error.

In ordinary Factor Analysis (FA), factors and errors are uncorrelated and identification rests on the following covariance restrictions:

$$\Sigma_Y = \Lambda \Sigma_X \Lambda^T + \Sigma_U, \tag{3.1}$$

where Σ_Z denotes the variance-covariance matrix of any random vector Z. Obviously parameters Λ, Σ_X and Σ_U are not identified from second-order restrictions (see Anderson and Rubin, 1956). First, restrictions are needed on the correlations between errors and it is usually assumed that Σ_U is diagonal. Second, Σ_X is not separately identified from Λ. If (Λ, Σ_X) satisfies (3.1), then so does $(\Lambda\Omega, I_K)$, where $\Omega\Omega^T = \Sigma_X$. The variance-covariance matrix of X is usually normalized to the identity matrix I_K. Thirdly, even if $\Sigma_X = I_K$ and Σ_U is diagonal, Λ is identified only up to an orthogonal matrix; that is, if Λ satisfies the covariance restrictions, then so does ΛP, for any orthonormal matrix P. Principal Component Analysis is the least-squares version of FA.[6]

[6] Given a sample $\mathbf{Y} = (Y_1, ..., Y_N)$ of observations, Principal Component Analysis estimates both Λ and factor realizations $\mathbf{X} = (X_1, ..., X_N)$ by non linear least squares under the constraint $\frac{1}{N}\mathbf{X}\mathbf{X}^T = I_K$. Principal components $\widehat{\mathbf{X}}$ are the first K eigenvectors of the $N \times N$ matrix $\mathbf{Y}^T\mathbf{Y}$ (corresponding to the K largest eigenvalues), and factor loadings are estimated by regressing $\mathbf{Y}$ on $\widehat{\mathbf{X}}$ by OLS. Common factors ΛX_n, $n = 1, ..., N$, are identified if matrix $\operatorname{Var}(Y)$ has no multiple eigenvalue but identifying factors X_n from factor loadings requires the arbitrary choice of a rotation P, even under the normalization $\operatorname{Var}(X) = I_K$. The asymptotic theory of principal components usually assumes a fixed number of measurements L but a large sample size N (see Anderson, 1984, and Lawley and Maxwell, 1971). In two recent papers, Bai and Ng (2002) and Bai (2003) study the case where both L and N tend to infinity.

In this chapter, we maintain the assumptions that $\Sigma_X = I_K$ and Σ_U is diagonal and we intensify the absence of correlations between factors, between errors and between factors and errors by making them independent. Moreover we assume that factors are non Gaussian with finite third and fourth-order moments.

Assumption A2 ***(Normalization)*** *Factors have unit variances.*

Assumption A3 ***(Independence)*** *All factor and error variables are mutually independent.*

Assumption A4 ***(Non gaussianity)*** *Factor variables X_k, $k = 1, ..., K$, are non Gaussian with finite third and fourth-order moments.*

In addition, in order to prove the semi-parametric identification results below, we shall require characteristic functions and cumulant generating functions to exist and to be smooth on all $\mathbb{R}$ and not only locally around the origin, as implied by Assumption A4. This assumption is yet not necessary for the parametric estimation procedures that we shall later develop.

Assumption A5 ***(Characteristic functions)*** *The characteristic functions of factors and errors are of class $\mathcal{C}^2$ on $\mathbb{R}$, and are nonvanishing almost everywhere.*

We shall say that a representation (Λ, X, U) is regular if it satisfies all previously listed assumptions.

3.2.1 Definitions

For all K, let us define the set of sign-permutation matrices as the set $\mathcal{S}_K$ of all products DP, where D is a diagonal matrix with diagonal components equal to 1 or -1 and P is a permutation matrix. For given values of L and K, let (Λ, X, U) be a regular representation. Clearly, for all $S \in \mathcal{S}_K$, $\left(\Lambda S, S^T X, U\right)$ is another regular representation. Hence, identification has to be defined modulo the set $\mathcal{S}_K$.

Note that the group $\mathcal{S}_K$ is a finite subgroup of the infinite orthogonal group $\mathcal{O}_K$, up to which identification is defined in ordinary or orthogonal Factor Analysis. The quotient group $\mathcal{O}_K/\mathcal{S}_K$ is thus also infinite. Proving identification results modulo $\mathcal{S}_K$, instead of modulo $\mathcal{O}_K$, will result in a considerable reduction of the model's indeterminacy.

We define semi-parametric identification as follows.

Definition 1 ***(Semiparametric identification)*** *A regular representation* (Λ, X, U) *is said identifiable if for every other regular representation* $(\widetilde{\Lambda}, \widetilde{X}, \widetilde{U})$ *there exists a matrix* S *in* $\mathcal{S}_K$ *such that:* $\widetilde{\Lambda} = \Lambda S$, $\widetilde{X} \stackrel{d}{=} S^T X$, *and* $\widetilde{U} \stackrel{d}{=} U$, *where* $\stackrel{d}{=}$ *means "equal in distribution."*

Semiparametric identification draws information on the finite-dimensional parameter Λ and the infinite-dimensional parameters that are the distributions of X and U from the whole distribution of observed measurements. For practical reasons, it is useful to understand how much of the model's structure can be identified from a finite set of parameters of this distribution. We thus also define parametric identification as follows.

Definition 2 ***(Parametric identification)*** *A regular representation* (Λ, X, U) *is said to be parametrically identified if their exists a finite vector of moments* $\mathbb{M}(Z)$, *defined for a vector of r.v.* Z, *such that, for every other regular representation* $(\widetilde{\Lambda}, \widetilde{X}, \widetilde{U})$, *moment equality:* $\mathbb{M}(\Lambda X + U) = \mathbb{M}(\widetilde{\Lambda}\widetilde{X} + \widetilde{U})$, *implies that there exists a sign-permutation matrix* $S \in \mathcal{S}_K$ *such that* $\widetilde{\Lambda} = \Lambda S$, $\mathbb{M}(\widetilde{X}) = \mathbb{M}\left(S^T X\right)$, *and* $\mathbb{M}(\widetilde{U}) = \mathbb{M}(U)$.

Unless otherwise specified, we shall simply say that the factor model is parametrically identified if it is identified from second, third and fourth-order moments of observed measurements.

3.2.2 Identifying restrictions

In this subsection, we develop some implications of the regularity assumptions in terms of cumulant generating functions and their derivatives.

Cumulant generating function. Denote the cumulant generating functions (the log of characteristic functions) of Y, X_k and U_ℓ as κ_Y, κ_{X_k} and κ_{U_ℓ}. The independence assumptions and the linear factor structure imply that, for almost all $t = (t_1, ..., t_L)^T \in \mathbb{R}^L$,

$$\kappa_Y(t) \equiv \ln\left[\mathbb{E}\exp\left(\sqrt{-1}\cdot t^T Y\right)\right] = \sum_{k=1}^{K} \kappa_{X_k}\left(\boldsymbol{\lambda}_k^T t\right) + \sum_{\ell=1}^{L} \kappa_{U_\ell}\left(t_\ell\right), \tag{3.2}$$

where ln denotes the principal branch of the logarithm.

Then, define the following sets of multi-indices:

$$\begin{aligned}
\overline{\Delta}_{L,p} &= \left\{\alpha = (\alpha_1, ..., \alpha_L) \in \{0, ..., p\}^L : |\alpha| = \alpha_1 + ... + \alpha_L = p\right\}, \\
\Delta_{L,p} &= \left\{\alpha = (\alpha_1, ..., \alpha_L) \in \{0, ..., p-1\}^L : |\alpha| = \alpha_1 + ... + \alpha_L = p\right\}.
\end{aligned}$$

Let $\#\overline{\Delta}_{L,p}$ (resp. $\#\Delta_{L,p}$) be the number of elements in $\overline{\Delta}_{L,p}$ (resp. $\Delta_{L,p}$). For $p = 2$: $\#\overline{\Delta}_{L,2} = \frac{L(L+1)}{2}$ and $\#\Delta_{L,2} = \frac{L(L-1)}{2}$.[7]

[7]Multi-indices are convenient ways to select p components of a vector of size L with repetition, *via* the following bijection

$$\begin{aligned}
\Psi_{L,p}: \ \overline{\Delta}_{L,p} &\longrightarrow \{1, ..., L\}^p \\
\alpha &\longmapsto (\ell_1, ..., \ell_p) = \Psi_{L,p}(\alpha)
\end{aligned}$$

where $(\ell_1, ..., \ell_p)$ is such that $\ell_1 \leq ... \leq \ell_p$ and $\mathbf{i}_{L,\ell_1} + ... + \mathbf{i}_{L,\ell_p} = \alpha$, vector $\mathbf{i}_{L,\ell}$ denoting the ℓth column of the identity matrix of dimension L. For example, $\alpha = (2,1,1) \in \overline{\Delta}_{3,4}$ corresponds to variable indices $(\ell_1, \ell_2, \ell_3, \ell_4) = (1,1,2,3)$, $\alpha = (0,2,2)$ to $(\ell_1, \ell_2, \ell_3, \ell_4) = (2,2,3,3)$, etc. To simplify the notation, we shall also denote as $\overline{\Delta}_{L,p}$ and $\Delta_{L,p}$ their image by $\Psi_{L,p}$.

Let $\alpha = (\alpha_1, ..., \alpha_L) \in \overline{\Delta}_{L,p}$. For any vector $x = (x_1, ..., x_L) \in \mathbb{R}^L$, define the monomial $x^\alpha = x_1^{\alpha_1} \cdots x_L^{\alpha_L}$. Then, assuming that derivatives exist, we have

$$\kappa_Y^{(\alpha)}(t) \equiv \partial_\alpha \kappa_Y(t) \equiv \frac{\partial^{|\alpha|} \kappa_Y(t)}{\partial t_1^{\alpha_1} \cdots \partial t_L^{\alpha_L}} = \sum_{k=1}^{K} \boldsymbol{\lambda}_k^\alpha \kappa_{X_k}^{(p)} \left(\boldsymbol{\lambda}_k^T t \right) + \sum_{\ell=1}^{L} \delta_{\alpha_\ell, p} \kappa_{U_\ell}^{(p)} (t_\ell), \qquad (3.3)$$

where $\kappa_{X_k}^{(p)}$ and $\kappa_{U_\ell}^{(p)}$ are the pth derivative of κ_{X_k} and κ_{U_ℓ}, and δ_{ij} is the Kronecker delta (= 1 if $i = j$ and = 0 if $i \neq j$).

Cumulants. For any multi-index α, one defines a multivariate cumulant as

$$\kappa_\alpha (Y) = \frac{\kappa_Y^{(\alpha)}(0)}{\left(\sqrt{-1}\right)^{|p|}}.$$

Let $\mathbf{i}_\ell$, $\ell \in \{0, ..., L\}$, be the ℓth column of the $L \times L$ identity matrix. For any $p$tuple $(\ell_1, ..., \ell_p) \in \{0, ..., L\}^p$, we denote as $\text{Cum}\left(Y_{\ell_1}, ..., Y_{\ell_p}\right) \equiv \kappa_{\mathbf{i}_{\ell_1} + ... + \mathbf{i}_{\ell_p}} (Y)$ the multivariate cumulant of $\left(Y_{\ell_1}, ..., Y_{\ell_p}\right)$.

The second-order cumulants of zero-mean random variables are equal to their covariances:

$$\text{Cum}\,(Y_{\ell_1}, Y_{\ell_2}) = \mathbb{E}(Y_{\ell_1} Y_{\ell_2}), \qquad (3.4)$$

Third-order cumulants are:

$$\text{Cum}\,(Y_{\ell_1}, Y_{\ell_2}, Y_{\ell_3}) = \mathbb{E}(Y_{\ell_1} Y_{\ell_2} Y_{\ell_3}). \qquad (3.5)$$

And fourth-order cumulants:

$$\begin{aligned} \text{Cum}\,(Y_{\ell_1}, Y_{\ell_2}, Y_{\ell_3}, Y_{\ell_4}) &= \mathbb{E}(Y_{\ell_1} Y_{\ell_2} Y_{\ell_3} Y_{\ell_4}) - \mathbb{E}(Y_{\ell_1} Y_{\ell_2}) \mathbb{E}(Y_{\ell_3} Y_{\ell_4}) \qquad (3.6) \\ &\quad - \mathbb{E}(Y_{\ell_1} Y_{\ell_3}) \mathbb{E}(Y_{\ell_2} Y_{\ell_4}) - \mathbb{E}(Y_{\ell_2} Y_{\ell_3}) \mathbb{E}(Y_{\ell_1} Y_{\ell_4}). \end{aligned}$$

Taking $t = 0$ in equation (4.11) yields a set of restrictions on cumulants of factors and measurements:

$$\begin{aligned}\mathrm{Cum}\left(Y_{\ell_1},...,Y_{\ell_p}\right) &= \sum_{k=1}^{K}\left(\prod_{i=1}^{p}\lambda_{\ell_i,k}\right)\kappa_p\left(X_k\right)+\mathrm{Cum}\left(U_{\ell_1},...,U_{\ell_p}\right)\\ &= \sum_{k=1}^{K}\left(\prod_{i=1}^{p}\lambda_{\ell_i,k}\right)\kappa_p\left(X_k\right)+\delta_{\ell_1,...,\ell_p}\kappa_p\left(U_{\ell_1}\right), \qquad (3.7)\end{aligned}$$

where $\delta_{\ell_1,...,\ell_p}=1$ if $\ell_1=...=\ell_p$ and $=0$ otherwise, and $\kappa_p\left(Z\right)$ denotes the pth cumulant of a univariate random variable Z. If Z has zero mean,

$$\begin{aligned}\kappa_2\left(Z\right) &= \mathrm{Cum}\left(Z,Z\right)=\mathrm{Var}\left(Z\right)=\mathbb{E}Z^2,\\ \kappa_3\left(Z\right) &= \mathrm{Cum}\left(Z,Z,Z\right)=\mathbb{E}Z^3,\\ \kappa_4\left(Z\right) &= \mathrm{Cum}\left(Z,Z,Z,Z\right)=\mathbb{E}\left(Z^4\right)-3\mathbb{E}\left(Z^2\right)^2.\end{aligned}$$

Moment restrictions. It will prove convenient to write moment restrictions in matrix form, provided that the corresponding moments exist. Using (3.7) with $p=2$, second-order restrictions are equivalently rewritten as

$$\Sigma_Y=\Lambda\Lambda^T+\Sigma_U, \qquad (3.8)$$

where Σ_Y and Σ_U denote the variance-covariances matrices of Y and U.

Next, define the following matrices of third-order cumulants

$$\Gamma_Y\left(\ell\right)=\left[\mathrm{Cum}\left(Y_i,Y_\ell,Y_j\right);(i,j)\in\{1,...,L\}^2\right]\in\mathbb{R}^{L\times L},\quad \ell\in\{1...L\}. \qquad (3.9)$$

Third-order restrictions ($p=3$) imply that

$$\Gamma_Y\left(\ell\right)=\Lambda D_3\,\mathrm{diag}\left(\Lambda_\ell\right)\Lambda^T+\kappa_3\left(U_\ell\right)\mathrm{Sp}_{L,\ell}, \qquad (3.10)$$

where $\Lambda_\ell^T\in\mathbb{R}^{K\times 1}$ is the ℓth row of Λ, D_3 is the diagonal matrix with $\kappa_3\left(X_k\right)$ in the kth entry of the diagonal, and $\mathrm{Sp}_{L,\ell}$ is the $L\times L$ sparse matrix with only one 1 in position (ℓ,ℓ).

Let us also define the following matrices of fourth-order cumulants

$$\Omega_Y(\ell, m) = \left[\mathrm{Cum}(Y_i, Y_\ell, Y_m, Y_j); (i,j) \in \{1, ..., L\}^2\right] \in \mathbb{R}^{L \times L}, \quad (\ell, m) \in \overline{\Delta}_{L,2}. \tag{3.11}$$

Fourth-order restrictions ($p = 4$) imply that

$$\Omega_Y(\ell, m) = \Lambda D_4 \operatorname{diag}(\Lambda_\ell \odot \Lambda_m) \Lambda^T + \delta_{\ell m} \kappa_4(U_\ell) \mathrm{Sp}_{L,\ell}, \tag{3.12}$$

where D_4 is the diagonal matrix with $\kappa_4(X_k)$ in the kth entry of the diagonal, and $\odot$ is the Hadamard (element by element) matrix product.

3.2.3 Semiparametric identification

We now use restrictions (3.3) to derive necessary and sufficient conditions for the semiparametric identification of independent factor models. The next theorem, proved in the mathematical appendix, gives sufficient conditions for identification.

Theorem 3 ***(Sufficient conditions for semiparametric identification)*** *Let* (Λ, X, U) *be a regular representation. Let* $(\widetilde{\Lambda}, \widetilde{X}, \widetilde{U})$ *be an alternative regular representation. The following two propositions hold true:*

1. *Every column of* Λ *is a scalar multiple of a column of* $\widetilde{\Lambda}$.

2. *If the* $\frac{L(L-1)}{2} \times K$ *matrix* $Q(\Lambda) = [\lambda_{\ell 1}\lambda_{m1}, ..., \lambda_{\ell K}\lambda_{mK}; (\ell, m) \in \Delta_{L,2}]$, *where rows are stacked by increasing order of* (ℓ, m), $\ell < m$, *is full column rank, then* (Λ, X, U) *is identified.*

If factor variables are not normally distributed, then the matrix of factor loadings is identified whatever the number of factors. This result is well-known in the ICA literature, at least since Comon (1994). Moreover, it suffices that $\mathrm{rank}(Q(\Lambda)) = K$ for the distributions of factors and errors to be identified. A model with $\frac{L(L-1)}{2}$ factors is thus potentially identifiable when factors and errors are assumed independent instead of uncorrelated.

Theorem 3 is a straightforward generalisation of Eriksson and Koivunen's (2003) identification result for non noisy ICA. Proposition (i) of Theorem 3 follows from factor nongaussianity by a straightforward application of a result due to Kagan, Linnik and Rao (1973) that is stated in the mathematical appendix. Proposition (ii) of Theorem 3 easily follows from proposition (i).

We now show that the rank condition in proposition (ii) is generically necessary, that is: for a class of distribution functions dense in the set of continuous distribution functions. As far as we know, this is a new result. Let us first define the class of distributions divisible by a normal distribution.

Definition 4 ***(Distribution divisible by a normal)*** *Let X be a continuous random variable with density f and characteristic function φ. The distribution of X is divisible by a normal distribution if there exists $\sigma^2 > 0$ such that $\widetilde{\varphi}(t) = \varphi(t)\exp\left(\frac{\sigma^2 t^2}{2}\right)$ is the characteristic function of a random variable $\widetilde{X}$.*

The distribution of a variable X is divisible by a normal if and only if $X \overset{d}{=} \widetilde{X} + \mathcal{N}(0, \sigma^2)$, where $\mathcal{N}(0, \sigma^2)$ is a normal r.v. with mean 0 and variance σ^2. The set of distributions divisible by a normal is dense in the set of continuous distribution functions.[8]

For a representation (Λ, X, U) to be identifiable, the next theorem shows that either $Q(\Lambda)$ is full-column-rank or it is not, but then at least some of the factor and error variables must not be divisible by a normal distribution.

[8] Let X be a continuous random variable. Let $X_n = X + \mathcal{N}(0, \sigma^2)$. Then $X_n \overset{d}{\to} X$ when $\sigma^2 \to 0$.

Theorem 5 ***(Necessary condition for semiparametric identification)*** *Let (Λ, X, U) be a representation. Suppose that $Q(\Lambda)$ is not full-column-rank and that the distributions of factors and errors are divisible by normal distributions. Then (Λ, X, U) is not identifiable.*

We refer the reader to the mathematical appendix for a proof of Theorem 5. We show that under the assumptions of Theorem 5, for any representation (Λ, X, U) such that $Q(\Lambda)$ is not full-column-rank and the distributions of factors X and errors U are divisible by normal distributions, then one can construct another representation $(\widetilde{\Lambda}, \widetilde{X}, \widetilde{U})$ that is not equal to (Λ, X, U) up to a sign-permutation matrix and that still verifies the equality $\Lambda X + U \overset{d}{=} \widetilde{\Lambda}\widetilde{X} + \widetilde{U}$.

3.2.4 Parametric identification of factor loadings in the noise-free case $(U = 0)$

We here derive parametric identification results based on the first four moments of the data. The identification proofs are constructive, and will be used for estimation in the next section.

We first consider the case of factor models without errors. In this case, second, third and fourth-order restrictions (3.8), (3.10), (3.12) imply that matrix Λ satisfies simultaneously

$$
\begin{aligned}
\Sigma_Y &= \Lambda\Lambda^T, && (3.13)\\
\Gamma_Y(\ell) &= \Lambda D_3 \operatorname{diag}(\Lambda_\ell)\Lambda^T, \quad \ell \in \{1...L\}, && (3.14)\\
\Omega_Y(\ell, m) &= \Lambda D_4 \operatorname{diag}(\Lambda_\ell \odot \Lambda_m)\Lambda^T, \quad (\ell, m) \in \overline{\Delta}_{L,2}. && (3.15)
\end{aligned}
$$

Left and right-multiplying (3.13), (3.14) and (3.15) by $\Sigma_Y^{-1/2}$ and $\Sigma_Y^{-T/2}$, respectively, where $\Sigma_Y^{-1/2}\Sigma_Y\Sigma_Y^{-T/2} = I_K$, one obtains:

$$\begin{aligned}\Sigma_Y^{-1/2}\Gamma_Y(\ell)\,\Sigma_Y^{-T/2} &= VD_3\,\mathrm{diag}\,(\Lambda_\ell)\,V^T, \quad \ell \in \{1...L\},\\ \Sigma_Y^{-1/2}\Omega_Y(\ell,m)\,\Sigma_Y^{-T/2} &= VD_4\,\mathrm{diag}\,(\Lambda_\ell \odot \Lambda_m)\,V^T, \quad (\ell,m) \in \overline{\Delta}_{L,2}.\end{aligned}$$

where $V = \Sigma_Y^{-1/2}\Lambda$ is orthogonal; that is: $VV^T = I_K$. Therefore, V solves a joint diagonalization problem. Theorem 6 below gives conditions for the solution to this joint diagonalization problem to be unique.

Theorem 6 ***(Parametric identification in the noise-free case)*** *Assume (i) $U = 0$, (ii) $K \leq L$ and (iii) Λ has rank K.*

If (iv) at most one factor variable has zero kurtosis excess, then factor loadings are identified from second and fourth-order moment restrictions (3.13) and (3.15).

If (iv') at most one factor variable has zero skewness, then factor loadings are identified from second and third-order moment restrictions (3.13) and (3.14).

If (iv") for any couple of factors indices (k,k'), $(\kappa_3(X_k), \kappa_3(X_{k'}), \kappa_4(X_k), \kappa_4(X_{k'})) \neq 0$, then factor loadings are identified from second, third and fourth-order moment restrictions (3.13), (3.14) and (3.15).

The proof is in the mathematical appendix. Theorem 6 shows that high order moments are a source of identification in noise-free factor models. This insight has been widely used in the ICA literature. For instance, Cardoso and Souloumiac (1993) use restrictions (3.13) and (3.15) as the basis of their JADE algorithm. In usual applications of ICA methods, factors are thought to be symmetric. For this reason, third-order information is *a priori* neglected in this literature. However, there is no strong argument in favour of discarding third-order moments of the data in econometrics. It is yet true that the variables of interest are often transformed to make them as much Gaussian as possible. For instance, by taking the logarithm of income, one obtains a distribution which is close to being normal, at least as far as the skewness and kurtosis are concerned. However, there can still be enough non normality in the multivariate distribution of the data for factor loadings and factor moments to be well identified. The application in Section 3.5 will provide an illustration of this remark.

3.2.5 Parametric identification of error moments

In the "noisy" case ($U \neq 0$), the previous identification results apply, provided that the first moments of error variables are identified. We here give conditions under which these moments are identified. Two cases are distinguished, depending on whether all fourth-order cumulants of factors are zero or not.

First case: all factor distributions are kurtotic

Let Ω_Y be the $\frac{L(L+1)}{2} \times \frac{L(L-1)}{2}$ matrix of *all* fourth-order cumulants of the data, defined by

$$\Omega_Y = \left[\text{Cum}\left(Y_i, Y_j, Y_\ell, Y_m\right); (i,j) \in \overline{\Delta}_{L,2}, (\ell, m) \in \Delta_{L,2}\right] \in \mathbb{R}^{\frac{L(L+1)}{2} \times \frac{L(L-1)}{2}}. \quad (3.16)$$

The rows of Ω_Y are indexed by $(i,j) \in \overline{\Delta}_{L,2}$ and the columns are indexed by $(\ell, m) \in \Delta_{L,2}$, i.e. $(i,j) \in \{1, ..., L\}^2$, $i \leq j$, and $(\ell, m) \in \{1, ..., L\}^2$, $\ell < m$. The factor structure implies that

$$\Omega_Y = \overline{Q} D_4 Q^T, \quad (3.17)$$

where

$$Q \equiv Q(\Lambda) = [\lambda_{\ell k}\lambda_{mk}; (\ell, m) \in \Delta_{L,2}, k \in \{1, ..., K\}] \in \mathbb{R}^{\frac{L(L-1)}{2} \times K}, \quad (3.18)$$

$$\overline{Q} \equiv \overline{Q}(\Lambda) = [\lambda_{\ell k}\lambda_{mk}; (\ell, m) \in \Delta_{L,2}, k \in \{1, ..., K\}] \in \mathbb{R}^{\frac{L(L+1)}{2} \times K}. \quad (3.19)$$

We first show that, under the assumption that all factors have kurtosis excess, it suffices that Q be full column rank for the first four error moments to be identified from the first four moments of the data.

Lemma 7 *Assume that (i)* $K \leq \frac{L(L-1)}{2}$*, (ii)* Q *has rank* K *and (iii) factor variables have non zero kurtosis excess. Then the following propositions hold true.*

1. *Matrix* Ω_Y *has rank* K*.*

2. *Let* $\overline{C} \in \mathbb{R}^{\frac{L(L+1)}{2} \times \left(\frac{L(L+1)}{2} - K\right)}$ *be a basis of the null space of* Ω_Y^T*; that is: the columns of* $\overline{C}$ *are linearly independent and* $\Omega_Y^T \overline{C} = 0$*. The first four moments of* U_ℓ*,* $\ell \in \{1, ..., L\}$*, satisfy the linear restrictions:*

$$\overline{C}^T \operatorname{vech}(\Sigma_Y) = \sum_{\ell=1}^{L} \operatorname{Var}(U_\ell)\, \overline{C}_{(\ell,\ell)}, \tag{3.20}$$

$$\overline{C}^T \operatorname{vech}(\Gamma_Y(\ell)) = \kappa_3(U_\ell)\, \overline{C}_{(\ell,\ell)}, \tag{3.21}$$

$$\overline{C}^T \operatorname{vech}(\Omega_Y(\ell,\ell)) = \kappa_4(U_\ell)\, \overline{C}_{(\ell,\ell)}, \tag{3.22}$$

where $\overline{C}^T_{(\ell,\ell)}$ *denotes the* (ℓ,ℓ)*th row of* $\overline{C}$*, when the* $\frac{L(L+1)}{2}$ *rows of* $\overline{C}$ *are indexed by* $\overline{\Delta}_{L,2}$*, and where* vech *is the linear matrix operator stacking all* $\frac{L(L+1)}{2}$ *non redundant elements of a symmetric matrix.*[9]

3. *Matrix* $\left[\overline{C}_{(1,1)}, ..., \overline{C}_{(L,L)}\right]$ *is full rank and* $\operatorname{Var}(U_\ell)$*,* $\kappa_3(U_\ell)$ *and* $\kappa_4(U_\ell)$ *are uniquely defined by identification restrictions (3.20), (3.21) and (3.22).*

The proof is in Section C.1.4 of the mathematical appendix. The following theorem then follows straightforwardly.

Theorem 8 ***(Sufficient conditions for parametric identification when*** $K \leq L$***)*** *Assume that (i)* $K \leq \min\left\{L, \frac{L(L-1)}{2}\right\}$*, (ii)* Λ *is full column rank, (iii)* Q *has rank* K*, and (iv) factor variables have non zero kurtosis excess. Then, factor loadings are identified from second and fourth-order moments.*

[9]Let $A = [a_{ij}]$ be a $L \times L$ matrix. Then $\operatorname{vech}(A) = [a_{ij}; i \leq j] \in \mathbb{R}^{\frac{L(L+1)}{2} \times 1}$, ordering couples (i,j) by increasing order.

Theorem 6 shows that the maximal number of factors for which Λ can be identified (up to column sign and permutation) is $K = 1$ if $L = 2$, and $K = L$ if $L \geq 3$.

Second case: all factor distributions are either skewed or kurtotic

We now consider the problem of identifying factor loadings, in the "noisy" factor model, when some or all factor distributions have zero kurtosis excess.

Let

$$\Omega_Y(j) = \left[\mathrm{Cum}(Y_i, Y_j, Y_\ell, Y_m); i \in \{1, ..., L\}, (\ell, m) \in \Delta_{L,2}\right] \in \mathbb{R}^{L \times \frac{L(L-1)}{2}}. \quad (3.23)$$

The rows of Ω_Y are indexed by $i \in \{1, ..., L\}$ and the columns are indexed by $(\ell, m) \in \Delta_{L,2}$, i.e. $\ell < m$. The factor structure implies that

$$\Omega_Y(j) = \Lambda \,\mathrm{diag}(\Lambda_j)\, D_4 Q^T. \quad (3.24)$$

Let also Γ_Y be the $L \times \frac{L(L-1)}{2}$ matrix of third-order cumulants of the data defined by

$$\Gamma_Y = \left[\mathrm{Cum}(Y_i, Y_\ell, Y_m); i \in \{1, ..., L\}, (\ell, m) \in \Delta_{L,2}\right] \in \mathbb{R}^{L \times \frac{L(L-1)}{2}}, \quad (3.25)$$

The rows of Γ_Y are indexed by $i \in \{1, ..., L\}$ and the columns are indexed by $(\ell, m) \in \Delta_{L,2}$, i.e. $(\ell, m) \in \{1, ..., L\}^2$, $\ell < m$. The factor structure implies that

$$\Gamma_Y = \Lambda D_3 Q^T. \quad (3.26)$$

Lastly, let Ξ_Y be the $L \times \frac{L(L-1)(L+1)}{2}$ matrix of *all* third and fourth-order cumulants of the data, obtained by stacking matrices Γ_Y, $\Omega_Y(1)$,..., $\Omega_Y(L)$ columnwise:

$$\Xi_Y = \left[\Gamma_Y, \Omega_Y(1), ..., \Omega_Y(L)\right]. \quad (3.27)$$

We first establish a set of linear restrictions on error moments.

Lemma 9 *Assume that (i)* $K \leq \min\left\{L, \frac{L(L-1)}{2}\right\}$*, (ii)* Λ *and* Q *are full column rank* K *and (iii) every factor distribution is either skewed or kurtotic. Then the following propositions hold true.*

1. Ξ_Y *has rank* K*.*

2. *Let* $C \in \mathbb{R}^{L\times(L-K)}$ *be a basis of the null space of* Ξ_Y^T*; that is: the columns of* C *are linearly independent, and* $\Xi_Y^T C = 0$*. Let* C_ℓ^T *denote the* ℓ*th row of* C*. The second, third and fourth-order moments of* U_ℓ*, for all* $\ell \in \{1, ..., L\}$*, satisfy the linear restrictions:*

$$C^T \begin{pmatrix} \mathbb{E}(Y_1 Y_\ell) \\ \vdots \\ \mathbb{E}(Y_L Y_\ell) \end{pmatrix} = \operatorname{Var}(U_\ell)\, C_\ell, \tag{3.28}$$

$$C^T \begin{pmatrix} \mathbb{E}(Y_1 Y_\ell^2) \\ \vdots \\ \mathbb{E}(Y_L Y_\ell^2) \end{pmatrix} = \kappa_3(U_\ell)\, C_\ell. \tag{3.29}$$

and

$$C^T \begin{pmatrix} \mathbb{E}(Y_1 Y_\ell^3) - 3\mathbb{E}(Y_1 Y_\ell)\,\mathbb{E}(Y_\ell^2) \\ \vdots \\ \mathbb{E}(Y_L Y_\ell^3) - 3\mathbb{E}(Y_L Y_\ell)\,\mathbb{E}(Y_\ell^2) \end{pmatrix} = \kappa_4(U_\ell)\, C_\ell. \tag{3.30}$$

Lemma 9 is not sufficient to identify error moments if $K = L$, as in this case matrix C is zero. We thus require additional assumptions on Λ.

Lemma 10 *Assume, in addition to the conditions of Lemma 9, that (i)* $K \leq L-1$*, and (ii) every submatrix of* Λ *made of a selection of* $L-1$ *rows has rank* K*. Then, no column of* C *is nil (*$C_\ell \neq 0$*,* $\forall \ell$*) and* $\operatorname{Var}(U_\ell)$*,* $\kappa_3(U_\ell)$ *and* $\kappa_4(U_\ell)$ *are identified.*

The proofs are in Section C.1.5 of the mathematical appendix. The following theorem then follows immediately.

Theorem 11 ***(Sufficient conditions for parametric identification when*** $K \leq L-1$***)*** *Assume that (i)* $K \leq L-1$*, (ii) every submatrix of* Λ *made of a selection of* $L-1$ *rows has rank* K*, (iii) matrix* Q *has rank* K*, (iv) every factor distribution is either skewed or kurtotic. Then, factor loadings are parametrically identified from second, third and fourth-order moments.*

As a special case, if all factors are skewed then factor loadings are parametrically identified from second and third-order moments.

Corollary 12 ***(Sufficient conditions for parametric identification from second and third-order moments when*** $K \leq L-1$***)*** *Assume that (i)* $K \leq L-1$*, (ii) every submatrix of* Λ *made of a selection of* $L-1$ *rows has rank* K*, (iii) matrix* Q *has rank* K*, and (iv) all factor distributions are skewed. Then, factor loadings are parametrically identified from second and third-order moments.*

For example, consider the case of $L=2$ and $K=1$ and factor X_1 has a non symmetric distribution:

$$\begin{cases} Y_1 = \lambda_{11} X_1 + U_1, \\ Y_2 = \lambda_{21} X_1 + U_2, \end{cases}$$

and $\mathbb{E}\left(X_1^3\right) \neq 0$. One easily finds:

$$\begin{aligned} \lambda_{11} &= \sqrt{\mathbb{E}(Y_1 Y_2) \frac{\mathbb{E}(Y_1 Y_1 Y_2)}{\mathbb{E}(Y_1 Y_2 Y_2)}}, \\ \lambda_{21} &= \sqrt{\mathbb{E}(Y_1 Y_2) \frac{\mathbb{E}(Y_1 Y_2 Y_2)}{\mathbb{E}(Y_1 Y_1 Y_2)}}. \end{aligned}$$

Interestingly, the ratio of the two factor loadings is then

$$\frac{\lambda_{21}}{\lambda_{11}} = \frac{\mathbb{E}(Y_1 Y_2 Y_2)}{\mathbb{E}(Y_1 Y_1 Y_2)}. \tag{3.31}$$

Replacing expectations by sample means, we obtain a consistent estimator of $\frac{\lambda_{21}}{\lambda_{11}}$ which is the coefficient of the regression Y_2 on Y_1 with no intercept, by 2SLS, using

$Y_1 Y_2$ as an instrument for Y_1. This is the estimator of the measurement error model that was proposed by Geary (1942). Interestingly, the quasi-JADE estimator that we shall propose in the next section also satisfies equation (3.31). The estimators introduced in this chapter can thus be interpreted as a generalization of Geary's IV estimator.

3.3 Estimation

We start by discussing the issue of estimating the number of factors.

3.3.1 Estimating the number of factors K

Estimating K when $K \leq \frac{L(L-1)}{2}$ and all factors are kurtotic. Assuming that Q is full column rank and that factor variables show kurtosis excess, then matrix Ω_Y has rank K (see Lemma 7). For any i.i.d. sample, let $\widehat{\Omega}_Y$ be the empirical counterpart of Ω_Y, obtained by replacing expectations by sample means. We use the sequential testing procedure developed by Robin and Smith (2000) to estimate the rank of Ω_Y.[10]

Monte Carlo simulations show that the rank test, applied to matrix Ω_Y alone, suffers from substantial size distortions (see the simulations in the next section). Assuming $K \leq L$, the factor structure provides additional rank conditions that can be used to improve the test's properties. We propose the following refinement.

Consider matrices $\Omega_Y(\ell, m)$ for all $(\ell, m) \in \Delta_{L,2}$ ($\ell < m$). They satisfy the restrictions:

$$\Omega_Y(\ell, m) = \Lambda D_4 \operatorname{diag}(\Lambda_\ell \odot \Lambda_m) \Lambda^T.$$

Let $w = (w_{1,2}, ..., w_{L-1,L})$ be a vector of $\frac{L(L-1)}{2}$ positive weights. Then,

$$\Omega_{Y,w} \equiv \sum_{(\ell,m)\in\Delta_{L,2}} w_{\ell,m} \Omega_Y(\ell, m) = \Lambda D_4 \operatorname{diag}\left(Q^T w\right) \Lambda^T. \tag{3.32}$$

[10]Robin and Smith's rank test is described in Appendix C.4.

As no column of Q is identically zero, matrix $\Omega_{Y,w}$ has rank K for almost all w.

It seems natural to weight cumulant matrices more if they are more precise. We therefore suggest to choose $w_{\ell,m}$ equal to the inverse of the average of asymptotic variances of the components of the empirical estimate $\widehat{\Omega}_Y(\ell,m)$ of $\Omega_Y(\ell,m)$. These variances can be computed by standard bootstrap.

Estimating K when $K \leq L$ and all factors are skewed or kurtotic. Assuming that Λ and Q are full column rank and that factor variables have non zero skewness, then matrix Γ_Y has rank K (see Lemma 9). One can thus apply the rank test to any root-N estimator $\widehat{\Gamma}_Y$.

More generally, one can use the following version of the rank test, which uses third and fourth-order information of the data. Lemma 9 shows that, assuming that Λ and Q are full column rank (so that $K \leq L$) and that each factor distribution is either skewed or kurtotic, matrix Ξ_Y has rank K. One can thus test the rank of any root-N consistent estimator $\widehat{\Xi}_Y$.

Again, one can refine the test and account for moments' variability, as matrix

$$\Xi_{Y,w} = \Gamma_Y + \sum_{j=1}^{L} w_j \Omega_Y(j) = \Lambda \left[D_3 + D_4 \operatorname{diag}\left(\Lambda^T w\right)\right] Q^T \tag{3.33}$$

has rank K, for almost all weight $w = (w_1, ..., w_L)^T \in \mathbb{R}^L$.[11] We suggest to set w_j equal to the average of the variances of the components of $\widehat{\Gamma}_Y$ divided by the average of the variances of the components of $\widehat{\Omega}_Y(j)$.

[11] This is because the set

$$\{w \in \mathbb{R}^L, \kappa_3(X_k) + \kappa_4(X_k)\left(\sum_{j=1}^{L} w_j \lambda_{jk}\right) = 0\}$$

has measure zero in $\mathbb{R}^L$, for all $k = 1...K$.

3.3.2 Cardoso and Souloumiac's JADE procedure

Assuming no noise, factor loadings satisfy the following system of matrix equations:

$$\Omega_Y(\ell,m) = \Lambda D_4(\ell,m)\Lambda^T, \quad (\ell,m)\in\overline{\Delta}_{L,2}, \tag{3.34}$$

$$\Sigma_Y = \Lambda\Lambda^T, \tag{3.35}$$

for diagonal matrices $D_4(\ell,m)$ (see Section 3.2.4).

In an influential paper, Cardoso and Souloumiac (1993) propose the following procedure to estimate Λ using this system of restrictions.

1. "Whiten" the data, i.e. compute $\widetilde{Y} = P^{-}Y$, where P is a $L\times L$ such that $PP^T = \Sigma_Y$ (for example, a Cholesky decomposition) and A^{-} is a generalized inverse of P, e.g. $P^{-} = \left[P^TP\right]^{-1}P^T$.

2. Compute $\Omega_{\widetilde{Y}}(\ell,m)$, for all $(\ell,m)\in\overline{\Delta}_{L,2}$. These matrices satisfy the restrictions:

   $$V^T\Omega_{\widetilde{Y}}(\ell,m)V = D_4(\ell,m),$$

   where $V = P^{-}\Lambda$ is an orthonormal matrix of dimensions K.

3. Compute V as an orthonormal matrix minimizing the sum of squares of the off-diagonal elements of matrices $V^T\Omega_{\widetilde{Y}}(\ell,m)V$. Cardoso and Souloumiac (1993) develop a simple and efficient algorithm to perform this optimisation (using Jacobi rotations), that is detailed in Section C.2 of the Appendix.[12]

To apply this algorithm on a sample $\{Y_1,...,Y_N\}$ of i.i.d. observations, replace expectations by sample means. The theoretical restrictions then only hold approximately but the joint diagonalisation algorithm still delivers an orthonormal matrix $\widehat{V}$

[12]A MATLAB code of the JADE algorithm is available on Cardoso's web page: http://www.tsi.enst.fr/ cardoso/Algo/Jade/jadeR.m.

such that all matrices $\widehat{V}^T \widehat{\Omega}_{\widetilde{Y}}(\ell, m) \widehat{V}$ are approximately diagonal. An estimate of Λ is then simply obtained as $\widehat{\Lambda} = \widehat{P}\widehat{V}$. Cardoso and Souloumiac (1993) call JADE this empirical procedure (Joint Approximate Diagonalisation of Eigenmatrices).

The JADE algorithm has several attractive properties. As it uses *all* fourth-order cumulants of the data, it is much less sensitive to spectrum degeneracy than single diagonalization algorithms (see Cardoso, 1999). Moreover, the cost to pay for these efficiency gains is reasonable, as Jacobi rotation-based algorithms are fast to converge. Lastly, JADE is *equivariant* in the sense that changing Y into WY, for any invertible matrix W, changes $\widehat{\Lambda}$ into $W\widehat{\Lambda}$.

3.3.3 Asymptotic theory for JADE

As far as we know, there is no derivation of the asymptotic properties of JADE in the ICA literature. This section aims at filling this gap.

To proceed, let $\widehat{A}_1, ..., \widehat{A}_J$ be root-N consistent and asymptotically normal estimators of J symmetric $K \times K$ matrices $A_1, ..., A_J$. Construct $\widehat{A} = \left[\widehat{A}_1, ..., \widehat{A}_J\right]$ and $A = [A_1, ..., A_J]$ by concatenation. Let $\mathbb{V}_A$ be the asymptotic variance of $N^{1/2}\,\mathrm{vec}(\widehat{A})$. The JADE estimator is

$$\widehat{V} = \arg\min_{V \in \mathcal{O}_K} \sum_{j=1}^{J} \mathrm{off}(V^T \widehat{A}_j V),$$

where $\mathrm{off}(M) = \sum_{i \neq j} m_{ij}^2$ and $\mathcal{O}_K$ is the set of orthonormal $K \times K$ matrices.

Assume that there exists $V \in \mathcal{O}_K$ such that, for all $j = 1, ..., J$, $V^T A_j V = D_j$, where D_j is the diagonal matrix with diagonal elements $d_{j1}, ..., d_{jK}$. Define the $K \times K$ matrices:

$$R(D_j) = \left[\frac{(d_{jk} - d_{jm})}{\sum_{j'=1}^{J} (d_{j'k} - d_{j'm})^2}; \quad (k, m) \in \{1, ..., K\}^2 \right].$$

Lastly, let W be the following $K^2 \times JK^2$ matrix:

$$W = [\text{diag}(\text{vec}(R(D_1))), ..., \text{diag}(\text{vec}(R(D_J)))] .$$

We show the following result in Appendix C.3.

Theorem 13 *Assume that* $\sum_{j=1}^{J}(d_{jk} - d_{jm})^2 \neq 0$ *for all* $k \neq m$. *Then*

$$N^{1/2}\left(\text{vec}(\widehat{V}) - \text{vec}(V)\right) \underset{N\to\infty}{\overset{L}{\to}} \mathcal{N}\left(0, \mathbb{V}_V\right),$$

where:

$$\mathbb{V}_V = (I_K \otimes V)W(I_J \otimes V^T \otimes V^T)\mathbb{V}_A(I_J \otimes V \otimes V)W^T(I_K \otimes V^T). \tag{3.36}$$

Let us consider the particular case of $J = 1$. In this case, (3.36) yields the well-known expression for the variance-covariance matrix of the eigenvectors of a symmetric matrix (*e.g.* Anderson, 1963). The diagonal coefficients of matrix W are equal to $1/(d_{1k} - d_{1m})$, for $k \neq m$. The variance of eigenvectors thus increases when two eigenvalues of A_1 get close to each other.

In the general case of more than one matrix ($J > 1$), a precise estimation requires $\sum_j (d_{jk} - d_{jm})^2$ not to be close to zero, for all indices (k, m). Now, the larger J and the less likely it is that $d_{jk} = d_{jm}$ for all j. Cardoso (1999) already noted that joint diagonalisation algorithms seemed less sensitive to the presence of multiple roots than usual diagonalisation techniques.[13] Theorem 13 allows to better understand the conditions granting a good precision.

Basing identification on fourth-order moments, indices are $j = (\ell, m)$, and matrices A_j and D_j are of the form: $\Omega_Y(\ell, m)$ and $D_4 \,\text{diag}\,(\Lambda_\ell \odot \Lambda_m)$, respectively. If there exist k, k' such that $d_{jk} = d_{jk'}$ for all j, it must be that

$$\lambda_{\ell k}\lambda_{mk}\kappa_4(X_k) = \lambda_{\ell k'}\lambda_{mk'}\kappa_4(X_{k'}),$$

[13] See also the asymptotic distribution of estimators of *common* principal components derived by Flury (1984, 1986).

for all ℓ, m. This cannot happen if at most one factor has zero kurtosis excess and the columns of Λ are not proportional to each other.

This result is not surprising, as the variance of eigenvector estimators blows up when the model is not identified. Non identification arises in PCA when the variance of the vector of measurements has multiple eigenvalues (there are then obviously many possible choices for a basis of the corresponding eigenspace). In ICA this happens when two columns of the matrix of factor loadings are proportional or when factor distributions lack skewness and/or kurtosis excess. We shall produce Monte-Carlo simulations to illustrate this point.

Practical remark. In practice, we do *not* recommend to use formula (3.36) to compute standard errors. Instead, we suggest to compute standard errors or coverage intervals by standard bootstrap (with appropriate recentering). The reason is that the expression in (3.36) involves second moments of third and/or fourth-order moments of the data, which are difficult to estimate precisely. Our simulations show extremely imprecise estimates of matrix $\mathbb{V}_A$, even with very large samples (more than $10,000$ observations). In contrast, bootstrap provides good approximations of the true variance-covariance matrix of the JADE estimator.

3.3.4 The quasi-JADE algorithm

The JADE algorithm is only valid if there is no noise ($U = 0$). However, Lemmas 7 and 9 show that the first four moments of error variables can be estimated independently of factor loadings. Given error moments, one can then apply JADE.

We call quasi-JADE the following procedure.

1. Estimate matrices C and/or $\overline{C}$ of Lemmas 7 and 9. These matrices are easily obtained by Singular Value Decomposition of matrices Ω_Y and Ξ_Y.

2. Estimate error variances $\mathrm{Var}(U_\ell)$, third-order cumulants $\kappa_3(U_\ell)$ and/or fourth-order cumulants $\kappa_4(U_\ell)$ using the restrictions in Lemmas 7 and 9. One should impose the non negativity of error variances, as well as the positive semi-definiteness of matrix $\Sigma_Y - \Sigma_U$.

3. Proceed to the joint diagonalisation (i.e. steps 2 and 3 of the JADE algorithm) of matrices

$$P^- \left[\Gamma_Y(\ell) - \kappa_3(U_\ell)\,\mathrm{Sp}_{L,\ell}\right] P^{-T} \text{ and/or } P^- \left[\Omega_Y(\ell, m) - \delta_{\ell m}\kappa_4(U_\ell)\,\mathrm{Sp}_{L,\ell}\right] P^{-T},$$

   where P is a full column rank $L \times K$ matrix such that $\Sigma_Y - \Sigma_U = PP^T$. We suggest to compute P as the first K columns of the Cholesky decomposition of matrix $\Sigma_Y - \Sigma_U$. Let V be the orthonormal matrix of joint eigenvectors. Then $\Lambda = PV$.

4. Estimate factor cumulants $\kappa_3(X_k)$ and $\kappa_4(X)$ by OLS from restrictions:

$$\begin{aligned}
\left[V^T P^- \left[\Gamma_Y(\ell) - \kappa_3(U_\ell)\,\mathrm{Sp}_{L,\ell}\right] P^{-T} V\right]_{k,k} &= \lambda_{\ell k}\kappa_3(X_k), \\
\left[V^T P^- \left[\Omega_Y(\ell, m) - \delta_{\ell m}\kappa_4(U_\ell)\,\mathrm{Sp}_{L,\ell}\right] P^{-T} V\right]_{k,k} &= \lambda_{\ell k}\lambda_{mk}\kappa_4(X_k),
\end{aligned}$$

   where $[A]_{i,j}$ denotes the (i, j) entry of matrix A.

Quasi-JADE is only marginally more complicated to implement than JADE,[14] and is almost as fast to converge. However, allowing for errors has a cost. Whereas JADE is equivariant, quasi-JADE is not. In practice, we suggest to apply quasi-JADE to normalized measurements, by dividing each Y_ℓ by its standard error.

Efficiency improvements. As the original JADE algorithm, quasi-JADE is obviously not efficient. First, it operates a sequence of minimum distance estimations

[14] GAUSS codes for quasi-JADE are given in the Appendix to this chapter.

instead of estimating all parameters jointly. Second, it does not use the optimal metric in these minimum distance problems. Third, it does not use all the structural moment restrictions. For example, the diagonal matrices $D_4(\ell, m)$ in (3.12) are related to Λ but we do not use this restriction.

A natural alternative to our approach would be to use all cumulant restrictions (3.8)-(3.10)-(3.12) in estimation. However, these restrictions are highly nonlinear polynomial equations, which are difficult to solve using standard gradient algorithms or any other general-purpose solving technique. We shall make this point more precise in the simulation section. Second, there is considerable evidence that the optimal metric does not outperform the identity metric in finite samples (see Altonji and Segal, 1994, 1996).

Nevertheless, there is scope for efficiency improvements. For instance, one can use Generalised Least Squares instead of OLS to estimate error cumulants in Step 2 of the algorithm. Likewise, one can weight the matrices to diagonalise in Step 3. Weights can be some measure of estimation precision, as outlined in 3.3.1. In simulations, we found that this method yielded little efficiency gains. On real data, however, we found slightly different results that we shall present in section 3.5. Note that this weighting procedure is *ad hoc*. Issues regarding the optimal weighting of cumulant matrices, based on asymptotic results such as (3.36), are left for future research.

3.4 Monte-Carlo simulations

In this section, we study the finite-sample properties of our estimators by numerical simulations. We first consider the estimation of Λ given the true value of K, the number of factors. Then, we present simulations for estimating K.

N	500	1000	5000	10000
λ_{11}	2.03 (.28)	2.03 (.17)	2.01 (.09)	2.01 (.06)
λ_{21}	.95 (.23)	.99 (.14)	1.00 (.07)	1.00 (.05)
λ_{31}	.95 (.23)	.99 (.15)	.99 (.07)	1.00 (.05)
λ_{12}	.98 (.23)	.98 (.15)	1.00 (.06)	1.00 (.05)
λ_{22}	2.05 (.27)	2.03 (.19)	2.01 (.08)	2.01 (.07)
λ_{32}	.97 (.23)	.98 (.17)	1.00 (.06)	1.00 (.05)
λ_{13}	.97 (.23)	.98 (.15)	.99 (.06)	1.00 (.05)
λ_{23}	.97 (.23)	.98 (.16)	1.00 (.06)	1.00 (.05)
λ_{33}	2.06 (.27)	2.02 (.19)	2.01 (.09)	2.00 (.05)
$\text{Var}(U_1)$	.77 (.59)	.87 (.43)	.96 (.20)	.98 (.16)
$\text{Var}(U_2)$	.76 (.57)	.87 (.43)	.98 (.20)	.98 (.17)
$\text{Var}(U_3)$	.74 (.56)	.86 (.42)	.96 (.20)	.98 (.16)

Table 3.1: Quasi-JADE based on the 2nd, 3rd and 4th moment restrictions of Lemma 7 (log-normal factors, standard normal errors, $\Lambda = \Lambda_1$)

3.4.1 Estimation of factor loadings

Table 3.1 presents the results of 1000 simulations of the model with centered and standardized log-normal factors, standard normal errors and Λ equal to

$$\Lambda_1 \equiv \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix}.$$

Results are given for various sample sizes N. Monte Carlo standard deviations of estimates are given between brackets. Estimation is based on all second, third and fourth-order moments of the data and uses the restrictions of Lemma 7.

Table 3.1 shows some evidence of finite sample bias. However, in general the bias is small and rapidly decreases as N increases. By comparison, small sample biases are much larger and convergence is much slower for empirical cumulants. Table 3.2 shows the means and standard deviations of the empirical skewness and kurtosis of a standardised log-normal variate, for various sample sizes.[15] The striking contrast between Tables 3.1 and 3.2 suggests that our algorithm does a good job at extracting

[15]Means and variances were computed from 1000 independent drawings, for each sample size N.

N	500	1000	5000	10000	∞
κ_3	4.51 (1.98)	5.01 (2.36)	5.73 (2.65)	5.89 (2.02)	6.18
κ_4	36.1 (38.4)	48.6 (62.4)	77.0 (132.3)	83.3 (104.7)	110.9

Table 3.2: Empirical skewness and excess kurtosis of a log-normal random variable

the relevant information from higher-order moments of the data, while being relatively immune to the imprecision of their estimation in finite samples.

We then study the robustness of our algorithm to the magnitude of the noise. In Table 3.3, we compare the performance of quasi-JADE to standard JADE. We run the simulations with normal errors, log-normal factors, a sample size of $N = 1000$ and $\Lambda = \Lambda_1$. The standard deviation of errors can take four values: 0.1, 0.5, 1 and 2. We see that the performances of quasi-JADE deteriorate as the signal-to-noise ratio decreases. However, the bias remains limited even for rather large error variances. By comparison, JADE presents much larger biases. In particular, while our quasi-JADE algorithm yields consistent estimates of factor loadings, the inconsistency of ordinary (noise-free) ICA is severe, even when the magnitude of the error variances is not especially large (for example for a variance of one; which here implies that $\operatorname{Var}(U_\ell)/\operatorname{Var}(Y_\ell) = 20\%$).

Next, we compare these results with Minimum Distance based on the complete set of moment restrictions. The estimation is based on second and fourth-order restrictions:

$$\begin{cases} \Sigma_Y = \Lambda\Lambda^T + \Sigma_U, \\ \Omega_Y = \overline{Q} D_4 Q^T, \end{cases}$$

where Ω_Y is the 6×3 matrix of fourth-order cumulants of Y given by (3.16) and where Q and $\overline{Q}$ depend on Λ.

In all the simulations that we performed, Full Minimum Distance proved to be highly unstable. Maximisation with respect to the whole set of parameters (Λ, Σ_U, D_4)

JADE

$\text{Var}(U_\ell)$	.01	.25	1	4
$\widehat{\lambda}_{11}$	2.00 (.07)	2.11 (.08)	2.36 (.12)	2.81 (.46)
λ_{21}	1.00 (.11)	1.00 (.12)	.95 (.24)	.72 (.86)
λ_{31}	1.00 (.11)	1.03 (.14)	1.08 (.22)	1.05 (.77)
λ_{12}	1.00 (.11)	1.00 (.12)	.97 (.24)	.78 (.86)
λ_{22}	2.00 (.07)	2.11 (.07)	2.37 (.12)	2.86 (.32)
λ_{32}	1.00 (.12)	1.03 (.13)	1.08 (.22)	1.08 (.76)
λ_{13}	1.00 (.11)	.87 (.13)	.61 (.20)	.16 (.69)
λ_{23}	1.00 (.11)	.87 (.12)	.62 (.20)	.15 (.67)
λ_{33}	2.00 (.08)	2.02 (.09)	2.13 (.16)	2.52 (.43)

quasi-JADE

$\text{Var}(U_\ell)$	.01	.25	1	4
λ_{11}	1.98 (.12)	2.01(.13)	2.03 (.17)	2.02 (.44)
λ_{21}	1.00 (.15)	.99 (.12)	.99 (.14)	.95 (.31)
λ_{31}	1.00 (.16)	.99 (.13)	.99 (.15)	.95 (.32)
λ_{12}	1.00 (.16)	.99 (.13)	.98 (.15)	.97 (.33)
λ_{22}	1.97 (.11)	2.02 (.11)	2.03 (.19)	2.02 (.41)
λ_{32}	.99 (.16)	.99 (.13)	.98 (.17)	.97 (.32)
λ_{13}	1.00 (.16)	1.00 (.14)	.98 (.15)	.96 (.32)
λ_{23}	1.00 (.16)	1.00 (.13)	.98 (.16)	.96 (.32)
λ_{33}	1.98 (.11)	2.02 (.11)	2.02 (.19)	2.01 (.42)
$\text{Var}(U_1)$	.04 (.11)	.18 (.22)	.87 (.43)	3.77 (.98)
$\text{Var}(U_2)$	.04 (.11)	.17 (.23)	.87 (.43)	3.77 (.94)
$\text{Var}(U_3)$	.04 (.11)	.17 (.22)	.86 (.42)	3.77 (.97)

Table 3.3: Robustness to noise of JADE and quasi-JADE (log-normal factors, standard normal errors, $N = 1000$, $\Lambda = \Lambda_1$)

converged (numerically) in none of the cases that we considered. To obtain a more stable algorithm, admittedly at the cost of lower efficiency, we treated the coefficients of D_4 as nuisance parameters. Precisely, we minimised the Minimum Distance norm, evaluated at $(\Lambda, \Sigma_U, D_4(\Lambda))$, with respect to (Λ, Σ_U) alone and where $D_4(\Lambda)$ is such that

$$\text{vec}\,[D_4(\Lambda)] = \left(Q \otimes \overline{Q}\right)^{-} \text{vec}(\Omega_Y).$$

Note that using the optimal metric to estimate $D_4(\Lambda)$ from restriction $\Omega_Y = QD_4\overline{Q}^T$ given Λ yielded even greater instability. Incorporating third-order moment restrictions into the algorithm had the same effect.

Table 3.4 presents simulation results with log-normal factors, normal errors and $\Lambda = \Lambda_1$. Results are presented conditional on numerical convergence.[16] Starting values were chosen equal to the true parameters. First, we find that, conditional on numerical convergence, Full Minimum Distance is slightly more efficient than our algorithm in finite sample. This result was to be expected, as our algorithm uses only a subset of the moment conditions implied by the factor model. However, the difference in variances is not large, especially when looking at factor loadings. Second, this efficiency gain is obtained at two costs. The first one is numerical instability, which is illustrated by the final row of Table 3.4. When error variances are larger ($\text{Var}(U_\ell) = 4$), maximisation failed to converge in 157 cases out of 1000. The second cost is computing time, which increases rapidly with the number of factors.

Next, we investigate the sensitivity of our algorithm to the amount of factor kurtosis. The sample size is $N = 1000$. Errors are standard normal variables. To vary the kurtosis, we generate factors as mixtures of two independent normals. Let $W_1 \sim N(0, 1/2)$, and let $\rho \in]0, 1[$. Define $W_2 \sim N(0, (2-\rho)/(2-2\rho))$, indepen-

[16] We declared numerical convergence achieved when the gradient of the GMM criterion was inferior to 10^{-3} in absolute value after 5000 GMM iterations.

Var(U_ℓ)	.01	.25	1	4
λ_{11}	2.03 (.12)	2.04 (.14)	2.04 (.17)	2.02 (.43)
λ_{21}	.98 (.10)	.98 (.10)	.98 (.12)	.97 (.28)
λ_{31}	.98 (.10)	.98 (.11)	.99 (.13)	.98 (.28)
λ_{12}	.99 (.10)	.99 (.11)	.99 (.13)	.96 (.26)
λ_{22}	2.04 (.13)	2.04 (.12)	2.03 (.17)	2.04 (.44)
λ_{32}	.99 (.10)	.99 (.11)	.99 (.13)	.96 (.27)
λ_{13}	.99 (.11)	.98 (.11)	.98 (.13)	.96 (.28)
λ_{23}	.98 (.10)	.99 (.10)	.99 (.13)	.95 (.27)
λ_{33}	2.04 (.13)	2.04 (.13)	2.03 (.18)	2.00 (.42)
Var(U_1)	-.09 (.32)	.11 (.37)	.86 (.44)	3.75 (1.28)
Var(U_2)	-.11 (.33)	.11 (.34)	.87 (.42)	3.63 (2.27)
Var(U_3)	-.12 (.34)	.12 (.35)	.88 (.45)	3.78 (1.09)
% convergence	99.9%	100.0%	99.8%	84.3%

Table 3.4: Minimum Distance estimator based on 2nd and 4th order moments ($K = 3$, log-normal factors, normal errors, $V(U) = .25$)

dent of W_1. Then, it is straightforward to see that X, defined as W_1 with probability ρ and W_2 with probability $1 - \rho$, has variance one and its kurtosis excess is $\kappa_4(\rho) = 3\rho/(4(1-\rho))$. Table 3.5 displays Monte Carlo simulation results for values of ρ yielding kurtosis equal to $\frac{1}{2}$, 2, 5, 10 and 100. In the first column of Table 3.5, we report the results corresponding to factors following a (standardised) uniform distribution over $[-1, 1]$. The uniform distribution is platykurtic, with $\kappa_4 = -6/5$. The last column shows the results for (standardised) log-normal factors, the kurtosis excess of which is equal to $e^4 + 2e^3 + 3e^2 - 6 \approx 110$. Overall, we find that the impact of kurtosis on the performance of the algorithm is far from negligible. The closer the kurtosis excess is to zero, the greater the estimator's bias and the lower its precision.

We now set $K < L$ and compare the quasi-JADE procedures based on the restrictions of Lemma 7 and 9. The estimator based on the restrictions of Lemma 7 uses all second, third and fourth-order moment restrictions while the estimator based on the restrictions of Lemma 9 only uses second and third-order moments and assumes

ρ	(Uniform)	2/5	4/7	20/23	40/43	400/403	(Lognormal)
κ_4	-6/5	1/2	1	5	10	100	≈110
λ_{11}	1.94 (.48)	1.66 (.78)	1.76 (.74)	2.03 (.33)	2.01 (.26)	2.01 (.19)	2.03 (20)
λ_{21}	.91 (.48)	.97 (.71)	.94 (.63)	.97 (.30)	.98 (.21)	.99 (.16)	.98 (.15)
λ_{31}	.92 (.48)	1.00 (.69)	.96 (.65)	.97 (.29)	.97 (.21)	.98 (.17)	.98 (.16)
λ_{12}	.97 (.49)	1.00 (.71)	.98 (.65)	.96 (.30)	.98 (.21)	.99 (.19)	.98 (.16)
λ_{22}	1.98 (.44)	1.71 (.69)	1.83 (.64)	2.02 (.35)	2.02 (.26)	2.01 (.18)	2.03 (.18)
λ_{32}	.98 (.49)	1.00 (.72)	.95 (.66)	.97 (.30)	.98 (.20)	.99 (.18)	.98 (.16)
λ_{13}	.96 (.49)	1.12 (.74)	1.05 (.70)	.97 (.29)	.99 (.20)	.99 (.17)	.98 (.15)
λ_{23}	.94 (.49)	1.12 (.75)	1.05 (.69)	.97 (.29)	.98 (.19)	.99 (.18)	.98 (.15)
λ_{33}	1.97 (.43)	1.83 (.57)	1.89 (.56)	2.03 (.32)	2.03 (.25)	2.02 (.18)	2.03 (.20)
$\text{Var}(U_1)$	.71 (.65)	.92 (.84)	.76 (.79)	.77 (.63)	.88 (.53)	.92 (.40)	.86 (.44)
$\text{Var}(U_2)$	.75 (.65)	.89 (.83)	.69 (.78)	.75 (.64)	.83 (.55)	.93 (.40)	.87 (.43)
$\text{Var}(U_3)$	.74 (.66)	.93 (.82)	.76 (.80)	.77 (.64)	.84 (.53)	.91 (.40)	.86 (.44)

Table 3.5: Quasi-Jade with factors of increasing kurtosis (factors are normal mixtures, standard normal errors, $N = 1000$, $\Lambda = \Lambda_1$)

that all factors are skewed. Table 3.6 reports simulations with log-normal factors, standard normal errors with variance 1, and matrix Λ is equal to:

$$\Lambda_2 \equiv \begin{pmatrix} 2 & 1 \\ 1 & 2 \\ 1 & 1 \end{pmatrix}. \tag{3.37}$$

Table 3.6 shows, quite surprisingly, that fourth-order moments yield rather small additional efficiency gains. This illustrative table suggests that an algorithm based on third-order moments only could do well in practice, provided that there is enough skewness in the data. On the other hand, adding moment restrictions (and there can be a lot of fourth-order moment restrictions) does not seem to increase the bias, which is reassuring.

Lastly, we investigate the finite-sample performance of our algorithm when the number of measurements and factors increases. Table 3.7 illustrates the cases $L = K = 5$ and $L = K = 10$, respectively. In both cases, Λ has entries equal to 2 everywhere on the diagonal, and equal to one everywhere else. We only report the

N	500	500	1000	1000	5000	5000
Cumulants	2,3,4	2,3	2,3,4	2,3	2,3,4	2,3
λ_{11}	1.95 (.28)	1.93 (.32)	1.98 (.19)	1.97 (.24)	2.00 (.08)	2.00 (.08)
λ_{21}	1.96 (.30)	1.91 (.37)	1.99 (.16)	1.96 (.23)	1.00 (.09)	2.00 (.05)
λ_{31}	.97 (.23)	.98 (.25)	.98 (.17)	.98 (.20)	1.00 (.08)	1.00 (.08)
λ_{12}	2.02 (.24)	2.03 (.27)	2.01 (.17)	2.01 (.20)	1.00 (.08)	2.00 (.08)
λ_{22}	1.02 (.28)	1.05 (.32)	1.00 (.18)	1.02 (.22)	2.00 (.09)	1.00 (.08)
λ_{32}	2.01 (.12)	1.99 (.14)	2.01 (.10)	2.00 (.11)	1.00 (.05)	2.00 (.05)
Var(U_1)	.98 (.21)	1.01 (.16)	.98 (.15)	1.00 (.13)	.97 (.09)	1.00 (.06)
Var(U_2)	.94 (.21)	.99 (.20)	.96 (.15)	1.00 (.15)	.97 (.08)	1.00 (.07)
Var(U_3)	.94 (.22)	1.00 (.20)	.96 (.15)	1.00 (.15)	.98 (.09)	1.00 (.07)

Table 3.6: Comparing the two quasi-JADE algorithms based on Lemma 7 and 9 (log-normal factors, standard normal errors, $\Lambda = \Lambda_2$)

estimates of the first column of Λ and the variance of the first error, the other estimates being qualitatively similar. These simulations show that the performances of our algorithm are only moderately affected by the number of factors/measurements. We view this as quite remarkable a result as a hundred of factor loadings is certainly a significant number of parameters to estimate given that no explanatory variable is observed. In comparison, the Minimum Distance algorithm discussed above turned out to be infeasible in practice for L as low as five, the computing time becoming prohibitive.

3.4.2 Estimation of the number of factors

We here report a Monte-Carlo study of the rank tests detailed in 3.3.1.

We first compute the empirical size of the test based on matrix Ω_Y for various values of factor kurtosis. The simulation scheme is the same as for the results reported in Table 3.5. The true value of Λ is Λ_2 (as in (3.37)) and we test $K = 2$ against $K = 3$.

Table 3.8 shows substantial size distortion. This especially happens when the kurtosis excess is low (in absolute value) – that is, when fourth-order cumulants

N	$L=K=5$ 500	1000	5000	$L=K=10$ 500	1000	5000
λ_{11}	2.06 (.41)	2.03 (.28)	2.01 (.13)	1.85 (.72)	1.97 (.56)	2.00 (.27)
λ_{21}	.95 (.35)	.98 (.25)	.99 (.12)	.89 (.52)	.90 (.43)	.98 (.22)
λ_{31}	.95 (.34)	.98 (.24)	1.00 (.12)	.88 (.53)	.90 (.45)	.98 (.23)
λ_{41}	.95 (.35)	.98 (.24)	.99 (.11)	.88 (.53)	.92 (.43)	.98 (.22)
λ_{51}	.95 (.34)	.98 (.24)	.99 (.12)	.88 (.53)	.90 (.43)	.98 (.22)
λ_{61}				.88 (.54)	.91 (.43)	.98 (.22)
λ_{71}				.89 (.53)	.90 (.44)	.98 (.22)
λ_{81}				.88 (.52)	.90 (.44)	.98 (.23)
λ_{91}				.87 (.53)	.91 (.44)	.98 (.23)
$\lambda_{10,1}$				.88 (.52)	.89 (.44)	.98 (.22)
$\text{Var}(U_1)$	.58 (.56)	.81 (.44)	.95 (.20)	.40 (.55)	.49 (.53)	.88 (.28)

Table 3.7: Increasing the number of factors and measurements (log-normal factors, standard normal errors)

ρ $\kappa_4(\rho)$	- -6/5	2/5 1/2	4/7 1	20/23 5	40/43 10	400/403 100
$\alpha=.10$	.90	.73	.82	.87	.85	.62
$\alpha=.20$	.79	.57	.67	.74	.69	.43
$\alpha=.30$	.67	.44	.54	.61	.57	.29
$\alpha=.40$	.58	.33	.42	.50	.45	.19
$\alpha=.50$	.47	.24	.32	.40	.35	.11
$\alpha=.60$	.37	.16	.22	.32	.26	.05
$\alpha=.70$	.27	.10	.13	.24	.19	.02
$\alpha=.80$	.20	.05	.08	.15	.11	.01
$\alpha=.90$	.10	.02	.04	.06	.04	.00

Table 3.8: Size of the rank test based on Ω_Y for increasing kurtosis (factors of normal mixtures, errors are Gaussian, $N=1000$, $\Lambda=\Lambda_2$)

contain very little information on the factor structure – or large – that is, when fourth-order moments are imprecisely estimated. However, for reasonable values of kurtosis excess,[17] the risk of misestimating the number of factors exists but remains limited.

Fourth-order moments of unbounded distributions are imprecisely estimated because there is a non negligible probability of drawing values that are much higher than the mode of the distribution, around which most of the distribution is concentrated (peakedness). For the lognormal distribution, for example, drawing one very large value displaces the fourth-order moment to the right of its theoretical value. However, the lognormal distribution is positively skewed and there is thus a bigger probability of drawing small values, so that most of the time the fourth-order moment is underestimated. Since an excessively long tail yields imprecise estimates of higher-order moments, one may be willing to trade a bit of bias against increased precision. We thus experimented with trimming and effectively found that a certain amount of trimming (of measurement variables) improved the size of the test, but too much trimming deteriorated it. As it is impossible to say what is the "optimal" amount of trimming without knowing the model, data trimming is hardly advisable in practice.

In Section 3.3.1, we proposed to improve the size properties of the rank test by considering a weighted average of cumulant matrices $\Omega_Y(\ell, m)$ – i.e. $\Omega_{Y,w}$ in equation (3.32) – instead of Ω_Y. Table 3.9 provides a comparison of rank tests based on different cumulant matrices. We focus on the most difficult case of log-normal factors, normal

[17]Stock returns are well-known for presenting high kurtosis. The S&P 500 daily returns for 1986 to 1996 have an extremely high kurtosis of about 111. This can be ascribed to the October 1987 stock market crash (Duffie and Pan, 1997). However, between January 1969-December 2004, Lin and Hung (2005), report, for daily 1-, 30-, 100- and 300-day return data on the S&P 500 index, kurtosis values of 36.02, 5.80, 3.77 and 2.99.

Matrix	Ω_Y	$\sum_{\ell<m} w_{\ell m}\Omega_Y(\ell,m)$	Γ_Y
α =.10	.56	.87	.90
α =.20	.34	.71	.79
α =.30	.20	.56	.69
α =.40	.12	.44	.58
α =.50	.08	.32	.48
α =.60	.05	.21	.38
α =.70	.02	.13	.29
α =.80	.01	.06	.16
α =.90	.00	.01	.07

Table 3.9: Size of the rank test applied to various matrices: Ω_Y, $\sum_{\ell<m} w_{\ell m}\Omega_Y(\ell,m)$ and Γ_Y (log-normal factors, standard normal errors, $N = 1000$, $\Lambda = \Lambda_2$)

errors and a sample size of 1000. The first column reports the size of the rank test based on Ω_Y, the second column corresponds to matrix $\Omega_{Y,w}$, and the third and last column refers to matrix Γ_Y (third-order cumulants). The weighting scheme definitely improves the size of the test of $K = 2$ against $K = 3$. However, the rank test still underrejects noticeably, in particular when the theoretical probability of rejection is low. Finally, third-order moments are more precisely estimated and, consequently, the empirical size of the rank test based on Γ_Y is close to the nominal size (third column).

This confirms that applying the characteristic root test to matrices of higher-order cumulants should be done with care. However, the results in Tables 3.8 and 3.9 show that, for reasonable magnitudes of skewness and kurtosis excess, the size properties of the CR test based on third and fourth-order cumulant matrices are satisfactory.

We end this section by a study of the power of the rank test based on $\Omega_{Y,w}$. Table 3.10 display empirical power computations for various levels of kurtosis. The true value of Λ is Λ_1 and we test $K = 2$ against $K = 3$. For low significance values (α less than 10%) the power of the test is good even if factors are excessively leptokurtic. For intermediate values of the kurtosis excess, the power is good whatever the α level.

ρ	-	2/5	4/7	20/23	40/43	400/403
$\kappa_4(\rho)$	-6/5	1/2	1	5	10	100
α =.10	.99	.81	.81	1.00	1.00	.89
α =.20	.99	.63	.66	1.00	1.00	.80
α =.30	.98	.68	.51	.99	1.00	.72
α =.40	.97	.36	.39	.99	1.00	.64
α =.50	.96	.26	.29	.98	.99	.56
α =.60	.94	.18	.22	.96	.98	.47
α =.70	.93	.11	.16	.92	.96	.35
α =.80	.89	.06	.10	.86	.90	.22
α =.90	.83	.02	.04	.72	.77	.12

Table 3.10: Power of the improved rank test, Ω_Y, Factors with increasing Kurtosis (standard normal errors, $N = 1000$, $\Lambda = \Lambda_2$)

3.5 Application to the returns to schooling

In this section, we apply our methodology to the estimation of the returns to schooling. We consider the relationship between wage and education. Chamberlain and Grilliches (1975, 1977) provide insightful examples of the use of factor models in this context. We first construct a second measure of educational attainment, and we estimate a one-factor model to correct for measurement error in the first education measure. We then apply the methods of this paper and estimate a second factor.

3.5.1 The data

We use data from the French Labor Force Survey of 1995. This is a large and representative cross-section of the French labor force which provides detailed information on individual education. We exclude women, out-of-employment individuals, and workers with missing data for either (monthly) wages, hours worked or education. We trim the sample of the first and last percentiles of the wage, hour and education data. We finally obtain a sample of 21,794 workers.

We divide monthly wages by hours worked to obtain wage rates. We define Y as

	Wage Y	Years of Schooling D	Diploma D^*
Mean	0	17.7	17.6
Standard error	.29	2.64	2.17
Skewness	.29	.61	.61
Kurtosis	.079	-.015	.18
		Covariances	
Y	0.086	0.304	0.284
D	0.304	6.95	4.33
D^*	0.284	4.33	4.71

Table 3.11: Moments of the variables

the residual of the regression of wage rates on a set of regressors, including a quartic in age. We construct two education variables. The first one is the "age at the end of school", which broadly corresponds to the number of years of schooling (minus 6) in France. This variable, denoted as D, is the usual regression variable in most studies of the returns to schooling. The second one (say "diploma") codes the highest diploma obtained by the individual into 16 categories (no diploma, elementary level, middle school, high school, college, plus various declinations of these different levels into vocational and non vocational). To make this variable continuous and comparable to D, we construct a new variable, D^*, equal to the median value of D by diploma.

Table 3.11 shows the moments of the three variables of interest. The correlation between D and D^* is only 0.76, indicating that both measures of education are correlated, yet not perfectly. The OLS coefficients of the separate regressions of Y on D and on D^* are 0.044 and 0.060, respectively. The second measure yields a slightly higher return.

The two education variables are only slightly negatively skewed and exhibit little kurtosis excess. Yet, the joint distribution of (Y, D, D^*) displays a statistically significant amount of skewness and kurtosis. To check that, we estimate the three

	Γ_Y	Ω_Y	$\Omega_{Y,w}$
Rank	0	0	0
Statistic	29994	3646	20.2
Critical value .05	57.40	386.1	2.20
p-value	.00	.00	.00
Rank	1	1	1
Statistic	114.0	491.0	2.34
Critical value .05	7.74	45.4	.12
p-value	.00	.00	.00
Rank	2	2	2
Statistic	1.10	36.0	.185
Critical value .05	1.32	7.62	.0091
p-value	.072	.00	.00

Table 3.12: Rank tests

characteristic roots of matrices Γ_Y and Ω_Y, as well as their bootstrap standard errors.[18] These estimates are: 1.17 (1.14,1.20), .07 (.06,.08) and .007 (.001,.014) for the three CRs of Γ_Y, and .38 (.28,.49), .15 (.12,.18) and .04 (.03,.05) for those of Ω_Y. These results are confirmed by the CR test applied to matrices Γ_Y and Ω_Y and reported in Table 3.12. The null hypothesis that Γ_Y has rank 2 is not rejected by the data at the 5% level. The test rejects the hypothesis that the rank of Ω_Y is less than 3 at the 1% level. There is thus evidence that the joint distribution of (Y, D, D^*) is not normal.

3.5.2 Estimation results

We start by estimating the matrix of factor loadings under the assumption that $K = 1$. Factor loadings can then be estimated from covariance calculations only. We report the PCA estimates in the first column of Table 3.13 (PCA). The implied return to

[18] As in the rest of this section, 5%-95% confidence intervals are computed by 500 bootstrap replications with appropriate recentering. Confidence intervals are given between brackets.

	$K=1$ PCA	$K=1$ quasi-JADE(4)	$K=1$ quasi-JADE(3,4)	$K=2$ quasi-JADE(4)	$K=2$ quasi-JADE(3,4)
$\widehat{\lambda}_{11}$	.141 (.138,.145)	.154 (.136,166)	.142 (.137,.148)	.172 (.146,.200)	.166 (.145,.182)
$\widehat{\lambda}_{21}$	2.15 (2.12,2.19)	2.09 (2.02,2.18)	2.13 (2.09,2.20)	2.05 (1.96,2.16)	2.09 (2.02,2.19)
$\widehat{\lambda}_{31}$	2.01 (1.98,2.03)	2.05 (1.95,2.14)	2.03 (1.96,2.11)	2.02 (1.93,2.12)	2.02 (1.93,2.10)
$\frac{\widehat{\lambda}_{11}}{\widehat{\lambda}_{21}}$	6.6%	7.4%	6.7%	8.5%	7.9%
$\widehat{\lambda}_{12}$	-	-	-	-.138 (-.212,-.067)	-.136 (-.209,-.040)
$\widehat{\lambda}_{22}$	-	-	-	.360 (.009,.561)	.316 (.091,.459)
$\widehat{\lambda}_{32}$	-	-	-	.475 (.310,.660)	.381 (.131,.484)
$\widehat{V}(U_1)$	.066 (.065,.067)	.052 (.041,.070)	.066 (.060,.069)	.038 (.000,.060)	.040 (.010,.063)
$\widehat{V}(U_2)$	2.31 (2.22,2.40)	2.56 (2.06,2.90)	2.43 (2.04,2.65)	2.61 (1.85,3.04)	2.50 (1.92,2.84)
$\widehat{V}(U_3)$	.672 (.604,.745)	.426 (.000,.850)	.586 (.177,.867)	.385 (.000,.766)	.500 (.089,.889)

Table 3.13: Factor loadings and error variances (quasi-JADE(4): uses second and fourth-order moments; quasi-JADE(3,4): uses second, third and fourth-order moments)

education, as measured by $\frac{\lambda_{11}}{\lambda_{21}}$ is .066, higher than the return estimated by OLS but comparable to the OLS estimate of the regression of Y on D^*. We find that X_1 accounts for 23% of the variance of wages, 67% of the variance of D but 86% of the variance of D^*. These results are consistent with D^* being a "better" measure of educational attainment than D.[19]

We then estimate the one-factor model using high-order moments of the data. Columns 2 and 3 of Table 3.13 present the estimates of the vector of factor loadings using the quasi-JADE algorithm. In column 2, we report the results for the version of the algorithm using second and fourth-order cumulants and the restrictions of Lemma 7. In column 3, second, third and fourth-order cumulants are used and the restrictions of Lemmas 7 and 9 are combined. The results of all three columns are remarkably similar.

Next, we turn to the estimation of the two-factor model, reported in the last two

[19]Note that PCA yields the same estimate of $\frac{\lambda_{11}}{\lambda_{21}}$ as instrumenting D by D^* in the 2SLS regression of Y on D.

	$K=1$ quasi-JADE(4)	$K=1$ quasi-JADE(3,4)	$K=2$ quasi-JADE(4)	$K=2$ quasi-JADE(3,4)
$\kappa_3(X_1)$	-	1.34 (1.29,1.39)	-	1.17 (1.08,1.30)
$\kappa_3(X_2)$	-	-	-	.087 (-.709,6.10)
$\kappa_4(X_1)$	.612 (.391,.854)	.741 (.354,1.02)	.627 (.439,.768)	.665 (.445,.841)
$\kappa_4(X_2)$	-	-	13.6 (3.58,196)	15.5 (4.28,580)

Table 3.14: Factor cumulants (quasi-JADE(4): uses second and fourth-order moments; quasi-JADE(3,4): uses second, third and fourth-order moments)

columns of Table 3.13. The estimates of factor loadings associated to the first factor are very close to the values estimated using the one-factor model. The second factor is positively correlated with the number of years of schooling D and is negatively correlated with the wage Y.

We then performed a test of overidentifying restrictions, based on the JADE criterion (sum of squares of the off-diagonal elements of the jointly diagonalized matrices). We bootstrapped the test statistic 500 times to compute p-values. We found p-values of 26% and 27% for the two versions of quasi-JADE, when $K = 2$ was assumed. Thus, according to this criterion, at all conventional levels, the data do not reject the validity of the overidentifying restrictions imposed in quasi-JADE.

Notice that, using third-order moments only, we obtained very imprecise estimates (not reported). This is because the second factor is found to have a nearly symmetric distribution. We report in Table 3.14 the estimates of factor cumulants. The results show that the first factor is skewed to the left, with rather small kurtosis. Moreover, the second factor shows little skewness but displays much kurtosis excess. This implies that the second factor is essentially identified from fourth-order moments of the data.

Finally, we tried to investigate the existence of a third factor without success. The estimates were far too imprecise. In any case, if a third factor exists, it has very little explanatory power on individual earnings.

3.5.3 Interpretation

The factor structure is consistent with the interpretation that overspecialization in education is counterproductive as far as market value is concerned.

Another way of interpreting these results is as follows. We obtain the following factor structure:

$$\begin{cases} Y = .17X_1 - .14X_2 + U_1 \\ D = 2X_1 + .4X_2 + U_2 \\ D^* = 2X_1 + .4X_2 + U_3 \end{cases} \tag{3.38}$$

Let $E = 2X_1 + .4X_2$. View E as the "true" education measure and U_2 and U_3 as measurement errors. The number of years of education, D, faces large measurement error ($\text{Var}(U_2) = 2.6$ and $\text{Var}(E) = 5.6$) in comparison to the other education measure based on the highest diploma obtained ($\text{Var}(U_3) = .4$).[20] This explains why the OLS estimate of the regression of Y on D^* is less biased toward zero than the OLS estimator of the regression of Y on D (6% vs 4.4%).

Now, rewrite the wage equation as

$$\begin{aligned} Y &= .17\left(\frac{E - .4X_2}{2}\right) - .14X_2 + U_1 \\ &= 0.085E - 0.174X_2 + U_1. \end{aligned} \tag{3.39}$$

Factor X_2 can be interpreted as the unobserved factor correlating E and the error term in the wage equation ($-0.174X_2 + U_1$). The negative correlation may result from marginal costs of education increasing faster than returns to education across individuals. Now this correlation is small (about 7%). This explains why the returns to education is about 8.5% when we introduce a second factor, and is about 7% when we only control for measurement error (the one-factor model) whereas the OLS

[20] These results are in line with previous evidence for graduate students. Ashenfelter and Mooney (1968) stress that the number of years of education completed is a very incomplete indicator of the education level for this population. Recently, Hamermesh and Donald (2004) find highly significant differences in returns accross majors, even after controling for non-response bias.

estimate is 4.4%. Most of the OLS bias is due to measurement error. Still, a fraction of the bias (about 20%) is due to unobserved heterogeneity. Interestingly, in this case unobserved heterogeneity and measurement error biases point in the same direction.

Of course, equation (3.39) results from changing (X_1, X_2) into (E, X_2), which is only one possible rotation, among many others, of (X_1, X_2). There are thus other possible structural interpretations. For example, one can also change (X_1, X_2) into (E, V) where V is chosen orthogonal (to the second order) to E, for example: $V = X_1 - 5X_2$. In which case

$$Y = 0.068E + 0.033V + U_1.$$

This equation yields a returns of 6.8%, which exactly corresponds to the PCA estimate (regress Y on D, instrumenting D by D^*).

3.6 Conclusion

It is well known that non normality is an important source of identification in linear measurement error models. In this chapter, we extend this insight to general linear independent factor models. We prove that $L(L-1)/2$ factors can be generically identified from a set of L measurement. Contrary to ordinary Factor Analysis, identification is unambiguously defined up to sign and permutation normalisations.

We also prove that second, third and/or fourth-order moments of the data provide sufficient information to identify and estimate the first four moments of at most L factors. We then extend and adapt a well-known technique of Independent Component Analysis (ICA), Cardoso and Souloumiac's (1993) JADE algorithm, to construct estimators of factor loadings in the case where errors are not negligible. We propose a multi-step procedure (quasi-JADE) in which we estimate error moments in a first stage, and then apply Cardoso and Souloumiac's approximate joint diagonalisation algorithm.

The independent factor structure generates many overidentifying restrictions on higher-order moments. This may explain the encouraging Monte Carlo simulation results that we obtained. In contrast with previous evidence on the use of higher-order moments for estimation,[21] we find, for sufficiently non symmetric and/or kurtotic data, small biases and precise estimates, even in relatively small samples. The estimation methodology is then applied to earnings and education data. Besides the common factor that IV and PCA estimates already reveal (explaining the bias toward zero of the OLS estimate of the returns to the number of years of education on individual earnings), quasi-ICA reveals an interesting second factor that is negatively correlated with earnings and positively correlated with education. This is evidence that there exist individual characteristics which are valued by the education institution but not by the labour market.

In the future, we plan to pursue two directions of research. First, this chapter leaves many methodological questions unanswered. In particular, efficiency issues concerning the quasi-JADE estimators, as well as the properties of the tests of the number of factors, seem worth investigating further. Moreover, it would be interesting to extend existing algorithms to deal with more factors than measurements ($K > L$). In the ICA literature, this case is referred to as *overcomplete* ICA. De Lathauwer (2003) presents an algorithm comparable to JADE that works for $K > L$ in the case of complex measurements. In the real case, the one of interest in econometrics, we are not aware of similar semi-parametric methods.

The second direction of research concerns the extension of the method of this

[21] See the results reported in Madansky (1959), and the survey by Aigner *et al.* (1984).

paper to the case of a very large number of measurements. Bai and Ng (2002) and Bai (2003) provide extensive analyses of the PCA estimator in this case. Financial and macroeconomic applications motivate the need to extend ICA methods in this direction.

Chapter 4

Nonparametric Estimation of Factor Distributions in Linear Independent Factor Models

4.1 Introduction

In this chapter, we consider linear factor models of the form:

$$Y = \Lambda X + U, \tag{4.1}$$

where Y is a vector of L observed outcomes, X is a vector of K unobservable common factors, U is a vector of L independent errors and Λ is a $L \times K$ matrix of parameters (factor loadings). The critical assumption is that all components of X and U are mutually independent.

In econometrics, linear factor models are widely used as a dimensionality-reduction device. In particular, unobserved heterogeneity is often assumed to follow a linear factor structure, as in the case of the one-factor error component model:

$$y_{it} = x_{it}^T b + u_i + v_{it}, \quad i = 1, ..., N, \quad t = 1, ..., L, \tag{4.2}$$

where v_{it} is an i.i.d. error independent of the individual factor u_i.

This chapter focuses on the nonparametric identification estimation of the distributions of factors and errors in (4.1). Throughout the chapter, we assume that a root-N consistent estimator of the matrix of factor loadings Λ is available, and focus on the identification and estimation of factor and error distributions. Under suitable assumptions, estimators of Λ can be derived from covariance restrictions only, as in orthogonal factor analysis. In addition, higher-order information can be used to relax some of the limitations of the traditional approach (as in chapter 3).

One possibility to estimate factor densities is to use flexible distributional families, as the family of normal mixtures. This is the approach followed by Carneiro, Hansen and Heckman (2003) in the context of a structural Roy model of education choice. There are two main difficulties with this method: First, it is difficult to estimate the degree of mixing (the number of components in a mixture model). Second, estimation along these lines involves multiple integrations, and computer-intensive techniques such as MCMC.

The alternative approach we adopt in this chapter uses deconvolution methods (*e.g.* Stefanski and Caroll, 1990). The idea of deconvolution is to proceed in two steps: estimate the characteritic function of the variable of interest, then recover its density by inverse Fourier transform. Linear independent factor models imply particularly simple relationships between the characteristic functions of factors and measurements. Thus, it is natural to use deconvolution in this context.

The original deconvolution approach deals with the simple case with one factor and one error, the distribution function of which is known. Subsequently, several generalizations have been proposed in the literature. Horowitz and Markatou (1996), Li and Vuong (1998) and Hall and Yao (2003) have proposed estimators to deal with one-factor models of the form (4.2). Related methods have been applied by Li (2002), Schennach (2004) and Hu and Ridder (2005) to the problem of estimating nonlinear models with measurement error, and by Linton and Whang (2002) to estimation with aggregated data. To our knowledge, however, there is no readily available result allowing to deal with the more general case of a multidimensional structure of individual heterogeneity of the form:

$$y_{it} = x_{it}^T b + a_{t1} u_{1i} + a_{t2} u_{2i} + ... + v_{it}.$$

This framework includes as particular case moving average error structures like:

$$y_{it} = x_{it}^T b + u_{it} + a u_{i,t-1} + v_{it},$$

which have $L-1$ common factors $u_{i1}, ..., u_{i,L-1}$ and L errors $au_{i0} + v_{i1}, v_{i2}, ..., v_{i,L-1}$ and $u_{iL} + v_{iL}$. It is to be noted that there is even no general identification result in this case.

Our contribution is twofold: Firstly, we show that factor and error distributions are nonparametrically identified, under general conditions. Secondly, we propose a class of estimators based on the identification proof. Our estimators rely on the second derivative of the cumulant generating function (c.g.f.) of the data, and can be used to recover factor and error densities by integration and deconvolution. It can be seen as a generalization of Li and Vuong's (1998) estimator to the multi-factor case.

We then address two main issues. The general asymptotic theory of deconvolution estimators being an extremely difficult topic (*e.g.* Fan, 1991), we do not derive the asymptotic distribution of our estimator. Still, we prove that our estimator converges uniformly to the true density when the sample size tends to infinity. Moreover, we provide some upper bounds for the convergence rates in special cases.

Our proof of convergence builds on results from statistical learning theory. We show how to prove the convergence of second derivatives of empirical characteristic functions to their means, extending the Glivenko-Cantelli Theorem. Our approach is conceptually simpler than Li and Vuong's (1998). Moreover, it is not clear how to extend their use of Von Mises differential calculus to our case.

Then, we study the practical behavior of the estimator. As the assumption of independence yields many overidentifying restrictions, one has to choose among a large class of estimators. We provide intuitions to pick up one special estimator in this class. We also discuss the choice of the amount of smoothing in practice.

Our findings are consistent with the ones of the deconvolution literature. Convergence rates are slower than root-N, and can be very slow in special cases. We also illustrate our approach by means of Monte Carlo simulations. We find that the shape of factor distributions strongly influences the finite-sample performance of the estimator. As a special case, worse performance is achieved when the distributions are skewed or kurtotic.

We lastly apply our methodology to the two-factor model of returns to schooling discussed in chpater 3. Estimating the distributions allows us to assign two factor means to each individual in the sample, thus allowing to correlate the unobserved factors to observed covariates.

The outline of the chapter is as follows. First, we provide a short review of the literature. In Section 2, we discuss identification. In Section 3, we propose a class of nonparametric estimators and we study their asymptotic properties in Section 4. Sections 5 and 6 present some simulations and the application. Lastly, Section 7 concludes.

4.2 Review of the literature

The deconvolution problem. One aims at estimating the distribution of X given the distribution of U and a sample of i.i.d. observations of a random variable Y such that $Y = X + U$. It is assumed that X and U are independent and absolutely continuous. Let φ_Y, φ_Y and φ_U denote the characteristic functions (c.f.'s) of Y, X and U. Assume that φ_U is nonvanishing everywhere. Then,

$$\varphi_X(t) = \frac{\varphi_Y(t)}{\varphi_U(t)},$$

and the probability density function (pdf) of X, say f_X, follows as the inverse Fourier transform of φ_X:

$$\begin{aligned} f_X(x) &= \frac{1}{2\pi}\int \exp(-itx)\varphi_X(t)dt, \\ &= \frac{1}{2\pi}\int \exp(-itx)\frac{\varphi_Y(t)}{\varphi_U(t)}dt. \end{aligned} \tag{4.3}$$

As noted by several authors (e.g. Horowitz, 1998), this argument proves identification, but cannot be directly used for estimation since the above integral does not necessarily converge when the characteristic functions are replaced by their empirical analogs. Consistent estimation of f_X therefore requires some smoothing, resulting in low convergence rates.

Generalizing the results of Carroll and Hall (1988), Fan (1991) shows that the convergence rate of the deconvolution estimator crucially depends on the *smoothness* of the distributions of X and U; that is: on the tails of their characteristic functions. In particular, slow rates can be achieved in the case of normally distributed error.

Repeated measurements. Now, let there be two measurements of X:

$$\begin{cases} Y_1 = X + U_1, \\ Y_2 = X + U_2, \end{cases}$$

with X, U_1 and U_2 independent. Let us also assume that measuremnts are centered, and X, U_1 and U_2 have zero mean. Kotlarski (1967) shows that the distributions of all three variables X, U_1 and U_2 are identified by the distribution of $Y = (Y_1, Y_2)$.

Various proofs of Kotlarski's result can be found in the literature (see Rao, 1992, p.21). Relying on these proofs, several authors have proposed consistent estimators of the factor and error distributions. Horowitz and Markatou (1996) focus on the case where the d.f.'s of U_1 and U_2 are identical and symmetric. Then, the information contained in the three univariate distributions of Y_1, Y_2 and $Y_2 - Y_1$ is enough to construct consistent estimators of f_X and $f_{U_1} = f_{U_2}$. Li and Vuong (1998) consider the more general case of nonsymmetric distributions for U_1 and U_2.

We now describe Li and Vuong's estimator in some details as this chapter extends their approach to the multi-factor case. The c.f. of $Y = (Y_1, Y_2)$ is

$$\varphi_Y(t_1, t_2) = \varphi_X(t_1 + t_2)\varphi_{U_1}(t_1)\varphi_{U_2}(t_2). \tag{4.4}$$

Hence,

$$\frac{\partial \ln \varphi_Y(0,t)}{\partial t_1} = (\ln \varphi_X)'(t),$$

as $\left(\ln \varphi_{U_1}\right)'(0) = i\mathbb{E}U_1 = 0$. Li and Vuong estimate $\varphi_X(t)$ as

$$\widehat{\varphi}_X(t) = \exp \int_0^t \frac{\partial \ln \widehat{\varphi}_Y(0,u)}{\partial t_1} du, \tag{4.5}$$

where $\widehat{\varphi}_Y(t) = \frac{1}{N}\Sigma_{i=1}^N e^{ity_i}$ is a consistent estimate of $\varphi_Y(t)$ from an i.i.d. sample $(y_1, .., y_N)$. The pdf of X is recovered by inverse Fourier transform, as in the deconvolution problem. Li and Vuong derive the convergence rate of their estimator in several cases, depending on the smoothness of the distributions. Li (2002) uses this estimator in the context of nonlinear errors-in-variables models.

In a recent paper, Hall and Yao (2003) build an alternative uniformly consistent estimator of f_X from the distributions of Y_1, Y_2 and $Y_1 + Y_2$. Condition (4.4) indeed implies the following restriction:[1]

$$\frac{\varphi_Y(t,t)}{\varphi_Y(t,0)\varphi_Y(0,t)} = \frac{\varphi_{Y_1+Y_2}(t)}{\varphi_{Y_1}(t)\varphi_{Y_2}(t)} = \frac{\varphi_X(2t)}{\varphi_X(t)^2}. \tag{4.6}$$

Function $h(t) = \frac{\varphi_{Y_1+Y_2}(t)}{\varphi_{Y_1}(t)\varphi_{Y_2}(t)}$ has an immediate empirical counterpart and f_X can be estimated either as a discrete approximation verifying restriction (4.6) or by inverse Fourier transform of the analytical solution to equation (4.6), that is:

$$\varphi_X(t) = h\left(\frac{t}{2}\right) \prod_{j=1}^{\infty} h\left(\frac{t}{2^{j+1}}\right)^{2j}. \tag{4.7}$$

We shall not follow this route in this chapter as it is not easy to see how to extend this approach to the multi-factor case.

[1] This restriction is not exactly identical to the one exploited by Hall and Yao. The difference is not essential however.

4.3 Identification

In this section, we study the identification of factor densities. Next section will be devoted to their estimation. We shall most of the time eliminate the distinction between factors and errors, because, as far the identification and estimation of their distribution is concerned, the distinction is not essential. So, we now consider the case of DGPs of the form: $Y = AX$, where

1. $Y = (Y_1, ..., Y_L)^T$ is a vector of $L \geq 2$ zero-mean real-valued random variables (where ^{T} denotes the matrix transpose operator),

2. $X = (X_1, ..., X_K)^T$ is a random vector of K real valued, mutually independent and non degenerate random variables with zero means and finite variances,

3. $A = [a_{ij}]$ is a known $L \times K$ matrix of scalar parameters such that any two columns are linearly independent,

4. the characteristic functions of factors are non vanishing and two times differentiable everywhere.

In the factor analysis literature, variables Y_ℓ are called "measurements", and parameters a_{ij} are called "factor loadings". If the kth column of A (say, $A_{[\cdot,k]}$) contains only one nonzero element, variable X_k is called an "error" and otherwise, it is called a "factor". In the various examples that we shall consider, we use $U_1, ..., U_L$ to denote error variables, reserving the notation X for common factors and A is of the form $A = (\Lambda, I_L)$.

Without any restriction on A, we have shown in chapter 3 that $K = \frac{L(L-1)}{2}$ is the maximal number of factor distributions that can be identified in a linear independent factor model with L measurements and L errors, if factor and error distributions are divisible by the normal. We shall show that this limit is not binding if A is restricted.

If two columns of A are proportional, say $A_{[\cdot,k]} = \alpha A_{[\cdot,j]}$, then $A_{[\cdot,k]}X_k + A_{[\cdot,j]}X_j = A_{[\cdot,j]}\left(\alpha X_k + X_j\right)$ and there is obviously no way to separately identify the distribution of X_k from the distribution of X_j.

Lastly, instead of assuming A known, we can only assume that a root-N consistent estimator of A is available.

4.3.1 Identifying restrictions

Under assumption 4, cumulant generating functions are well defined and two times differentiable everywhere. Let us denote the characteristic functions of Y and X_k as φ_Y and φ_{X_k}, and their cumulant generating functions as $\kappa_Y = \ln \varphi_Y$ and $\kappa_{X_k} = \ln \varphi_{X_k}$. The independence assumptions and the linear factor structure imply that, for all $t = (t_1, ..., t_L) \in \mathbb{R}^L$,

$$\kappa_Y(t) \equiv \ln \left[\mathbb{E} \exp\left(it^T Y\right)\right] = \sum_{k=1}^{K} \kappa_{X_k}\left(t^T A_{[\cdot,k]}\right). \qquad (4.8)$$

Next, let us denote as $\partial_\ell \kappa_Y(t)$ the ℓth partial derivative of $\kappa_Y(t)$ and as $\partial^2_{\ell m}\kappa_Y(t)$ the second-order partial derivative of $\kappa_Y(t)$ with respect to t_ℓ and t_m:

$$\partial_\ell \kappa_Y(t) = i\frac{\mathbb{E}\left[Y_\ell e^{it^T Y}\right]}{\mathbb{E}\left[e^{it^T Y}\right]}, \qquad (4.9)$$

$$\partial^2_{\ell m}\kappa_Y(t) = -\frac{\mathbb{E}\left[Y_\ell Y_m e^{it^T Y}\right]}{\mathbb{E}\left[e^{it^T Y}\right]} + \frac{\mathbb{E}\left[Y_\ell e^{it^T Y}\right]}{\mathbb{E}\left[e^{it^T Y}\right]}\frac{\mathbb{E}\left[Y_m e^{it^T Y}\right]}{\mathbb{E}\left[e^{it^T Y}\right]}. \qquad (4.10)$$

Let $\Delta_{L,2} = \left\{(\ell, m) \in \{1, ..., L\}^2, \ell \leq m\right\}$ be a set of $L(L+1)/2$ bidimensional indices. Let $\nabla\kappa_Y(t)$ denote the L-dimensional gradient vector and let $\nabla^2\kappa_Y(t)$ denote the vector of all $\frac{L(L+1)}{2}$ non redundant second-order partial derivatives arranged in lexicographic order of (ℓ, m) in $\Delta_{L,2}$. Lastly, for any $\tau = (\tau_1, ..., \tau_K) \in \mathbb{R}^K$, denote as

$$\begin{aligned}
\boldsymbol{\kappa}_X(\tau) &= \left(\kappa_{X_1}(\tau_1), ..., \kappa_{X_K}(\tau_K)\right)^T, \\
\boldsymbol{\kappa}'_X(\tau) &= \left(\kappa'_{X_1}(\tau_1), ..., \kappa'_{X_K}(\tau_K)\right)^T, \\
\boldsymbol{\kappa}''_X(\tau) &= \left(\kappa''_{X_1}(\tau_1), ..., \kappa''_{X_K}(\tau_K)\right)^T,
\end{aligned}$$

the K-dimensional vectors of factor cumulant generating functions and corresponding first and second derivatives.

First differentiating equation (4.8) yields:

$$\nabla \kappa_Y(t) = A\boldsymbol{\kappa}'_X\left(A^T t\right) = \sum_{k=1}^{K} \kappa'_{X_k}\left(t^T A_{[\cdot,k]}\right) A_{[\cdot,k]}. \tag{4.11}$$

In general, $K > L$ as there are L errors and at least one common factor. So there are more function κ'_{X_k} than $\partial_\ell \kappa_Y$. To obtain an invertible system, we differentiate one more time:

$$\nabla^2 \kappa_Y(t) = Q(A)\,\boldsymbol{\kappa}''_X\left(A^T t\right), \tag{4.12}$$

where $Q(A)$ is the matrix operator that changes $A = [a_{ij}] \in \mathbb{R}^{L\times K}$ into the $\frac{L(L+1)}{2} \times K$ matrix which generic element is $a_{\ell k} a_{mk}$, when the row index is $(\ell, m) \in \Delta_{L,2}$ and the column index is $k \in \{1, ..., L\}$. In the sequel, we write Q for $Q(A)$ to simplify the notations.

If system (4.12) is invertible for all t in $\mathbb{R}^L$, then identification is guaranteed as we now explain. However, it can be that the system is not invertible for given t, and yet that (4.12) has a unique solution as a system of functional equations. We shall give an example at the end of this section.

4.3.2 Regular case: Q full column rank

Assume Q full column rank. This requires in particular that $K \leq \frac{L(L+1)}{2}$. We can invert equation (4.12) as

$$\boldsymbol{\kappa}_X'' \left(A^T t\right) = Q^- \nabla^2 \kappa_Y (t), \tag{4.13}$$

where Q^- is a generalised inverse of Q.[2]

This equation provides a set of overidentifying restrictions which can be exploited to yield an expression for κ_{X_k}'' and κ_{X_k}.

Let $\mathcal{T}_k = \left\{t \in \mathbb{R}^L | t^T A_{[\cdot,k]} = 1\right\}$. $\mathcal{T}_k$ is not empty as there is at least one non zero element in $A_{[\cdot,k]}$. Let $(Q^-)_{[k,\cdot]}$ denote the kth row of Q^-. Then, for all $t \in \mathcal{T}_k$ and $\tau_k \in \mathbb{R}$,

$$\kappa_{X_k}'' (\tau_k) = \left(Q^-\right)_{[k,\cdot]} \nabla^2 \kappa_Y (\tau_k t).$$

integrating with respect to τ_k, using the constants of integration: $\kappa_{X_k}' (0) = i\mathbb{E}X_k = 0$ and $\kappa_{X_k} (0) = 0$, yields

$$\kappa_{X_k} (\tau_k) = \int_0^{\tau_k} \int_0^u \left(Q^-\right)_{[k,\cdot]} \nabla^2 \kappa_Y (vt)\, dvdu. \tag{4.14}$$

It then follows from (4.14) that factor cumulant generating functions are identified. This in turn implies that factor characteristic functions, and hence factor densities, are identified. Moreover, equation (4.14) can be used for estimation, as we shall explain in the next section.

Example: the measurement error model. Consider the measurement error model:

$$\begin{cases} Y_1 = X + U_1 \\ Y_2 = aX + U_2, \end{cases}$$

[2]That is: $Q^- = \left(Q^T W Q\right)^{-1} Q^T W$, for a symmetric, positive definitive matrix W.

with $a \neq 0$ and $X \in \mathbb{R}$. A root-N consistent estimator of a could be obtained using third-order moments of Y by regressing Y_2 on Y_1 using $Y_1 Y_2$ as instrument, provided that the distribution of X is skewed (Reiersol, 1950). We here assume that a is known.

Then,

$$\kappa_Y(t_1, t_2) = \kappa_X(t_1 + a t_2) + \kappa_{U_1}(t_1) + \kappa_{U_2}(t_2),$$

and

$$\begin{pmatrix} \partial^2_{11}\kappa_Y(t_1,t_2) \\ \partial^2_{12}\kappa_Y(t_1,t_2) \\ \partial^2_{22}\kappa_Y(t_1,t_2) \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ a & 0 & 0 \\ a^2 & 0 & 1 \end{pmatrix} \begin{pmatrix} \kappa''_X(t_1 + a t_2) \\ \kappa''_{U_1}(t_1) \\ \kappa''_{U_2}(t_2) \end{pmatrix},$$

or

$$\begin{pmatrix} \kappa''_X(t_1 + a t_2) \\ \kappa''_{U_1}(t_1) \\ \kappa''_{U_2}(t_2) \end{pmatrix} = \begin{pmatrix} 0 & \frac{1}{a} & 0 \\ 1 & -\frac{1}{a} & 0 \\ 0 & -a & 1 \end{pmatrix} \begin{pmatrix} \partial^2_{11}\kappa_Y(t_1,t_2) \\ \partial^2_{12}\kappa_Y(t_1,t_2) \\ \partial^2_{22}\kappa_Y(t_1,t_2) \end{pmatrix}.$$

The restriction:

$$\kappa''_X(t_1 + a t_2) = \frac{1}{a}\partial^2_{12}\kappa_Y(t_1,t_2), \forall (t_1,t_2),$$

implies that, for all τ and t_1,

$$\kappa''_X(\tau) = \frac{1}{a}\partial^2_{12}\kappa_Y\left(t_1, \frac{1}{a}\tau - \frac{1}{a}t_1\right),$$

and

$$\begin{aligned} \kappa_X(\tau) &= \frac{1}{a}\int_0^\tau \int_0^u \partial^2_{12}\kappa_Y\left(t_1, \frac{1}{a}v - \frac{1}{a}t_1\right) dv du, \\ &= \int_0^\tau \left[\partial_1\kappa_Y\left(t_1, \frac{1}{a}u - \frac{1}{a}t_1\right) - \partial_1\kappa_Y\left(t_1, -\frac{1}{a}t_1\right)\right] du. \end{aligned}$$

Setting $t_1 = 0$ yields Li and Vuong's (1998) solution:

$$\kappa_X(\tau) = \int_0^{\tau_1} \partial_1\kappa_Y\left(0, \frac{1}{a}u\right) du.$$

So, Li and Vuong's estimator is particular for two reasons: First, in general, the double integrals of second-order derivatives of κ_Y in (4.14) will not simplify into a simple integral of first derivatives. Second, the choice of which component of t to fix a priori is arbitrary. It does not matter if one is only interested to show identification. Yet, overidentifying restrictions could be used to improve the estimation.

4.3.3 Irregular case: Q not full column rank

As A is known, matrix Q being full column rank is not a necessary condition for identification. For example, consider the following factor model with two factors and two measurements:

$$\begin{cases} Y_1 = X_1 + X_2 + U_1 \\ Y_2 = X_1 + aX_2 + U_2, \end{cases}$$

with $a \neq 1$ and $a \neq 0$ for the columns of

$$A = \begin{pmatrix} 1 & 1 & 1 & 0 \\ 1 & a & 0 & 1 \end{pmatrix}$$

not to be two-by-two proportional. In this example, $K = 4 > \frac{L(L+1)}{2} = 3$. So $Q \in \mathbb{R}^{3\times 4}$ cannot be full column rank.

The cumulant generating function of Y is

$$\kappa_Y(t_1, t_2) = \kappa_{X_1}(t_1 + t_2) + \kappa_{X_2}(t_1 + at_2) + \kappa_{U_1}(t_1) + \kappa_{U_2}(t_2)$$

and

$$\begin{cases} \partial^2_{11}\kappa_Y(t_1, t_2) = \kappa''_{X_1}(t_1 + t_2) + \kappa''_{X_2}(t_1 + at_2) + \kappa''_{U_1}(t_1) \\ \partial^2_{22}\kappa_Y(t_1, t_2) = \kappa''_{X_1}(t_1 + t_2) + 2a\kappa''_{X_2}(t_1 + at_2) + \kappa''_{U_2}(t_2) \\ \partial^2_{12}\kappa_Y(t_1, t_2) = \kappa''_{X_1}(t_1 + t_2) + a\kappa''_{X_2}(t_1 + at_2). \end{cases}$$

We now show that one can recover κ_{X_1} and κ_{X_2} from the third restriction if the factor variances are known. Constraining either $t_1 + t_2$ or $t_1 + at_2$ to be equal to zero yields the following expressions for κ''_{X_1} and κ''_{X_2}:

$$\begin{cases} \kappa''_{X_1}(\tau_1) = \partial^2_{12}\kappa_Y\left(-\frac{a\tau_1}{1-a}, \frac{\tau_1}{1-a}\right) - a\kappa''_{X_2}(0) \\ \kappa''_{X_2}(\tau_2) = \frac{1}{a}\partial^2_{12}\kappa_Y\left(\frac{\tau_2}{1-a}, -\frac{\tau_2}{1-a}\right) - \frac{1}{a}\kappa''_{X_1}(0), \end{cases}$$

where $\kappa''_{X_k}(0) = -\operatorname{Var}(X_k)$.

Assume that these variances are known equal to 1. Note that this assumption was not needed in the regular case, as factor variances are identified from (4.13), taking $t=0$. Using the fact that factors have zero mean, $\kappa'_{X_1}(0)=\kappa'_{X_2}(0)=0$ and one has

$$\begin{cases} \kappa_{X_1}(\tau_1)=\int_0^{\tau_1}\int_0^u \partial^2_{12}\kappa_Y(2v,-v)\,dvdu - a\frac{\tau_1^2}{2} \\ \kappa_{X_2}(\tau_2)=\int_0^{\tau_2}\int_0^u \partial^2_{12}\kappa_Y(-v,v)\,dvdu - \frac{1}{a}\frac{\tau_2^2}{2}. \end{cases}$$

4.4 Estimation

We first introduce the estimator of factor densities, and then discuss issues linked to its practical implementation.

4.4.1 The estimator

First step: Following most of the literature on deconvolution, given an i.i.d. sample of size N, we first estimate κ_Y and its derivatives by empirical analogs, replacing the mathematical expectations in (4.8), (4.9) and (4.10) by arithmetic means:

$$\widehat{\kappa}_Y(t)=\ln\left(\mathbb{E}_N\left[e^{it^TY}\right]\right), \tag{4.15}$$

$$\widehat{\partial_\ell\kappa_Y}(t)=i\frac{\mathbb{E}_N\left[Y_\ell e^{it^TY}\right]}{\mathbb{E}_N\left[e^{it^TY}\right]}=\partial_\ell\widehat{\kappa}_Y(t), \tag{4.16}$$

and

$$\widehat{\partial^2_{\ell m}\kappa_Y}(t)=-\frac{\mathbb{E}_N\left[Y_\ell Y_m e^{it^TY}\right]}{\mathbb{E}_N\left[e^{it^TY}\right]}+\frac{\mathbb{E}_N\left[Y_\ell e^{it^TY}\right]}{\mathbb{E}_N\left[e^{it^TY}\right]}\frac{\mathbb{E}_N\left[Y_m e^{it^TY}\right]}{\mathbb{E}_N\left[e^{it^TY}\right]}=\partial^2_{\ell m}\widehat{\kappa}_Y(t), \tag{4.17}$$

where $\mathbb{E}_N$ denotes the empirical expectation.

Second step: As the choice of t in $\mathcal{T}_k=\left\{t\in\mathbb{R}^L|t^TA_{[\cdot,k]}=1\right\}$, along which to perform the integration yielding $\kappa_{X_k}(\tau_k)$, is arbitrary, one can estimate κ_{X_k} by averaging solution (4.14) over a distribution of points in $\mathcal{T}_k$, that is,

$$\begin{aligned}\widehat{\kappa}_{X_k}(\tau) &= \int_0^\tau \int_0^u \left(Q^-\right)_{[k,\cdot]} \left(\int \nabla^2 \kappa_Y(vt)\, dW(t)\right) dvdu \\ &= \int_0^\tau \int_0^u \left(Q^-\right)_{[k,\cdot]} \left(\sum_{j=1}^p w_j \nabla^2 \kappa_Y(vt_j)\right) dvdu, \qquad (4.18)\end{aligned}$$

where $W = \sum_{j=1}^p w_j \delta_{t_j}$ is a symmetric, discrete probability distribution on $\mathcal{T}_k$.

Third step: We then estimate the factor distribution functions by inverse Fourier transform:

$$\begin{aligned}\widehat{f}_{X_k}(x) &= \frac{1}{2\pi} \int_{-T_N}^{T_N} \widehat{\varphi}_{X_k}(\tau) \exp(-i\tau x) d\tau \qquad (4.19) \\ &= \frac{1}{2\pi} \int_{-T_N}^{T_N} \exp(-i\tau x + \widehat{\kappa}_{X_k}(\tau)) d\tau,\end{aligned}$$

where T_N tends to infinity at a rate to be specified.

4.4.2 Practical issues

We here discuss two issues: the choice of T_N in practice, and the choice of the estimator of factor characteristic functions in the class of estimators given by (4.18).

Choice of T_N: To estimate T_N in practice, we apply the methodology in Diggle and Hall (1993) that we now present. In the context of a deconvolution problem $Y = X + U$ with X and U independent and (Y, U) observed, Diggle and Hall consider the choice of the trimming parameter T_N in

$$\widehat{f}_X(x) = \frac{1}{2\pi} \int_{-T_N}^{T_N} \exp(-itx) \frac{\widehat{\varphi}_Y(t)}{\widehat{\varphi}_U(t)} dt.$$

Approximating the Mean Integrated Squared Error $\int |\widehat{f}_X(x) - f_X(x)|^2 dx$ for large N and maximizing the expression with respect to T_N, they find that the optimal trimming parameter has to satisfy:

$$\varphi_Y(T_N) = N^{-1/2}. \tag{4.20}$$

For given t, $|\widehat{\varphi}_Y(t) - \varphi_Y(t)|$ is also of order $N^{-1/2}$. One thus cannot replace the unknown c.f. in (4.20) by its estimated counterpart. Assuming that $\varphi_Y(t) = \alpha t^{-\beta}$ for large enough $|t|$, Diggle and Hall then propose to proceed in three steps. First, estimate α and β by linear regression of $\ln|\varphi_Y(t)|$ on $\ln|t|$ on a range where this relationship is approximately linear. Then, substitute $\widehat{\alpha}T_N^{-\widehat{\beta}}$ into (4.20) to get T_N. Lastly, perform the integration by multiplying the c.f in the inverse Fourier transform by a "damping factor". This last step aims at reducing the oscillations that often characterize estimated c.f.'s in their tails.

In the case of model $Y = AX$, we propose to proceed analogously. Let $k \in \{1...K\}$, and $t = A^{-T}_{[.,k]} = \frac{A_k}{A_k^T A_k}$. Then

$$t^T Y = t^T AX = X_k + \sum_{m \neq k} t^T A_{[.,m]} X_m. \tag{4.21}$$

We suggest to proceed as if the d.f. of $\sum_{m \neq k} t^T A_{[.,m]} X_m$ in (4.21) were known. In this case, estimating the density of factor X_k reduces to a classical deconvolution problem, and the approach in Diggle and Hall (1993) can be applied.[3]

In practice, cumulant generating functions are approximately linear in $\ln|t|$ over a wide range around zero, yet badly estimated in the tails. An illustration is provided by Figure 4.1, which plots the logarithm of the characteristic function (in absolute value) of the wage data used in the empirical section. Index $\ln|t|$ is plotted on the x-axis.

[3] In particular, the "damping factor" we choose is given by:

$$d(v) = \mathbf{1}\{|v| < (1-\mu)T_N\} + \mathbf{1}\{(1-\mu)T_N \leq |v| \leq T_N\} \cdot \frac{1}{\mu}\left[1 - |v|/T_N\right],$$

where we set $\mu = .05$.

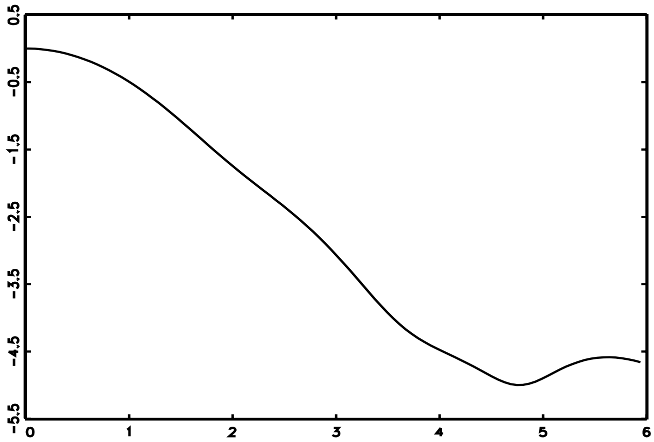


Figure 4.1: Cumulant generating function, wage data

Figure 4.1 is representative of many c.f.'s that we encountered in the simulations and in the application. The graph is approximately linear in $\ln |t|$ over $[1, 3]$, and becomes more erratic afterwards. For distributions of this sort we found the Diggle and Hall method very reliable given its simplicity.

Of course, some standard characteristic functions have tails decaying at a rate slower than polynomial. This is the case of the normal. For these distributions, we adapted the previous method by estimatating the c.g.f. on a range where it is approximately linear, and "close enough" to the erratic region. So doing, the estimate of β gets larger in absolute value than in the polynomial case, and the optimal T_N is smaller. As the original method, this approach involves the subjective judgement of the researcher. In practice, however, we found that it yielded fairly good results.

It is to be noted that we chose this method for its practical convenience. Indeed, the trimming parameter T_N is analogous to a bandwidth in nonparametric density estimation. Many alternative methods exist, such as cross validation or minimization of a suitable criterion.

Choice of the estimator: Formula (4.18) gives many possible way of estimating factor characteristic functions. We start by discussing the issue related to the choice of the direction(s) of integration, then turn to the weighting of different estimators.

From a practical point of view, the choice of which direction t along which to perform the integration is not irrelevant. Indeed, let us choose $t \in \mathcal{T}_k$. Then one can estimate the second derivatibve of the c.g.f. of factor X_k by using the empirical counterpart of:

$$\kappa''_{X_k}(\tau_k) = \left(Q^{-}\right)_{[k,\cdot]} \nabla^2 \kappa_Y(\tau_k t),$$

for each $\tau_k \in \mathbb{R}$.

In practice, as illustrated by Figure 4.1, emprirical characteristic functions are well estimated around the origin and badly estimated in the tails. Moreover, when c.f's are close to zero, cumulant generating functions and their derivatives are also extremely badly estimated.

It thus makes sense to choose t such that

$$\nabla^2 \kappa_Y(\tau_k t)$$

is well estimated on a maximal interval. A natural choice is to minimize the euclidian norm of:

$$\frac{t}{t^T A_{[.,k]}},$$

which yields, by Cauchy-Schwartz inequality:

$$t = \left(A_{[\cdot,k]}\right)^{-T} = \frac{A_{[\cdot,k]}}{A^T_{[\cdot,k]} A_{[\cdot,k]}}. \tag{4.22}$$

The simulation section will provide evidence that choosing t as in (4.22) can result in better estimation properties. This method could be improved. For instance, if matrix A is not known but a consistent estimator is available, then one could weight the generalized inverses of Q and $A_{[.,k]}$ by the inverse of their precision, as in Generalized Least Squares estimation.

Lastly, one could also use (4.18) to weight estimators corresponding to different indices $t \in \mathcal{T}_k$. Our experiments on simulated data did not show any improvement over the simple case where t is chosen according to (4.22). The explanation for this result could come from the fact that different estimators corresponding to different indices are very correlated over the range where they are precisely estimated. In that case (for fixed τ), averaging is known to yield little efficiency gains.

4.5 Asymptotic properties

4.5.1 Consistency theorem

We now proceed to show that $\widehat{f}_{X_k}$ is a uniformly convergent estimator of f_{X_k}, for $k = 1...K$, provided that the characteristic functions of factors and errors are nonvanishing everywhere. In addition, Li and Vuong (1998) and Hall and Yao (2003) assume bounded supports.[4]

However, as recently emphasized by Hu and Ridder (2005), the two assumptions of bounded support and nonvanishing characteristic functions are mutually exclusive. As the c.f.'s of most standard distributions are never zero, and as economic variables often have unbounded support, we relax the assumption of support boundedness. We refer to Hu and Ridder (2005) for insights on how to deal with the case where c.f.'s are nonvanishing *almost*-everywhere.

To prove the consistency of our estimators in the case of unbounded support, we extend Hu and Ridder's Lemma 1, which proves a uniform consistency result for the empirical characteristic function, to the case of the empirical mean of functions

[4] Horowitz and Markatou (1996) do not assume support boundedness. However, as pointed out by Hu and Ridder (2005), their proofs implicitely require this hypothesis.

$X_n \exp(it^T Y_n)$, for a sample of N i.i.d. couples (X_n, Y_n), where Y_n is a random vector and X_n a scalar random variable.

To proceed, let us denote by $|t| = \max_\ell |t_\ell|$, for a vector $t \in \mathbb{R}^L$. In the Appendix, we prove the following lemma.

Lemma 14 *Let X be a scalar random variable and let Y be a vector of L scalar random variables. Let $Z = \left(X, Y^T\right)^T$. Let F denote the c.d.f. of Z ($\mathbb{E}$ denotes the corresponding expectation operator) and let F_N (resp. $\mathbb{E}_N$) denote the empirical c.d.f. (resp. mean) corresponding to a sample $\mathbf{Z}_N \equiv (Z_1, ..., Z_N)$ of N i.i.d. draws from F. Assume that $\mathbb{E}X^2 \leq M_1 < \infty$ and that $\mathbb{E}|Y|^i < \infty$ for all $i \in \{1, ..., L\}$. Define $f_t(x, y) = x \exp(it^T y)$ for $t \in \mathbb{R}^L$. Lastly, assume that there exists a constant $M > 0$ and a positive and decreasing function $k(\varepsilon)$ such that for $\varepsilon > 0$ small enough*

$$\mathbb{E}[|X| \mathbf{1}\{|X| > Mk(\varepsilon)\}] \leq \varepsilon.$$

Then,

$$\sup_{|t| \leq T_N} |\mathbb{E}_N f_t - \mathbb{E} f_t| = O(\varepsilon_N) \quad a.s., \tag{4.23}$$

for all ϵ_N, T_N such that

$$\frac{\varepsilon_N}{k(\varepsilon_N)} = \left(\frac{N}{\ln N}\right)^{\gamma - \frac{1}{2}}, \text{ and } T_N = O(N^r),$$

for all $0 < \gamma < 1/2$ close enough to zero, and all $r > 0$.

Lemma 14 shows that the rate of convergence of the empirical mean of f_t depends on the tails of the distribution of X: the slower the rate at which $|X|$ decays to zero, the slower the rate of convergence. For example, if X is Gaussian:

$$\frac{1}{\sqrt{2\pi}\sigma} \int_K^\infty x e^{-\frac{x^2}{2\sigma^2}} dx = \frac{1}{\sqrt{2\pi}} \sigma e^{-\frac{K^2}{2\sigma^2}},$$

and one can take $k(\varepsilon) = \ln(1/\varepsilon)$. If X is Pareto:

$$\int_K^\infty x\frac{ab^a}{x^{a+1}}dx = \frac{ab^a}{a-1}\frac{1}{K^{a-1}}, \quad a > 1.$$

In which case, $k(\varepsilon) = (1/\varepsilon)^{a-1}$ works.

We now apply Lemma 14 to the first two derivatives of the c.f. of a vector of random variables Y: $\mathbb{E}\left(Y_\ell \exp(it^TY)\right)$ and $\mathbb{E}\left(Y_\ell Y_m \exp(it^TY)\right)$, for $\ell, m = 1...L$. This allows us to prove the following uniform consistency result for factor characteristic functions.

Theorem 15 *Suppose that there exists an integrable, decreasing function $g_Y : \mathbb{R}^+ \to [0,1]$, such that $|\varphi_Y(t)| \geq g_Y(|t|)$ as $|t| \to \infty$. Then, there exists $\varepsilon_N \downarrow 0$ and $T_N \to \infty$ such that*

$$\sup_{|\tau|\leq T_N} \left|\widehat{\varphi}_{X_k}(\tau) - \varphi_{X_k}(\tau)\right| = \frac{T_N^2}{g_Y(T_N)^3}O(\varepsilon_N) = o(1) \quad a.s.,$$

where ε_N is the minimum convergence rate satisfying the conditions of Lemma 14 for all functions f_t of the form $\exp\left(it^TY\right)$, $Y_\ell \exp\left(it^TY\right)$ and $Y_\ell Y_m \exp\left(it^TY\right)$, $\ell, m \in \{1, ..., L\}$, and T_N satisfies two constraints: $T_N = O(N^r)$ for some $r > 0$, and $\frac{T_N^2}{g_Y(T_N)^3}\varepsilon_N = o(1)$.

Note that assuming the existence of g_Y such that $|\varphi_Y(t)| \geq g_Y(|t|)$ is a simple implication of assuming the existence of g_X, mapping $\mathbb{R}^+$ onto $[0,1]$, decreasing and integrable, such that $|\varphi_X(\tau)| \geq g_X(|\tau|)$ as $|\tau| \to \infty$. It suffices to take $g_Y(|t|) = g_X(L|A||t|)$ for $|A| = \max_{ij}|a_{ij}|$. Indeed,

$$\begin{aligned} |\varphi_Y(t)| &= \left|\mathbb{E}\left[e^{it^TY}\right]\right| = \left|\mathbb{E}\left[e^{it^TAX}\right]\right| \\ &= |\varphi_X(A^Tt)| \\ &\geq g_X(\left|A^Tt\right|) \\ &\geq g_X(L|A||t|), \end{aligned}$$

where $|A| = \max_{i,j}(|a_{ij}|)$. Function g_Y inherits g_X's properties: it maps $\mathbb{R}^+$ onto $[0,1]$, it is decreasing and it is integrable, so that in particular $g_Y(|t|) \to 0$ when $|t| \to \infty$.

Therefore, to guarantee the existence of g_Y, it suffices to assume that the tails of factor c.f.'s are bounded from below.

The following theorem states that $\widehat{f}_{X_k}$ converges uniformly to f_{X_k} when the sample size tends to infinity. It is proved in the Appendix.

Theorem 16 *Suppose that there exists an integrable, decreasing function $g_X : \mathbb{R}^+ \to [0,1]$ such that $|\varphi_X(\tau)| \geq g_X(|\tau|)$ as $|\tau| \to \infty$. Suppose also that there exist K integrable functions $h_{X_k} : \mathbb{R}^+ \to [0,1]$ such that $h_{X_k}(|\tau|) \geq \left|\varphi_{X_k}(\tau)\right|$ as $|\tau| \to \infty$. Then, $\widehat{f}_{X_k}$ is a uniformly convergent estimator of the d.f. f_{X_k} of X_k, i.e.*

$$\sup_x \left|\widehat{f}_{X_k}(x) - f_{X_k}(x)\right| = \frac{T_N^3}{g_X(T_N)^3} O(\varepsilon_N) + O\left(\int_{T_N}^{+\infty} h_{X_k}(v)dv\right) = o(1) \quad a.s. \quad , \tag{4.24}$$

where ε_N and T_N are given by Theorem 15 applied to $g_Y(|t|) = g_X(L|A||t|)$.

4.5.2 Convergence rates

Theorem 16 shows that the convergence rate of our estimator depends on the shape of factor and measurement distributions in two different ways. Firstly, as emphasized by Lemma 14, the rate of convergence of derivatives of factor c.f.'s depends on the tails of factor distribution functions. Secondly, the rate of convergence of factor densities varies with the tail of factor characteristic functions, controled by functions g_X and h_{X_k}.

In the rest of this section, we illustrate the second type of dependence, with respect to the smoothness of factor distributions. For simplicity, we assume that the tails of the measurement d.f.'s do not affect the rate of convergence; as shown above, this is the case if factor tails decay at a faster than polynomial rate (which would correspond to Pareto tails).

Fan (1991) distinguishes two kinds of distributions according to the tails of their c.f. *Smooth* distributions correspond to polynomial g_X and h_{X_k} functions. Examples are the uniform, gamma or Laplace. In contrast, the tails of characteristic functions of *supersmooth* distributions decay at a faster, exponential, rate. For instance, normal distributions are *supersmooth*.

To illustrate how the rate of convergence vary with the degree of smoothness, we now derive the convergence rate of our estimator in some simple cases. We start by noticing that equation (4.24) yields a trade-off for choosing T_N: The larger T_N the smaller the weight of factor density outside the interval $[-T_N, T_N]$, hence the smaller the second term on the RHS of (4.24). However, the larger T_N the more imprecise the estimation of the characteristic function, hence the larger the first term in (4.24). It thus makes sense to look for an "optimal" trimming parameter T_N.

To proceed, let us assume that the term in $O(\varepsilon_N)$ in (4.24) is known, and call it Δ_N. Then we propose to choose T_N minimizing:

$$\frac{T_N^3}{g_X(T_N)^3}\Delta_N + \int_{T_N}^{+\infty} h_{X_k}(v)dv, \tag{4.25}$$

under the constraint: $\frac{T_N^2}{g_X(L|A|T_N)^3}\varepsilon_N = o(1)$.

Let us now suppose that the d.f. of all factors X_k are *smooth*, so that the tails of their c.f. decay at a polynomial rate. One can suppose, for all $k \in \{1...K\}$:

$$|\tau|^{-\alpha_k} \leq \left|\varphi_{X_k}(\tau)\right| \leq |\tau|^{-\beta_k}, \quad |\tau| \to \infty,$$

with $1 < \beta_k < \alpha_k$.

Let $k \in \{1...K\}$. One can take $h_{X_k}(|\tau|) = |\tau|^{-\beta_k}$. Moreover, as factors X_k are mutually independent:

$$|\varphi_X(\tau)| = \prod_{k=1}^{K} \left|\varphi_{X_k}(\tau_k)\right| \geq \prod_{k=1}^{K} |\tau_k|^{-\alpha_k} \geq |t|^{-\sum_{k=1}^{K}\alpha_k}.$$

One thus can take $g_X(|\tau|) = |\tau|^{-\alpha}$, with $\alpha = \sum_{k=1}^{K} \alpha_k$.

Then the convergence rate (4.25) becomes:

$$T_N^{3(1+\alpha)}\Delta_N + \frac{1}{\beta - 1} T_N^{1-\beta}. \tag{4.26}$$

Minimizing (4.26) with respect to T_N yields:

$$T_N = \left(3(1+\alpha)\Delta_N\right)^{-1/(2+3\alpha+\beta)}.$$

Now, Lemma 14 shows that, for all distribution functions but the ones with very fat tails (*e.g.* Pareto) one can choose

$$\Delta_N = \left(\frac{\ln N}{N}\right)^{1/2-\gamma},$$

where $0 < \gamma < 1/2$ is close enough to zero. In this case:

$$T_N = O\left(\left(\frac{N}{\ln N}\right)^{\frac{1/2-\gamma}{2+3\alpha+\beta}}\right). \tag{4.27}$$

Note that T_N tends to infinity with N.

Let us evaluate (4.25) at T_N given by (4.27). Then, up to a multiplicative constant:

$$\begin{aligned} \frac{T_N^3}{g_X(T_N)^3}\Delta_N &= \left(\frac{N}{\ln N}\right)^{\frac{3+3\alpha}{2+3\alpha+\beta}(1/2-\gamma)} \times \left(\frac{\ln N}{N}\right)^{1/2-\gamma}, \\ &= \left(\frac{\ln N}{N}\right)^{\frac{\beta-1}{2+3\alpha+\beta}(1/2-\gamma)}. \end{aligned}$$

As this last term is $o(1)$, so is $\frac{T_N^3}{g_X(L|A|T_N)^3}\Delta_N$. Moreover:

$$\int_{T_N}^{+\infty} |t|^{-\beta} dv = \frac{1}{\beta-1} T_N^{1-\beta} = \left(\frac{\ln N}{N}\right)^{\frac{\beta-1}{2+3\alpha+\beta}(1/2-\gamma)}.$$

This implies that:

$$\sup_x \left| \widehat{f}_{X_k}(x) - f_{X_k}(x) \right| = O\left(\left(\frac{\ln N}{N} \right)^{\frac{\beta-1}{2+3\alpha+\beta}(1/2-\gamma)} \right) \quad a.s. \tag{4.28}$$

Note that, as $\beta > 1$ this quantity is $o(1)$.

Two remarks are in order. First, the convergence rate in (4.28) is polynomial in N. This is typical in deconvolution problems when both factor and error densities are smooth.

Second, the convergence rate increases with β, and decreases with α. This is also consistent with the findings of the deconvolution literature: the convergence rate is higher, the fatter the tail of the factor distribution relative to the tail of the error distribution.

When the tail of the error characteristic function is considerably thiner than the tail of the factor c.f., the deconvolution literature shows that the convergence rate of the density estimator can be extremely slow. See Caroll and Hall (1988) and Horowitz and Markatou (1996) for estimators with logarithmic rates of convergence. We now show that this undesirable property carries over to our estimator.

We consider a case where factor X_k is smooth, and there exits one *supersmooth* factor other than X_k. So that $h_{X_k}(|\tau|) = |\tau|^{-\beta_k}$ is still a valid higher bound for the c.f. of X_k, but g_X is no longer polynomial. In the supersmooth case, one can take $g_X(|\tau|) = exp\left(-|\tau|^\alpha\right)$, with $\alpha > 0$.

Now, maximizing (4.25) with respect to T_N yields:

$$(3T_N^2 + 3\alpha T_N^{\alpha+2}) exp(3T_N^\alpha)\Delta_N - T_N^{-\beta} = 0.$$

Taking $\Delta_N = \left(\frac{\ln N}{N}\right)^{1/2-\gamma}$ as before, the "optimal" T_N satisfies:

$$T_N = O\left((\ln N)^{1/\alpha-\eta}\right),$$

for all $\eta > 0$ close enough to zero.[5]

Then the second term in (4.25) is *logarithmic* in N, and so is the convergence rate of $\widehat{\varphi}_{X_k}$.

Remark: It is tempting to use (4.27) as a guideline to choose T_N in practice. However, our experiments suggest that doing so, ones underestimates T_N. The reason might be that to choose the "optimal" T_N one maximizes an *upper bound* for the convergence rate. In practice, this upper bound may overestimate the true rate. We found much better estimation results by using the simple method outlined in 4.4.2 to choose T_N. Refining the bound further is an interesting issue that we leave for future research.

4.6 Monte-Carlo simulations

In this section, we study the finite-sample behavior of our density estimators.

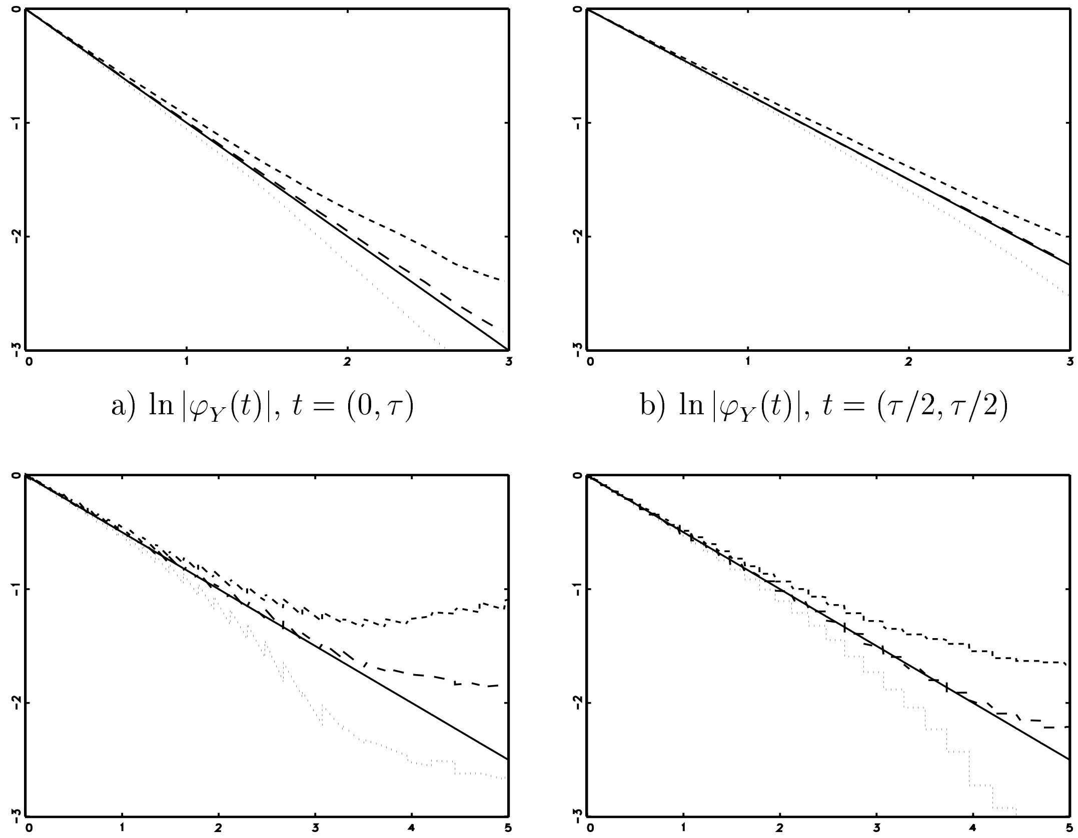
Characteristic functions: We first consider the estimation of factor characteristic functions. We illustrate the approach presented in 4.4.2 by means of the measurement error model:

$$\begin{aligned} Y_1 &= X_1 + U_1, \\ Y_2 &= X_1 + U_2, \end{aligned}$$

where $(X_1, U_1, U_2) \in \mathcal{N}(0, I_3)$. Figure 4.2, panels a) and b), presents estimates of the logarithm of the empirical characteristic function of Y:

$$\ln |\widehat{\varphi}_Y(t)| = \ln \left| \mathbb{E}_N \left[e^{it^T Y} \right] \right|,$$

[5]In this case, the constraint $\frac{T_N^2}{g(T_N)^3} \Delta_N = o(1)$ is binding. One has thus to redefine T_N even further as $O\left((\ln N)^{\frac{1/2-\gamma}{\alpha}} \right)$.



a) $\ln|\varphi_Y(t)|$, $t=(0,\tau)$ b) $\ln|\varphi_Y(t)|$, $t=(\tau/2,\tau/2)$

c) $\ln\left|\varphi_{X_1}(\tau)\right|$, $t=(0,\tau)$ d) $\ln\left|\varphi_{X_1}(\tau)\right|$, $t=(\tau/2,\tau/2)$

Figure 4.2: Estimation of characteristic functions, measurement error model

evaluated at $(t_1,t_2)=(0,\tau)$ and $(t_1,t_2)=(\tau/2,\tau/2)$, respectively, where τ^2, for $\tau\in\mathbb{R}^+$, is reported on the x-axis of the figure.

For each simulation in this section, we draw 100 independent realizations of Y. The thick line is the true factor characteristic function and the dashed line is the pointwise median of the 100 estimates. The dotted lines correspond to the pointwise first and ninth estimate deciles.

Here, the true value for the c.f. is represented by a line with slope -1 in panel a), and $-3/4$ in panel b). This is because the c.f. of the standard normal is $\exp(-t^2/2)$.

Figure 4.2 show that the true c.f.'s are well estimated over a wide range. However, the precision of the estimation in the tails is lower. This feature has important consequences in practice, as we now illustrate.

Then, in Figure 4.2, panel c), we report estimates the c.f. of X_1, obtained by the Li and Vuong method; that is: integrating along the direction $(0, \tau)$, for $\tau \in \mathbb{R}$. Panel d) of the same figure presents the estimate of factr c.f. obtained by the method of this chapter, where the direction of integration is $(\tau/2, \tau/2)$.

We draw two conclusions from the figure. First, the estimation of the c.f. of X_1 is much lower than that of the c.f. of Y. The reason is that small errors in the estimation of the tail of a c.f. translate into large errors for its derivative.

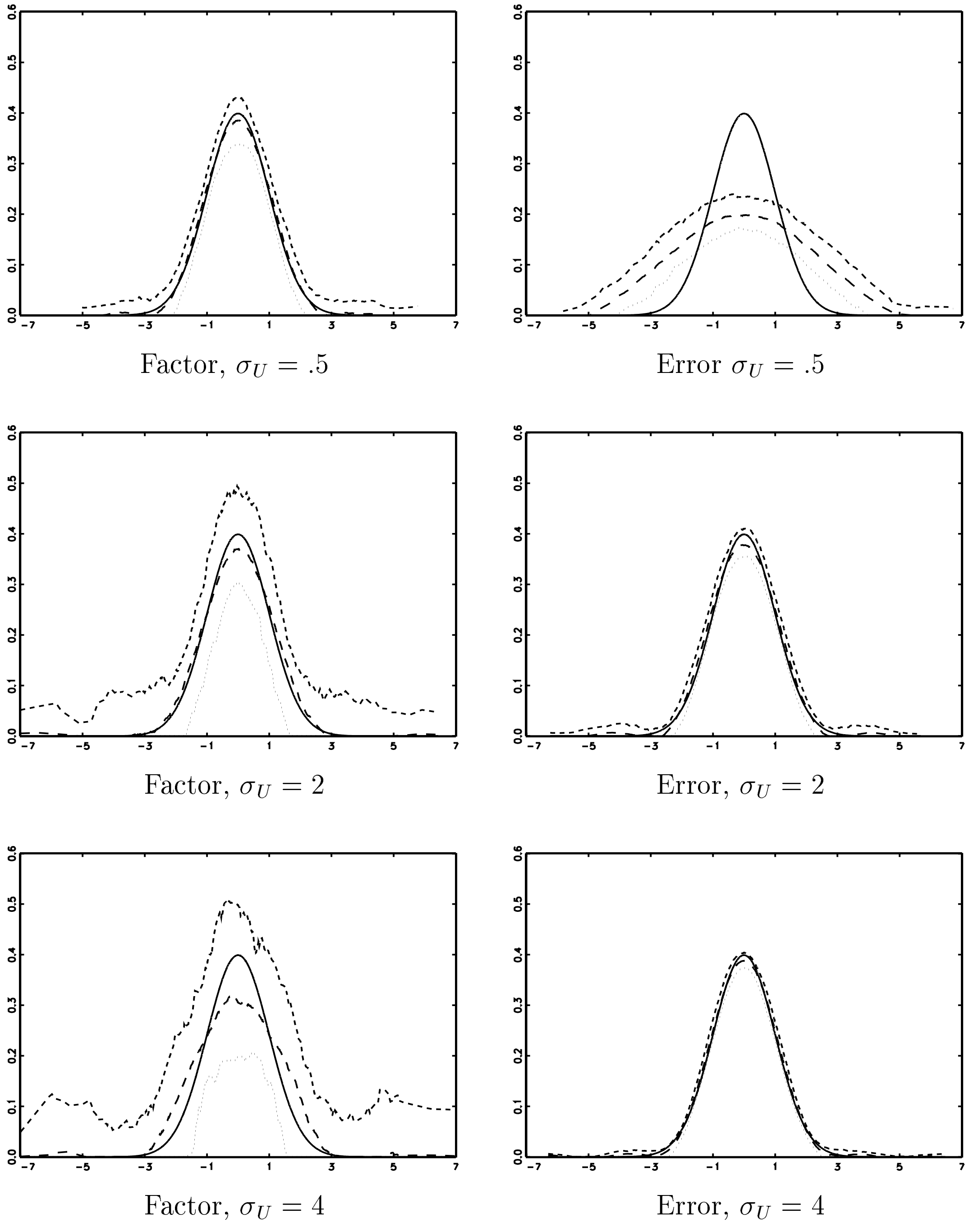
Second, the c.f. of X_1 is somewhat better estimated by the second method, using as direction of integration the vector of minimal euclidian norm. This observation is in line with the discussion in 4.4.2.

The normal case: Let us now consider the factor model: $Y = \Lambda X + U$, with $Y \in \mathbb{R}^3$ $(L = 3)$, $X \in \mathbb{R}^3$ is the vector of common factors and $U \in \mathbb{R}^3$ is the vector of errors. We set

$$\Lambda = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{pmatrix},$$

and assume that both factors and errors follow the same distribution. For reasons of symmetry, we shall present the estimation results for the first factor and first error component only.

In all figures, the left panel corresponds to the first factor X_1, and the right panel to the first error U_1. The thick line is the true factor (resp. error) distribution and the dashed line is the pointwise median of 100 estimates. As before, the dotted lines correspond to the pointwise first and ninth estimate deciles. Lastly, sample size is



Factor, $\sigma_U = .5$ | Error $\sigma_U = .5$

Factor, $\sigma_U = 2$ | Error, $\sigma_U = 2$

Factor, $\sigma_U = 4$ | Error, $\sigma_U = 4$

Figure 4.3: Density estimation in the normal case, factor model with $L = 3$, $K = 6$

$N = 1000$.

In this paragraph, factors follow the standard normal distribution, and errors are normally distrinbuted with different standard deviations σ_U. Figure 4.3 presents the estimation results for σ_U equal .5, 2 and 4, respectively.

The results illustrate that the estimation of the factor density is poorer, the larger the size of the noise. Note that for moderate error size (*e.g.* $\sigma_U = 2$) there is little finite sample bias. However, confidence bands are rather large. This result is consistent with the fact that the convergence rate of our estimator is *a priori* slower than root-N (see 4.5).

Smooth and supersmooth distributions: In Figure 4.4, both factors and errors follow the double exponential (Laplace) distribution with parameter 1, the density of which is $f_X(x) = 1/\sqrt{2}\exp(-\sqrt{2}|x|)$. The error standard deviation is 2.

The characteristic function of the Laplace distribution is $\varphi_X(t) = 1/(1 + \frac{1}{2}t^2)$. It is an example of a smooth distribution, as defined by Fan (1991).

Figure 4.4 shows that in this case the general shape of the double exponential is well reproduced, albeit not perfectly. In particular, the curvature of the d.f. around zero is fitted somewhat imprecisely.

In the two bottom panels of Figure 4.4, factors are still Laplace, and errors are normally distributed with a standard deviation of 2. In this case, factors are smooth and errors supersmooth. Therefore, we expect the convergence rate of our estimator to be extremely low (see Section 4.5).

Figure 4.4 suggests that the finite sample results are not so different from the previous case (smooth factors and errors). Thus, in our simulations the well-known problem of deconvoluting a d.f. in the presence of a smoother distribution does not seem to affect the estimation results very much.

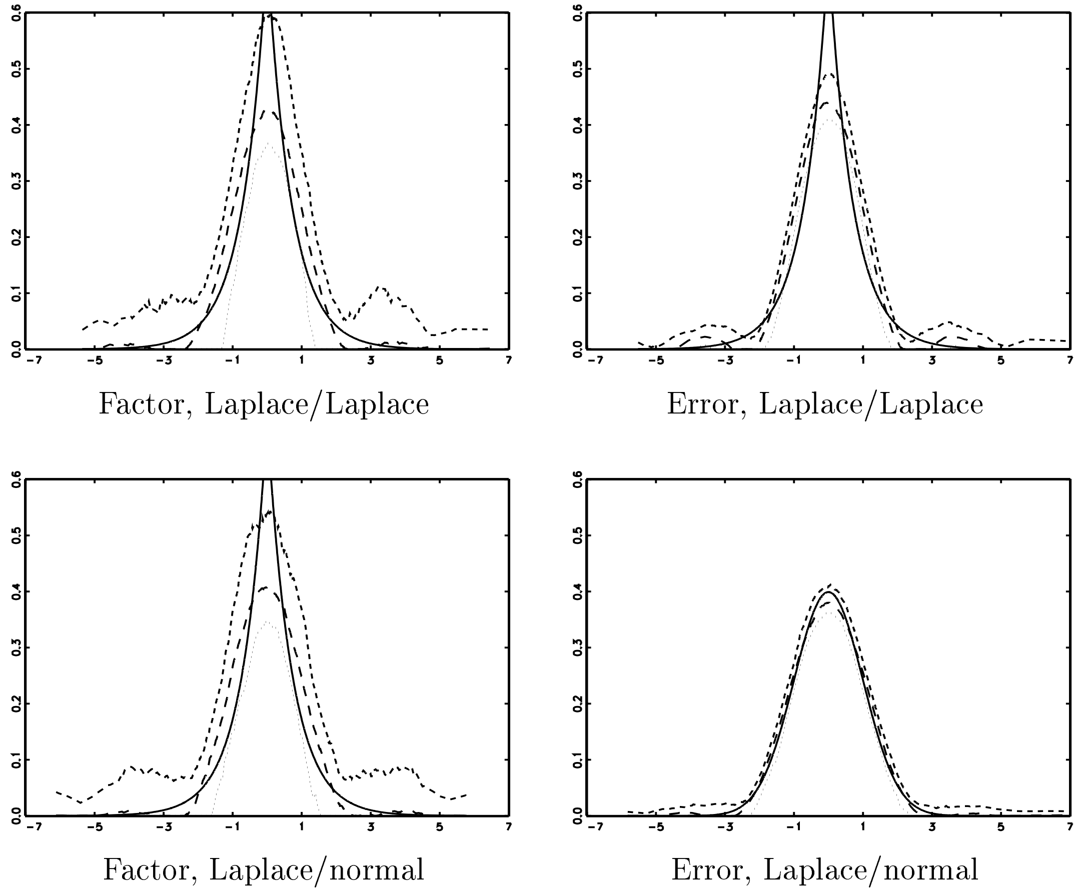


Figure 4.4: Density estimation in the Laplace/Laplace and Laplace/normal cases, factor model with $L = 3$, $K = 6$, $\sigma_U = 2$

Skewness and kurtosis: In the rest of this section, we study the ability of our estimator to deal with skewed and/or kurtotic distributions. In all remaining simulations, factors and errors follow the same distribution up to scale, and σ_U is 2.

In the four upper panels of Figure 4.5, factors are (standardized) gamma distributed, with parameters (5,1) and (2,1), respectively. For these values of the parameters, factor skewness is $2/\sqrt{5} \approx .89$ and $\sqrt{2}$, respectively, and factor kurtosis excess is 1.2 and 3.

The results suggest that our estimator captures skewness reasonably well. However, the estimation is less precise when skewness is larger (the second row of Figure 4.5).

Next, we turn to the impact of kurtosis. In the third row of Figure 4.5, factors are mixtures of independent normals.[6] The kurtosis excess of factor densities is: $\kappa = 100$.[7]

Factor densities are still globally well estimated in this case. However, neither the mode nor the tails of factor density is precisely estimated.

Lastly, we report in the last row of Figure 4.5 the simulation results for log-normally distributed factors. Their skewness and kurtosis excess are approximately 6.2 and 110, respectively. It is clear from the figure that factor densities are badly estimated in this case. In particular, the estimated tails present very large oscillations.

We conclude from this exercise that the finite-sample performance of our estimator critically depends on the shape of the distributions to be estimated.

Better finite-sample properties could certainly be achieved at the cost of complicating the calculation of the estimator, for instance by using splines or wavelets to

[6] More precisely, we construct factors as mixtures of two independent normals. Let $W_1 \sim N(0,1/2)$, and let $\rho \in]0,1[$. Define $W_2 \sim N(0,(2-\rho)/(2-2\rho))$, independent of W_1. Then it is straightforward to see that X define as the mixture of (W_1,ρ) and $(W_2,1-\rho)$ has variance one, and kurtosis excess $\kappa_4(\rho) = 3\rho/(4(1-\rho))$.

[7] That is: $\rho = 400/403$.

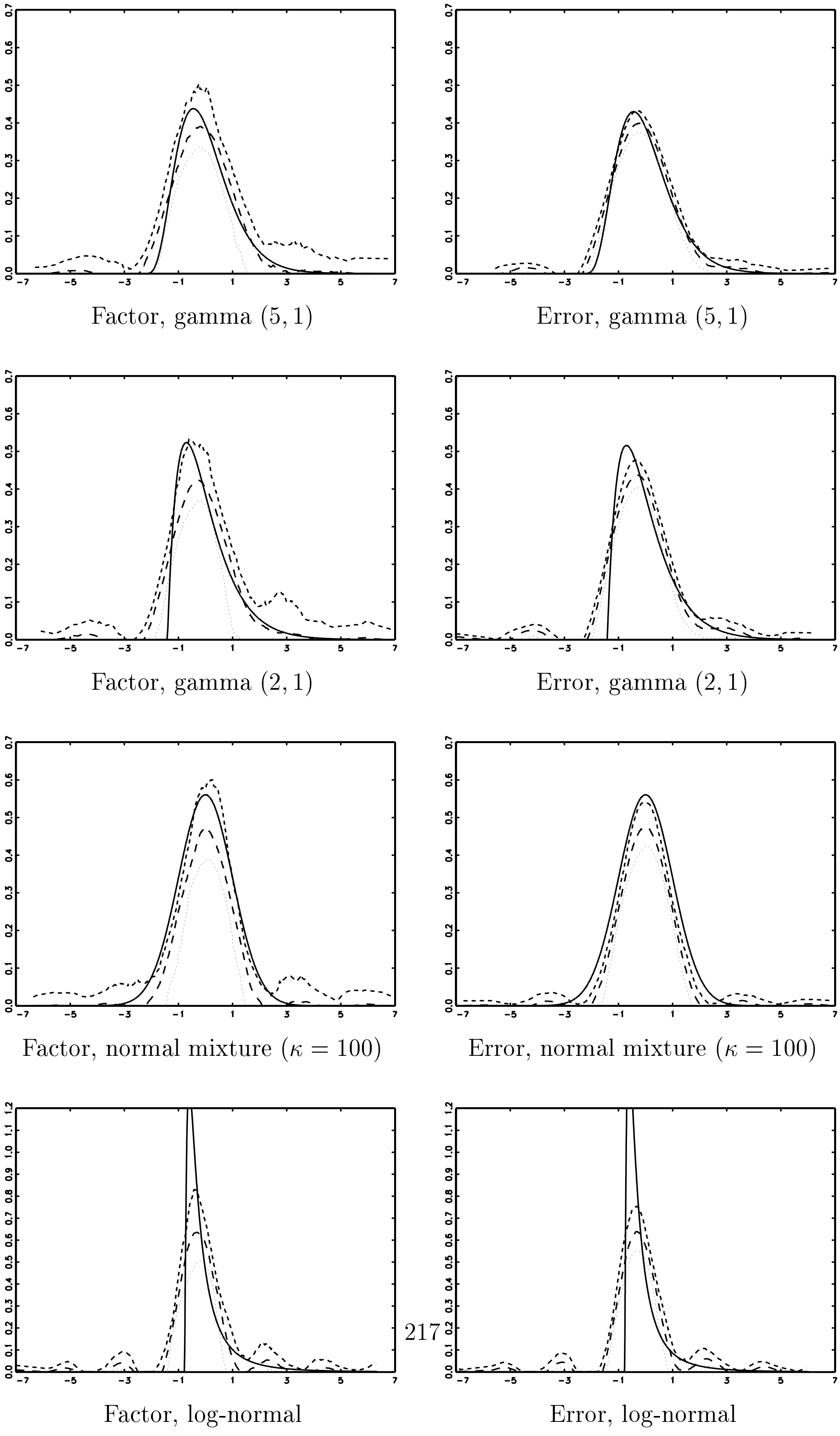


Figure 4.5: Density estimation in the presence of skewness and/or kurtosis, factor model with $L = 3$, $K = 6$, $\sigma_U = 2$

smoothen the estimator (*e.g.* Fan and Koo, 2002). We leave the comparison of these improvements to future research.

4.7 Application: a two-factor model of the returns to schooling

In this section, we apply our methodology to a linear factor model of three individual variables: wages and two measures of education. We use data from the French Labor Force Survey for 1995. This is a large and representative cross-section of the French labor force which provides detailed information on individual education. We exclude women, out-of-employment individuals, and workers with missing data for either (monthly) wages, hours worked or education. We trim the sample of the first and last percentiles of the wage, hour and education data. We finally obtain a sample of 21,794 workers. We divide monthly wages by hours worked to obtain wage rates. We define Y as the residual of the regression of wage rates on a quartic in age. The first education variable is the "age at the end of school", which broadly corresponds to the number of years of schooling (minus 6) in France. This variable, denoted as D, is the usual regression variable in most studies of the returns to schooling. The second one codes the highest diploma obtained by the individual into 16 categories (no diploma, elementary level, middle school, high school, college, plus various declinations of these different levels into vocational and non vocational). To make this variable continuous and comparable to D, we construct the variable D^* equal to the median value of D by diploma. Doing so, we obtain a variable that has a lower mean than D, the negative wedge corresponding to class repetition, hesitations about education tracks, etc.

In chapter 3, we have shown that the data can be described by the two factor model:

$$\begin{cases} Y = .17X_1 - .14X_2 + U_1, \\ D = 2.1X_1 + .40X_2 + U_2, \\ D^* = 2.0X_1 + .40X_2 + U_3, \end{cases}$$

where the standard errors of U_1, U_2, U_3 are .040, 2.5 and .50, respectively. We also find that X_1 has skewness 1.2 and kurtosis .70 and that X_2 has skewness .10 (insignificant from zero) and kurtosis 16.

The upper panel in Figure 4.6 shows the estimated density of the two factors X_1 and X_2. The thick line is the standard normal density. The dotted lines delimit the bootstrapped $10\% - 90\%$ confidence band.

The density of X_1 is well estimated, and seems signifiantly skewed. In contrast, the density of X_2 is imprecisly estimated. The very large oscillation in the left tail is of particular concern. This bad estimation is in line with our Monte-Carlo results, as the second factor represents around 1% of the total variance of Y. On the contrary, the first factor represents 57% of total variance. However, despite the lack of precision, the estimation result suggests that X_2 is significantly peaked. This is not inconsistent with the estimates of skewness and kurtosis reported above.

Then, in the second row of Figure 4.6 we plot the three error densities. The first and second error seem close to normally distributed. In contrast, the third error is badly estimated with large oscillations in the two tails. This is not surprising, as this variable is constructed from D^* which takes discrete values.

We now use these results to predict the values of the two factors for each individual in the sample. This allows us to correlate the unobserved factors with other (observed) labor market outcomes than the wage. To do so, we estimate:

$$\mathbb{E}(X_1|Y, D, D^*), \quad \text{and} \quad \mathbb{E}(X_2|Y, D, D^*), \tag{4.29}$$

and then correlate these quantities with three covariates:

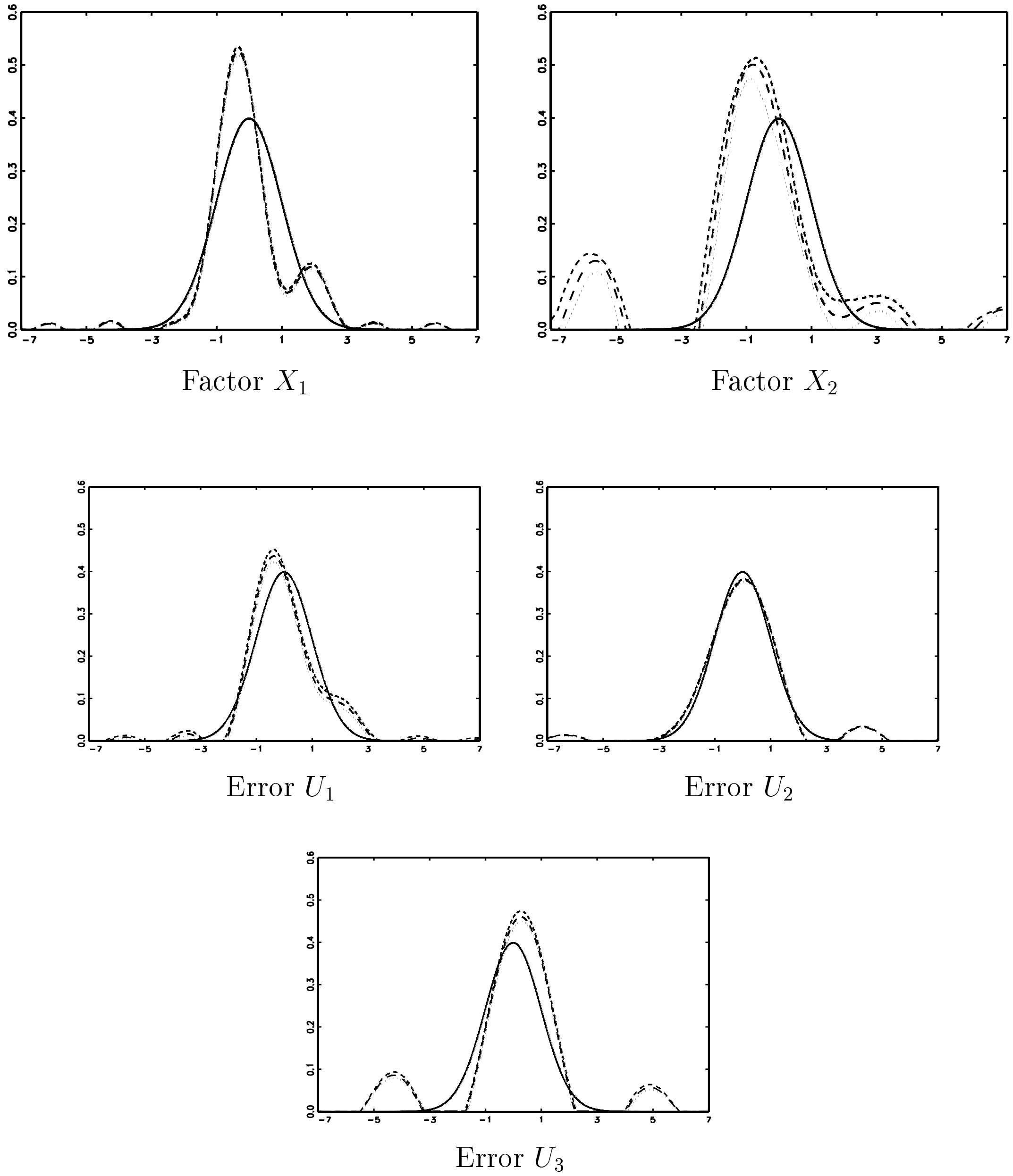


Figure 4.6: Estimates of factor and error densities in the two-factor model of returns to schooling

1. A dummy variable indicating if the individual works or not in the public sector (PUB).

2. A variable indicating the economic status of the son (CS): manual worker (1), intermediate profession (2) or manager (3).

3. A variable indicating the economic status of the father (CF), classified similarly as CS.[8]

The conditional expectations in (4.29) are functions of the whole distributions of factors and errors, not only of their first two moments as the following expressions show. For all integrable bivariate function g:

$$\mathbb{E}[g(X_1, X_2)|\,(Y, D, D^*) = (y, d, d^*)] = \frac{\int g(x_1, x_2) f_U(y, d, d^*|x_1, x_2) f_{X_1}(x_1) f_{X_2}(x_2) dx_1 dx_2}{\int f_U(y, d, d^*|x_1, x_2) f_{X_1}(x_1) f_{X_2}(x_2) dx_1 dx_2},$$

where:

$$f_U(y, d, d^*|x_1, x_2) \equiv f_{U_1}(y - \lambda_{11} x_1 - \lambda_{12} x_2) f_{U_2}(d - \lambda_{21} x_1 - \lambda_{22} x_2) f_{U_3}(d^* - \lambda_{31} x_1 - \lambda_{32} x_2).$$

To estimate (4.29), we replace the factor and error densities by their nonparametric estimates and set $g(x, y) = x$ and $g(x, y) = y$ for the two conditional expectations, respectively.

Having computed the expectations in (4.29), we correlate them with variables CS, CF and PUB. Table 4.1 displays the results. Overall, the public sector and high skill occupations attract individuals with higher X_1 and lower X_2. Moreover, public employees and high skill workers have on average higher wages and higher education (as measured by both D and D^*). As far as father's occupation is concerned, we find the usual human capital transfer effects: children of managers have on average higher

[8]The father's wage would have been another possible choice. However, it is not available in the French Labor Force Survey.

Table 4.1: Means of observed and unobserved covariates for various categories of individuals

	N	Y	D	D^*	X_1	X_2
Private	16270	-.03	17.6	17.5	-.06	.05
Public	5524	.09	18.0	17.9	.17	-.14
Low skill	14253	-.11	16.9	16.9	-.36	.13
Intermediate skill	5271	.14	18.6	18.5	.43	-.15
High skill	2070	.39	20.5	20.5	1.34	-.50
Low skill father	13024	-.02	17.2	17.3	-.16	-.02
Intermediate skill father	4090	.05	18.5	18.3	.29	.04
High skill father	1551	.18	20.3	19.9	1.02	.04

X_1 than children of manual workers. They also have higher wages and are bettter educated. Father's occupation is not correlated with X_2.

4.8 Conclusion

This chapter provides a generalization of the nonparametric estimator of Li and Vuong (1998) to the case of a general linear independent factor structure, allowing for any number of measurements, L, and at most $\frac{L(L+1)}{2}$ factors (including errors). The main lessons of the standard deconvolution literature carry over into the more general context that we consider in this paper. Convergence rates are slow; it is easier to identify the distribution of a smooth factor; and it is easier to identify the distribution of one factor if the other factor distributions are not smooth.

Our Monte Carlo results yield interesting insights. First, our simple trimming procedure manages to reduce the Gibbs oscillations in the tails of the estimated densities. Second, the finite-sample performance of our estimator critically depends on the shape of the distributions to be estimated (smoothness and tails properties).

We also find that it is easier to identify smooth distributions with little kurtosis excess. In Chapter 3, we have shown that skewness and peakedness are required for the matrix of factor loadings to be identified from higher-order moments. There is thus a tension between obtaining a precise estimate of factor loadings and a precise estimate of the distribution of factors. However, identifying the distributions of three factors and three errors from a panel of 3 observations for 1,000 individuals should be viewed as a considerably more complicated problem than the prototypical measurement error problem. Given the difficulty of the problem at hand, we view these simulation results as a confirmation that the nonparametric deconvolution approach can be successfully applied to a wide range of distributions.

Appendix A

Appendix of Chapter 1

A.1 Parametric copulas

In this appendix, we review the copulas mentioned in subsection 1.2.2. We refer the interested reader to Joe (1997), Nelsen (1999) and Drouet-Mari and Kotz (2001) for further reading.

Genest and MacKay (1986) introduce the Archimedean family of copulas. Let ϕ be a convex decreasing function from $]0,1]$ to $\mathbb{R}_+$ such that $\phi(1) = 0$. Then the bivariate function:

$$C(u,v) = \phi^{-1}(\phi(u) + \phi(v))$$

is a copula. Such Archimedean copulas are useful, to the extent that they reduce the study of bivariate distributions to the research of univariate functions, their generator ϕ. Below, we list the Archimedean copulas we use in the chapter:

- Frank:
$$\phi(u;\tau) = \ln\left(\frac{1-e^{-\tau}}{1-e^{-\tau u}}\right), \quad \tau > 0.$$
- Gumbel:
$$\phi(u;\tau) = (-\ln(u))^{\tau}, \quad \tau \geq 1.$$
- Clayton:
$$\phi(u;\tau) = \left(\frac{1}{u}\right)^{\tau} - 1, \quad \tau > 0.$$
- Joe:
$$\phi(u;\tau) = -\ln\left(1 - (1-u)^{\tau}\right), \quad \tau \geq 1.$$
- Log-copula (Genest and Rivest, 1993):
$$\phi(u;\tau) = \left(1 - \frac{\ln(u)}{\tau_1\tau_2}\right)^{\tau_1+1} - 1, \quad \tau = (\tau_1, \tau_2), \quad \tau_1 > 0, \quad \tau_2 > 0.$$

  This is the only two-parameter family we present in this chapter.

We also consider three non Archimedean copulas:

- Gaussian:
$$C(u,v;\tau) = \Phi_2\left(\Phi^{-1}(u), \Phi^{-1}(v); \tau\right)$$
where $\Phi_2(x,y;\tau)$ is the cdf of a couple of normal random variables $\mathcal{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \tau \\ \tau & 1 \end{pmatrix}\right)$.

- FGM (Farlie, Gumbel, Morgenstern):
$$C(u,v;\tau) = uv\left(1 + \tau(1-u)(1-v)\right), \quad -1 \le \tau \le 1.$$

- Plackett:
$$C(u,v;\tau) = \frac{1}{2}\tau^{-1}\left\{1 + \tau(u+v) - [(1+\tau(u+v))^2 - 4\tau(\tau+1)uv]^{1/2}\right\}, \quad \tau \ge -1.$$

A.2 The Plackett Copula

Plackett (1965) generalizes the independence condition for contingency tables. He shows that, if U and V are uniform r.v. on $[0,1]$, then the following equation:

$$\frac{P(U \le u, V \le v)P(U > u, V > v)}{P(U \le u, V > v)P(U > u, V \le v)} = \eta + 1 \quad \forall (u,v), \tag{A.21}$$

where $\eta > -1$ is a given constant, has one single solution. This solution is a copula, and writes:

$$C(u,v;\eta) = \frac{1}{2}\eta^{-1}\left\{1 + \eta(u+v) - [(1+\eta(u+v))^2 - 4\eta(\eta+1)uv]^{1/2}\right\}.$$

>From (A.21), η is a natural mobility index. More precisely, let us define the following ordering $\precsim_c$ on copulas, called the *concordance ordering* (e.g. Joe, 1997):

$$C_1 \precsim_c C_2 \quad \text{iff} \quad C_1(u,v) \le C_2(u,v), \quad \forall (u,v).$$

$\precsim_c$ is the first-order stochastic dominance ordering. It measures relative mobility: the ranks process governed by copula C_1 will be said more mobile than the ranks process governed by C_2 if $C_1 \precsim_c C_2$. Moreover, the concordance ordering possesses a lower and a upper bound (Fréchet, 1935). The lower bound C_L satisfies: $C_L(u,v) = \max(u+v-1, 0)$. The upper bound C_U satisfies: $C_U(u,v) = \min(u,v)$.

The Plackett copula satisfies the following properties (Joe, 1997):

1. $C(.,.;\eta_1) \precsim_c C(.,.;\eta_2)$ for all $\eta_2 > \eta_1$.
2. $C(.,.;\eta) \to C_L$ when $\eta \to -1$.
3. $C(.,.;\eta) \to C_U$ when $\eta \to \infty$.
4. $C(.,.;\eta) \to C^{\perp}$ when $\eta \to 0$, where $C^{\perp}(u,v) = uv$ is the independent copula.

Therefore, the Plackett copula is *mobility decreasing* with respect to its parameter, and the Plackett family covers the whole range of bivariate dependence, from immobility (C_U) to independence ($C^{\perp}$) and perfect mobility (C_L). These properties are shared by the Gaussian and the Frank copulas for instance, which makes these families good candidates for modelling empirical ranks processes.

A.3 A sequential EM algorithm for copula models

We consider finite discrete mixtures of latent distributions for first-order Markov earnings trajectories ($\mathbf{y}_n = (y_{nt}, t = 1, ..., T), n = 1, ..., N$). Let $z_n \in \{1, ..., K\}$ be a latent variable indicating which group individual n belongs to. The density of $\mathbf{y}_n$ given $z_n = k$ is

$$f(\mathbf{y}_n; \beta_k, \tau_k) = \prod_{t=1}^{T} f(y_{nt}; \beta_k) \cdot \prod_{t=1}^{T-1} c\left[F(y_{nt}; \beta_k), F(y_{n,t+1}; \beta_k); \tau_k\right]. \quad \text{(A.32)}$$

We denote as π_k the proportion of individuals in Group k in the population and β_k and τ_k are parameter vectors.

We start by making the following assumptions:

Assumptions:

1. The parameter space, i.e. the set of possible values of $\boldsymbol{\beta}$ and $\boldsymbol{\tau}$, is compact.

2. True weights are positive, $\pi_k^0 > 0$, for all $k = 1, ..., K$.

3. The true values of the parameters, $\boldsymbol{\beta}^0 = \left(\beta_1^0, ..., \beta_K^0\right)$, $\boldsymbol{\tau}^0 = \left(\tau_1^0, ..., \tau_K^0\right)$ and $\boldsymbol{\pi}^0 = \left(\pi_1^0, ..., \pi_K^0\right)$, belong to the interior of the parameter space.

4. The marginal densities $f(y; \beta)$ are continuously differentiable with respect to β and so are the copula densities $c(u, v; \tau)$ with respect to τ.

5. The unknown latent parameters $\boldsymbol{\beta} = (\beta_1, ..., \beta_K)$ of the marginal densities $f(y_{nt}; \beta_k)$ are identified, up to a permutation, from cross-section clusters, i.e. for all k,

$$f(y; \beta_k^0) = f(y; \beta_k^1), \ \forall y \Rightarrow \beta_k^0 = \beta_k^1.$$

6. The whole set of unknown latent parameters $\boldsymbol{\beta} = (\beta_1, ..., \beta_K)$ and $\boldsymbol{\tau} = (\tau_1, ..., \tau_K)$ and the weights $\boldsymbol{\pi}=(\pi_1, ..., \pi_K)$ are identified, up to a permutation, from couples of cross-sections, i.e.

$$\begin{aligned} \sum_{k=1}^{K} \pi_k^0 f(\mathbf{y}; \beta_k^0, \tau_k^0) \quad &= \quad \sum_{k=1}^{K} \pi_k^1 f(\mathbf{y}; \beta_k^1, \tau_k^1), \quad \forall \mathbf{y} = (y_1, ..., y_T) \\ &\Rightarrow \quad \boldsymbol{\beta}^0 = \boldsymbol{\beta}^1, \boldsymbol{\tau}^0 = \boldsymbol{\tau}^1, \boldsymbol{\pi}^0 = \boldsymbol{\pi}^1. \end{aligned}$$

A sequential EM algorithm for discrete mixtures of copulas To estimate the parameter vector $\theta = \{\beta_k, \tau_k, \pi_k,\ k = 1, ..., K\}$ we apply the following (modified) EM algorithm.

1. **(Expectation- or E-stage)** For an initial value $\theta^{(s)}$ of θ and all $k \in \{1, ..., K\}$, compute the posterior probability of $z_n = k$ given $\mathbf{y}_n$ and $\theta^{(s)}$ as

$$p_{kn}^{(s)} = \frac{\pi_k^{(s)} f(\mathbf{y}_n; \beta_k^{(s)}, \tau_k^{(s)})}{\sum_{\ell=1}^{K} \pi_\ell^{(s)} f(\mathbf{y}_n; \beta_\ell^{(s)}, \tau_\ell^{(s)})}. \tag{A.33}$$

2. **(Maximization- or M- stage)** Successively update $\beta_k^{(s)}$, $\tau_k^{(s)}$ and $\pi_k^{(s)}$ as

$$\beta_k^{(s+1)} = \arg\max_{\beta_k} \sum_{n=1}^{N} p_{kn}^{(s)} \sum_{t=1}^{T} \ln f(y_{nt}; \beta_k), \tag{A.34}$$

$$\tau_k^{(s+1)} = \arg\max_{\tau_k} \sum_{n=1}^{N} p_{kn}^{(s)} \sum_{t=1}^{T-1} \ln c \left[F\left(y_{nt}; \beta_k^{(s+1)}\right), F\left(y_{n,t+1}; \beta_k^{(s+1)}\right); \tau_k\right], \tag{A.35}$$

$$\pi_k^{(s+1)} = \frac{1}{N} \sum_{n=1}^{N} p_{kn}^{(s)}. \tag{A.36}$$

This algorithm differs from the standard EM algorithm since $\beta_k^{(s)}$ and $\tau_k^{(s)}$ are not simultaneously updated by maximizing $\sum_{n=1}^{N} p_{kn}^{(s)} \ln f(\mathbf{y}_n; \beta_k, \tau_k)$ jointly with respect to both β_k and τ_k, but sequentially. Proceeding this way renders the maximization stage considerably more tractable. This is because the copula part of $\ln f(\mathbf{y}_n; \beta_k, \tau_k)$ depends on β_k through the marginal cdf's $F(y_{nt}; \beta_k)$ and $F(y_{n,t+1}; \beta_k)$, and therefore in general in a very non linear way. The algorithm is easier to implement and also much faster.[1] This idea has been used many times under various forms to simplify the practical implementation of fixed-point algorithms. In particular, another type of sequential EM algorithm has recently been used in the context of finite mixtures of distributions by Arcidiacono and Jones (2003).

Notice that, unlike the standard case considered by Dempster *et al.* (1977), the sample likelihood does *not* increase at each new iteration of the sequential algorithm. This caveat has two important consequences. First, the numerical convergence of the sequential estimator is not guaranteed. However, the usual sufficient conditions for convergence are strong (Dempster *et al*, Wu 1983) and rarely verified in practice. Numerical convergence is rather assumed than formally proved in most cases. Second, the sequential estimator differs from the MLE.

[1]The slow numerical convergence of EM is usually thought to be the main drawback of the algorithm (see e.g. Redner and Walker, 1984). However, we found our sequential modification very fast to converge, even with large datasets.

The numerical limit of the algorithm. The sequential EM algorithm writes as a set of first-order equations:

$$\sum_{n=1}^{N}\left(p_k\left(\mathbf{y}_n;\theta^{(s)}\right)\sum_{t=1}^{T}\frac{\partial}{\partial\beta}\ln f(y_{nt};\beta_k^{(s+1)})\right)=0$$

$$\sum_{n=1}^{N}\left(p_k\left(\mathbf{y}_n;\theta^{(s)}\right)\sum_{t=1}^{T-1}\frac{\partial}{\partial\tau}\ln c\left[F\left(y_{nt};\beta_k^{(s+1)}\right),F\left(y_{n,t+1};\beta_k^{(s+1)}\right);\tau_k^{(s+1)}\right]\right)=0$$

$$\pi_k^{(s+1)}=\frac{1}{N}\sum_{n=1}^{N}p_k\left(\mathbf{y}_n;\theta^{(s)}\right)$$

where

$$p_k\left(\mathbf{y}_n;\theta^{(s)}\right)=\frac{\pi_k^{(s)}f(\mathbf{y}_n;\beta_k^{(s)},\tau_k^{(s)})}{\sum_{m=1}^{K}\pi_m^{(s)}f(\mathbf{y}_n;\beta_m^{(s)},\tau_m^{(s)})}. \tag{A.37}$$

We suppose that the sequence $\theta^{(s)}=(\beta_k^{(s)},\tau_k^{(s)},\pi_k^{(s)},\ k=1,...,K)$ converges numerically to the sequential estimator $\theta^N=(\beta_k^N,\tau_k^N,\pi_k^N,\ k=1,...,K)$, solution to the following system of equations:

$$\frac{1}{N}\sum_{n=1}^{N}\left(p_k\left(\mathbf{y}_n;\theta^N\right)\sum_{t=1}^{T}\frac{\partial}{\partial\beta}\ln f(y_{nt};\beta_k^N)\right)=0, \tag{A.38}$$

$$\frac{1}{N}\sum_{n=1}^{N}\left(p_k\left(\mathbf{y}_n;\theta^N\right)\sum_{t=1}^{T-1}\frac{\partial}{\partial\tau}\ln c\left[F\left(y_{nt};\beta_k^N\right),F\left(y_{n,t+1};\beta_k^N\right);\tau_k^N\right]\right)=0, \tag{A.39}$$

$$\pi_k^N=\frac{1}{N}\sum_{n=1}^{N}p_k\left(\mathbf{y}_n;\theta^N\right). \tag{A.310}$$

Pseudo-true value. The observations are i.i.d.. By the Weak Law of Large Numbers, the sample averages in (A.38), (A.39), (A.310) therefore converge to their theoretical expectation analogs uniformly in the parameters. The set of parameters to which θ^N belongs is compact. One can thus extract from (θ^N) a sub-sequence which converges to some limit $\theta^\infty=(\boldsymbol{\beta}^\infty,\boldsymbol{\tau}^\infty,\boldsymbol{\pi}^\infty)$, solution to the limiting system of equations:

$$\mathbb{E}\left(p_k\left(\mathbf{y}_n;\theta^\infty\right)\sum_{t=1}^{T}\frac{\partial}{\partial\beta}\ln f(y_{nt};\beta_k^\infty)\right)=0, \tag{A.311}$$

$$\mathbb{E}\left(p_k\left(\mathbf{y}_n;\theta^\infty\right)\sum_{t=1}^{T-1}\frac{\partial}{\partial\tau}\ln c\left[F\left(y_{nt};\beta_k^\infty\right),F\left(y_{n,t+1};\beta_k^\infty\right);\tau_k^\infty\right]\right)=0, \tag{A.312}$$

$$\pi_k^\infty=\mathbb{E}[p_k\left(\mathbf{y}_n;\theta^\infty\right)]. \tag{A.313}$$

Denoting as θ_0 the true set of DGP parameters, we now prove the consistency of θ^N by proving that $\theta^\infty=\theta_0$. Root-N consistency and asymptotic normality will follow from standard arguments once consistency has been proven (Hansen, 1982). We shall thus not develop this point further.

Consistency. Because we shall not be concerned here by non parametric identification issues, we take $T = 2$ for all trajectories. Allowing for greater values of T only makes the estimation more efficient but does not affect consistency given the identifying assumptions we have made.

>From assumption (6), we deduce that the true value $\theta_0 = \left(\boldsymbol{\beta}^0, \boldsymbol{\tau}^0, \boldsymbol{\pi}^0\right)$ of the parameters is the unique maximizer of the incomplete log-likelihood:

$$\mathbb{E}\left(\ln \sum_{k=1}^{K} \pi_k f(Y_1, Y_2; \beta_k, \tau_k)\right) \tag{A.314}$$

subject to the constraint $\sum_{k=1}^{K} \pi_k = 1$.

The likelihood being continuously differentiable and the true value θ_0 belonging to the interior of the parameter space, the first-order conditions of the Lagrangian are necessary for $\boldsymbol{\beta}^0, \boldsymbol{\tau}^0, \boldsymbol{\pi}^0$ to be the optimum. Writing the condition for π_k^0 we obtain:

$$\begin{aligned} \pi_k^0 &= \mathbb{E}\left(\frac{\pi_k^0 f(Y_1, Y_2; \beta_k^0, \tau_k^0)}{\sum_{k=1}^{K} \pi_k^0 f(Y_1, Y_2; \beta_k^0, \tau_k^0)}\right) \\ &= \mathbb{E} p_k(Y_1, Y_2; \boldsymbol{\beta}^0, \boldsymbol{\tau}^0, \boldsymbol{\pi}^0) \end{aligned} \tag{A.315}$$

and differentiating (A.314) with respect to τ_k yields

$$\mathbb{E}\left(p_k(Y_1, Y_2; \boldsymbol{\beta}^0, \boldsymbol{\tau}^0, \boldsymbol{\pi}^0) \frac{\partial \ln f(Y_1, Y_2; \beta_k^0, \tau_k^0)}{\partial \tau_k}\right) = 0$$

$$\Updownarrow$$

$$\mathbb{E}\left(p_k(Y_1, Y_2; \boldsymbol{\beta}^0, \boldsymbol{\tau}^0, \boldsymbol{\pi}^0) \frac{\partial \ln c\left[F(Y_1; \beta_k^0), F\left(Y_2; \beta_k^0\right); \tau_k^0\right]}{\partial \tau_k}\right) = 0. \tag{A.316}$$

It follows that $\boldsymbol{\tau}^0$ and $\boldsymbol{\pi}^0$ satisfy conditions (A.312) and (A.313) if $\boldsymbol{\beta}^0 = \boldsymbol{\beta}^\infty$.

The first-order condition for β_k:

$$\mathbb{E}\left(p_k(Y_1, Y_2; \boldsymbol{\beta}^0, \boldsymbol{\tau}^0, \boldsymbol{\pi}^0) \frac{\partial \ln f(Y_1, Y_2; \beta_k^0, \tau_k^0)}{\partial \beta_k}\right) = 0$$

does not yield a similar equation as condition (A.311) as $c\left[F(Y_1; \beta_k^0), F\left(Y_2; \beta_k^0\right); \tau_k^0\right]$ also depends on β_k^0:

$$\frac{\partial \ln f(Y_1, Y_2; \beta_k^0, \tau_k^0)}{\partial \beta_k} \neq \frac{\partial \ln f(Y_1; \beta_k^0)}{\partial \beta_k} + \frac{\partial \ln f(Y_2; \beta_k^0)}{\partial \beta_k}.$$

But, for $t = 1, 2$:

$$\begin{aligned} \mathbb{E}\left(p_k(Y_1, Y_2; \boldsymbol{\beta}^0, \boldsymbol{\tau}^0, \boldsymbol{\pi}^0) \frac{\partial \ln f(Y_t; \beta_k^0)}{\partial \beta_k}\right) &= \int \pi_k^0 f(y_1, y_2; \beta_k^0, \tau_k^0) \frac{\partial \ln f(y_t; \beta_k^0)}{\partial \beta_k} dy_1 dy_2 \\ &= \pi_k^0 \int f(y_t; \beta_k^0) \frac{\partial \ln f(y_t; \beta_k^0)}{\partial \beta_k} dy_t \\ &= 0. \end{aligned}$$

Hence:

$$\mathbb{E}\left(p_k(Y_1,Y_2;\boldsymbol{\beta}^0,\boldsymbol{\tau}^0,\boldsymbol{\pi}^0)\left[\sum_{t=1}^{2}\frac{\partial\ln f(Y_t;\beta_k^0)}{\partial\beta_k}\right]\right)=0. \tag{A.317}$$

Moreover, by assumption 5, β_k^0 is the unique solution to (A.317) if $\pi_k^0\neq 0$.

We have thus shown that $\boldsymbol{\beta}^0,\boldsymbol{\tau}^0,\boldsymbol{\pi}^0$ are the unique solution to the limiting equations (A.311), (A.312) and (A.313). The sequential EM estimator is therefore consistent.

Note that maximum likelihood in the M-stage of the algorithm can be replaced by pseudo maximum likelihood, so as to make the algorithm even more user-friendly, without hampering the consistency of the estimator. It is indeed easy to see that the consistency proof works just as fine if the ML programs in the sequential EM algorithm are replaced by Pseudo-ML programs using linear-exponential distributions instead of the true distribution. For example, weighted OLS regressions can be used to update group-specific means.

A.4 Detailed specification and estimation procedure

We here consider the problem of estimating the empirical model over each three-year panel separately. The data consist of N independent observations of employment states $\mathbf{e}_n=(e_{nt},\ t=1,...,T)$, earnings $\mathbf{y}_n=(y_{nt},\ t=1,...,T)$ (with $y_{nt}=\cdot$ if $e_{nt}=0$), time-varying exogenous attributes $\mathbf{x}_n=(x_{nt},\ t=1,...,T)$ and observed heterogeneity z_{0n}, for $n=1,...,N$.

The parameters to be estimated are:

- $\alpha=(\alpha_1,...,\alpha_{K_1})$, the parameters of the probability distribution of latent variable z_1:

  $$\Pr\{z_1=k_1|z_0\}\equiv\pi_{k_1}(z_{0n};\alpha)=\frac{\exp z_{0n}'\alpha_{k_1}}{\sum_{\ell=1}^{K_1}\exp z_{0n}'\alpha_\ell},\quad k_1\in\{1,...,K_1\},$$

  with the normalization: $\alpha_1=0$.

- For every $k_1\in\{1...K_1\}$, $\beta_{k_1}=\left(\beta_{1|k_1},...,\beta_{K_2|k_1}\right)$, the parameters of the probability distribution of variable z_2, conditional on $z_1=k_1$:

  $$\Pr\{z_2=k_2|z_1=k_1,z_0\}\equiv\pi_{k_2|k_1}(z_{0n};\beta_{k_1})=\frac{\exp z_{0n}'\beta_{k_2|k_1}}{\sum_{\ell=1}^{K_2}\exp z_{0n}'\beta_{\ell|k_1}},\quad k_2\in\{1,...,K_2\},$$

  with the normalization: $\beta_{1|k_1}=0$.

- μ_{k_1},ω_{k_1}, $k_1\in\{1,...,K_1\}$ the parameters of cross-section earnings distributions,

  $$F_t\left(y_t|x_t,z_1=k_1\right)=\Phi\left(\frac{y_t-x_t'\mu_{k_1}}{\sqrt{x_t'\omega_{k_1}}}\right).$$

- δ_{k_1}, $k_1 \in \{1, ..., K_1\}$ the parameters of the unconditional unemployment probability,
$$\Pr\{e_t = 0 | x_t, z_1 = k_1\} = \Phi\left(x_t' \delta_{k_1}\right).$$

- $\chi_{ik_1k_2}$, $i \in \{0, 1\}$, $k_1 \in \{1, ..., K_1\}$, $k_2 \in \{1, ..., K_2\}$, the parameters of the conditional unemployment probability,
$$\Pr\{e_{t+1} = 0 | e_t = i, x_t, z_1 = k_1, z_2 = k_2\} = \Phi\left(x_t' \chi_{ik_1k_2}\right).$$

- $\eta_{k_1k_2}$, $k_1 \in \{1, ..., K_1\}$, $k_2 \in \{1, ..., K_2\}$, the parameters of the Plackett copula densities:
$$\begin{aligned} c(u, v | x_t, z_1 = k_1, z_2 = k_2) &\equiv c\left(u, v; \exp\left(x_t' \eta_{k_1k_2}\right)\right) \\ &= \left(\left[1 + (u+v)\exp\left(x_t'\eta_{k_1k_2}\right)\right]^2 - 4\exp\left(x_t'\eta_{k_1k_2}\right)\left[1 + \exp\left(x_t'\eta_{k_1k_2}\right)\right] uv\right)^{-\frac{3}{2}} \\ &\quad \times \left[1 + \exp\left(x_t'\eta_{k_1k_2}\right)\right]\left[1 + (u + v - 2uv)\exp\left(x_t'\eta_{k_1k_2}\right)\right]. \end{aligned}$$

Note that we restrict the parameter of the Plackett copula, i.e. $\exp\left(x_t'\eta_{k_1k_2}\right)$, to be higher than zero, to exclude mean reversion.[2]

Let θ gather all parameters in one single vector. If all heterogeneity variables were observed one would maximize the following conditional likelihood (omitting the individual index n):

$$\begin{aligned} f(\mathbf{e}, \mathbf{y} | \mathbf{x}, z_0, z_1, z_2; \theta) \quad = \quad & \Pr\{e_1 = 0 | x_1, z_1\}^{1-e_1}\left[1 - \Pr\{e_1 = 0 | x_1, z_1\}\right]^{e_1} \\ & \times \prod_{t=1}^{T} f_t\left(y_t | x_t, z_1\right)^{e_t} \\ & \times \prod_{t=1}^{T-1} \Pr\{e_{t+1} = 0 | e_t, x_t, z_1, z_2\}^{1-e_{t+1}}\left[1 - \Pr\{e_{t+1} = 0 | e_t, x_t, z_1, z_2\}\right]^{e_{t+1}} \\ & \times \prod_{t=1}^{T-1} \left[c\left(F_t\left(y_t | x_t, z_1\right), F_{t+1}\left(y_{t+1} | x_{t+1}, z_1\right) | x_t, z_1, z_2\right)\right]^{e_t e_{t+1}}. \end{aligned}$$

The estimation algorithm updates the step-s value $\theta^{(s)}$ of the parameter vector θ by going through the following steps:

1. For all $k_1 \in \{1, ..., K_1\}$, compute the posterior probability of $z_{1n} = k_1$ given $\mathbf{y}_n, \mathbf{x}_n, z_{0n}$ and $\theta^{(s)}$ as
$$p_{k_1}^{z_1}\left(\mathbf{y}_n, \mathbf{x}_n, z_{0n}; \theta^{(s)}\right) = \sum_{\ell=1}^{K_2} p_{k_1,\ell}^{z_1,z_2}\left(\mathbf{y}_n, \mathbf{x}_n, z_{0n}; \theta^{(s)}\right),$$

[2] On the issue of negative dependence, see Shorrocks (1978).

where

$$p^{z_1,z_2}_{k_1,k_2}\left(\mathbf{y}_n,\mathbf{x}_n,z_{0n};\theta^{(s)}\right)$$
$$= \frac{\pi^{(s)}_{k_2|k_1}(z_{0n};\beta^{(s)}_{k_1})\pi^{(s)}_{k_1}(z_{0n};\alpha^{(s)})f(\mathbf{e}_n,\mathbf{y}_n|\mathbf{x}_n,z_{0n},z_{1n}=k_1,z_{2n}=k_2;\theta^{(s)})}{\sum_{k=1}^{K_1}\sum_{\ell=1}^{K_2}\pi^{(s)}_{\ell|k}(z_{0n};\beta^{(s)}_{k})\pi^{(s)}_{k}(z_{0n};\alpha^{(s)})f(\mathbf{e}_n,\mathbf{y}_n|\mathbf{x}_n,z_{0n},z_{1n}=k,z_{2n}=\ell;\theta^{(s)})}$$

is the posterior probability of $z_{1n}=k_1$ and $z_{2n}=k_2$ given $\mathbf{y}_n,\mathbf{x}_n,z_{0n}$ and $\theta^{(s)}$.

2. Update parameters μ_{k_1}, $k_1\in\{1,...,K_1\}$, by regressing y_{nt} on x_{nt} using OLS and weighting each observation by $p^{z_1}_{k_1}\left(\mathbf{y}_n,\mathbf{x}_n,z_{0n};\theta^{(s)}\right)$. Then, update ω_{k_1}, by regressing squared residuals on x_{nt}, again using weighted least squares.

3. Update δ_{k_1}, $k_1\in\{1,...,K_1\}$, as in the preceding step: estimate the PROBIT model of whether $i_1=0$ or not conditional on x_1, weighting each observation by $p^{z_1}_{k_1}\left(\mathbf{y}_n,\mathbf{x}_n,z_{0n};\theta^{(s)}\right)$.

4. Update the parameters $\chi_{ik_1k_2}$, $i\in\{0,1\}$, $k_1\in\{1,...,K_1\}$, $k_2\in\{1,...,K_2\}$, of the conditional unemployment transition probabilities, $\Pr\{e_{t+1}=0|e_t=i,x_t,z_1=k_1,z_2=k_2\}=\Phi\left(x_t'\chi_{ik_1k_2}\right)$, by running two PROBIT estimations weighting each observation by $p^{z_1,z_2}_{k_1,k_2}(\mathbf{y}_n,\mathbf{x}_n,z_{0n};\theta^{(s)})$.

5. Update $\eta_{k_1k_2}$ as:

$$\eta^{(s+1)}_{k_1k_2}=\arg\max_{\eta_{k_1k_2}}\left\{\sum_{n=1}^{N}p^{z_1,z_2}_{k_1,k_2}\left(\mathbf{y}_n,\mathbf{x}_n,z_{0n};\theta^{(s)}\right)\sum_{\substack{t=1\\ e_t\neq 0,\\ e_{t+1}\neq 0}}^{T-1}\ln c\left[F_t\left(y_t|x_t,z_1=k_1;\mu^{(s+1)}_{k_1},\omega^{(s+1)}_{k_1}\right),\right.\right.$$
$$\left.\left.F_{t+1}\left(y_{t+1}|x_{t+1},z_1=k_1;\mu^{(s+1)}_{k_1},\omega^{(s+1)}_{k_1}\right);\exp(x_t'\eta_{k_1k_2})\right]\right\}.$$

Due to the mathematical simplicity of the Plackett copula, this maximization is easy to perform.

6. Update α by solving the program

$$\alpha^{(s+1)}=\arg\max_{\alpha}\sum_{n=1}^{N}\sum_{k=1}^{K_1}p^{z_1}_{k}\left(\mathbf{y}_n,\mathbf{x}_n,z_{0n};\theta^{(s)}\right)\ln\pi_k(z_{0n};\alpha)$$
$$\text{s.t.}\quad\sum_{k=1}^{K_1}\pi_k(z_{0n};\alpha)=1,$$

In practice, we did not solve this ML program. To benefit from standard algorithms available in most statistical softwares (as STATA), we instead simulated for each n a set of D draws $z^{(1)}_{1n},...,z^{(D)}_{1n}$ of the latent group indicator z_{1n} from the multinomial distribution

$$\mathcal{M}\left(p^{z_1}_1\left(\mathbf{y}_n,\mathbf{x}_n,z_{0n};\theta^{(s)}\right),...,p^{z_1}_{K_1}\left(\mathbf{y}_n,\mathbf{x}_n,z_{0n};\theta^{(s)}\right)\right)$$

and computed the standard ML estimate:

$$\alpha^{(s+1)} = \arg\max_{\alpha} \sum_{n=1}^{N} \sum_{d=1}^{D} \sum_{k=1}^{K} \mathbf{1}\left(z_{1n}^{(d)} = k\right) \ln \pi_k^{z_1}(z_{0n}; \alpha).$$

Various experiments showed that $D = 1$ already gave a good approximation of the exact ML estimator.

7. Update β_{k_1}, for all k_1, by solving the program

$$\beta_{k_1}^{(s+1)} = \arg\max_{\beta_{k_1}} \sum_{n=1}^{N} \sum_{\ell=1}^{K_2} p_{k_1,\ell}^{z_1,z_2}\left(\mathbf{y}_n, \mathbf{x}_n, z_{0n}; \theta^{(s)}\right) \ln \pi_{\ell|k_1}(z_{0n}; \beta_{k_1})$$
$$\text{s.t.} \quad \sum_{\ell=1}^{K_2} \pi_{\ell|k_1}(z_{0n}; \beta_{k_1}) = 1,$$

by using the same simulation method.

A.5 Simulation

In this section of the appendix, we present our method to simulate earnings trajectories from date t_0 to date t_1. We proceed in six steps. For every individual in the sample:

1. Predict a group z_1 and a group z_2, based on the individual's permanent characteristics z_0 (education and cohort).

2. Predict an employment status at t_0, conditional on z_1 and x_{t_0}. For all $t_0 < t \leq t_1$, predict the sequence of employment status, based on z_1, z_2 and x_t.

3. Compute a wage by drawing from a log-normal distribution, conditional on z_1 and x_{t_0}. For unemployed individual, assign a virtual wage equal to the former value times a replacement ratio $0 < \lambda < 1$. If the individual is employed, then compute his rank at t_0 (conditional on employment).

4. Proceed recursively. For all $t_0 < t \leq t_1$ compute, if the individual is employed at $t-1$ and t, a rank at t by drawing from distribution $C\left(\cdot, \cdot; \exp\left(x_t' \eta_{k_1 k_2}\right)\right)$, the Plackett copula with parameter $\exp\left(x_t' \eta_{k_1 k_2}\right)$. To achieve this task, note that for a given $u \in [0,1]$, $\frac{\partial C}{\partial u}(u, v)$ is uniform(0,1). Therefore, for all u, draw a uniform(0,1) variable w and solve $\frac{\partial C}{\partial u}(u, v) = w$ with respect to v. For the Plackett copula, this implicit equation writes as a quadratic polynomial equation with two roots in $[0,1]$. Choose the highest or the lowest root with probability $\frac{1}{2}$. The result is the individual's rank at t, r_t. If the individual is employed at t but was unemployed at $t-1$ draw r_t in a uniform(0,1) distribution.

5. For all $t_0 < t \leq t_1$: if the individual is employed at t, draw a wage by computing $y_t = F_t^{-1}(r_t)$, where F_t is the cross-section cdf, conditional on z_1 and x_t.

6. For all $t_0 < t \leq t_1$: if the individual is unemployed at t and was employed at $t-1$, assign a virtual wage equal to the replacement ratio times his previous wage. If he was also unemployed at $t-1$ keep his virtual wage equal to its past value.

We then compute weighted means of the simulated log-earnings sequences, weighting time-t earnings, $t \in \{t_0, t_1\}$, by $\left(\frac{1}{1+r}\right)^{t-t_0}$.

Various parameters are critical in the simulation, and especially the replacement ratio λ. We take $\lambda = .60$ (Martin, 1996). Varying λ between .40 and .80 had a strong effect on the level of one-to-fifteen-year inequalities, but little influence on the evolution of inequality indices. Note that we neglect several important features, such as heterogeneity in the replacement ratio, nonstationary UI schemes and specific disutility of unemployment. Lastly, we take $r = .10$. Varying r between 0 and .20 had little effect on the results.

A.6 Estimation of conditional Spearman rho

Let $C(\cdot,\cdot|z)$ denote the copula of (y_t, y_{t+1}) conditional on some variable $Z = z$. $C(\cdot,\cdot|z)$ is the cdf of the couple of r.v.'s $(F(y_{nt}|z), F(y_{n,t+1}|z))$. To estimate $\rho\left(C(\cdot,\cdot|z)\right)$ from a sample $(\mathbf{y}_n)$ of i.i.d. observations, we use Hoeffding's formula:

$$\rho\left(C(\cdot,\cdot|z)\right) = 12\mathbb{E}\left[F(y_{nt}|z)F(y_{n,t+1}|z)|Z=z\right] - 3,$$

and we replace the conditional expectation by the sample mean of $F(y_{nt}|z)F(y_{n,t+1}|z)$. If z is unobserved heterogeneity, then we draw a value of z for each individual in the sample using the posterior probability that $Z_n = z$ conditional on the individual observation data. These posterior probabilities are obtained as a by-product of the estimation of the empirical model by the EM algorithm.

This procedure yields consistent estimates for the following reason. Let $\mathbf{y}_n$ denote the complete set of observations for individual n and let the posterior probability be $p_z(\mathbf{y}_n) = \frac{f(\mathbf{y}_n|z)P(z)}{f(\mathbf{y}_n)}$. Then:

$$\begin{aligned} plim_{N\to\infty}\frac{\sum_{n=1}^N F_t(y_{nt}|z)F_{t+1}(y_{n,t+1}|z)p_z(\mathbf{y}_n)}{\sum_{n=1}^N p_z(\mathbf{y}_n)} &= \frac{\mathbb{E}\left[F_t(Y_{nt}|Z)F_{t+1}(Y_{n,t+1}|Z)p_z(\mathbf{Y}_n)\right]}{\mathbb{E}\left[p_z(\mathbf{Y}_n)\right]} \\ &= \frac{\int F_t(y_{nt}|z)F_{t+1}(y_{n,t+1}|z)p_z(\mathbf{y}_n)f(\mathbf{y}_n)d\mathbf{y}_n}{\int p_z(\mathbf{y}_n)f(\mathbf{y}_n)d\mathbf{y}_n} \\ &= \mathbb{E}\left[F_t(Y_{nt}|z)F_{t+1}(Y_{n,t+1}|z)|Z=z\right], \end{aligned}$$

as $P(z) = \mathbb{E}\left[p_z(\mathbf{Y}_n)\right] = \int p_z(\mathbf{y}_n)f(\mathbf{y}_n)d\mathbf{y}_n$.

A.7 Parameter estimates

		Cross-section heterogeneity	
		$z_1 = 1$	$z_2 = 2$
Earnings			
	experience	.034 (.00091)	.030 (.00046)
Mean	squared experience	−.00068 (.000020)	−.00053 (.0000097)
	intercept	8.99 (.16)	8.43 (.0045)
	experience	−.0041 (.00060)	.00095 (.00026)
Variance	squared experience	.00011 (.000013)	−.000015 (.0000053)
	intercept	.16 (.063)	.058 (.0027)
Unemployment probability			
experience		−.042 (.0060)	−.084 (.0032)
squared experience		.0012 (.00012)	.0017 (.000069)
intercept		−1.77 (.066)	−.046 (.029)

Table F1: Cross-sectional earnings distribution - Parameter estimates

Independent variable	Estimate
Intercept	90 (8.4)
Education	
Junior high-school	72 (9.4)
Senior high-school	116 (11.1)
Some college	76 (14.6)
College +	−39 (34.0)
Cohort (year of entry into the labour market)	−.048 (.0043)
Cohort*Education	
Junior high-school	−.036 (.0048)
Senior high-school	−.057 (.0056)
Some college	−.035 (.0074)
College +	.024 (.017)
Pseudo-R^2	.53

Table F2: Distribution of cross-section heterogeneity given exogenous controls

	Mobility heterogeneity	
	$z_2 = 1$	$z_2 = 2$
$z_1 = 1$		
experience	−.015 (.0025)	−.051 (.0056)
intercept	6.44 (.14)	5.58 (.16)
$z_1 = 2$		
experience	.049 (.0050)	.048 (.0065)
intercept	3.43 (.092)	.87 (.32)

Table F3: Parameter η of the earnings copula (greater if less mobility) conditional on experience and unobserved heterogeneity)

	empl. to unempl.		unempl. to unempl.	
	$z_2 = 1$	$z_2 = 2$	$z_2 = 1$	$z_2 = 2$
$z_1 = 1$				
experience	.043 (.0044)	−.00017 (.0043)	.178 (.011)	.079 (.54)
intercept	−3.82 (.13)	−1.89 (.11)	−1.50 (.27)	−2.71 (23.4)
$z_1 = 2$				
experience	−.084 (.0039)	−.0070 (.0018)	.033 (.0026)	.060 (.028)
intercept	−.82 (.057)	−1.34 (.048)	.20 (.059)	−1.66 (.070)

Table F4: Conditional employment-unemployment transition probabilities

Independent variable	Cross-section heterogeneity	
	$z_1 = 1$	$z_1 = 2$
Intercept	27 (16.7)	−91 (3.4)
Education		
Junior high-school	38 (18.5)	22 (4.4)
Senior high-school	50 (17.4)	65 (8.2)
Some college	38 (18.8)	71 (40.7)
College +	35 (19.8)	−14 (95.4)
Cohort	−.014 (.0085)	.046 (.0018)
Cohort*Education		
Junior high-school	−.019 (.0094)	−.011 (.0022)
Senior high-school	−.025 (.0089)	−.033 (.0042)
Some college	−.019 (.0096)	−.036 (.021)
College +	−.018 (.10)	.0062 (.0048)
Pseudo-R^2	.025	.036

Table F5: Distribution of cross-section heterogeneity given exogenous controls

Appendix B

Appendix of Chapter 2

B.1 Data

The definition of jobs: We let individuals be in either one of the two following labor market states: employed or unemployed. Unemployment comprises self declared unemployment, inactivity, employment during less than 15 hours per week or with wages lower than the first percentile (which, for example, is around 235 Euros per month in France, that is 25% of the median wage). We drop every individual who experiences a self employment spell since we assume her trajectory (and especially her job mobility decisions) not to be governed by the same processes as those of workers in paid employment.

Attrition: Some of the observation periods are right censored, *i.e.* individuals do not always stay in the ECHP during the eight waves. We assume this right censoring to be exogenous to the wage, amenity and job mobility process.

Missing data: The problem of missing data is twofold: there can be non reported variables for a given wave where the individual is present or the individual can "disappear" from the survey during a year within his observation period and come back one year later. When it is possible, we impute missing data on wages and/ or amenity using the previous or following wave if the individual is still in the same job: we substitute the missing wage for the mean of the previous and following wage and draw the amenity from a binomial distribution weighting both the previous and following amenity with probability 0.5 (the amenity can change within a job). These substitutions affect less than a thousand observations (over *e.g.* more than 30 000 in France). For the few observations that still show missing data, we create two individuals out of one. This rather arbitrary treatment of less than 1% of our sample does not affect the consistency of the estimates and the loss of efficiency is likely to be small.

B.2 The estimation procedure

In this section of the Appendix, we detail the estimation procedure of the model presented in 2.3.1. We start by setting the notations. We let $i \in \{1...N\}$ denote individuals, and $t \in \{1...T\}$ denote time periods. Let e_{it} be the dummy variable indicating if individual i is employed at time t. The model allows for multivariate amenities $\boldsymbol{a}_{it} \in \{0,1\}^J$. In the empirical analysis, we take $J = 5$. As outlined in 2.3.1, the dynamics of employment, wages and amenities follow:

$$\begin{aligned}
(e_{it+1}, y_{it+1}, \mathbf{a}_{it+1} | e_{it} = 1) &= (0,0,0) \quad \text{if } z_{it}^e = 1, \\
&= (1, y_{it}^c, \mathbf{a}_{it}^c) \quad \text{if } z_{it}^e = 0 \text{ and } z_{it}^c = 1, \\
&= (1, y_{it}^*, \mathbf{a}_{it}^*) \quad \text{if } z_{it}^e = 0,\ z_{it}^c = 0 \text{ and } z_{it} = 1, \\
&= (1, y_{it}^r, \mathbf{a}_{it}^r) \quad \text{if } z_{it}^e = 0,\ z_{it}^c = 0 \text{ and } z_{it} = 0, \\
(e_{it+1}, y_{it+1}, \mathbf{a}_{it+1} | e_{it} = 0) &= (1, y_{it}^n, \mathbf{a}_{it}^n) \quad \text{if } z_{it}^n = 1, \\
&= (0,0,0) \quad \text{if } z_{it}^n = 0,
\end{aligned}$$

where z_{it} indicates voluntary mobility, and $(^c)$, $(^n)$ and $(^e)$ superscripts refer to constrained job-to-job mobility, non employment-to-job and job-to-non employment transitions, respectively. Moreover, to avoid confusions we have used superscript $(^r)$ to index within-job wage and amenity distributions, which appear in equations (2.1)-(2.2) in 2.3.1.

All equations in section 2.3.1 are easily adapted to the case of multivariate amenities. For instance, wage and amenity offers (2.3)-(2.4) become:

$$\begin{aligned}
a_{j,it+1}^* &= \mathbf{1}\left\{\alpha_{j,a}^* x_{it} + \beta_{j,1a}^* \theta_{1it} + \beta_{j,2a}^* \theta_{2it} + u_{j,ait+1}^* > 0\right\} \quad \forall j \in \{1..J\}, \\
y_{it+1}^* &= \boldsymbol{\rho}^* \mathbf{a}_{it+1}^* + \alpha_y^* x_{it} + \beta_{j,y}^* \theta_{1it} + u_{yit+1}^*.
\end{aligned} \tag{B.21}$$

In (B.21), $\boldsymbol{\rho}^* = (\rho_1^*, ..., \rho_J^*)$ is the vector of compensating differentials in job offers associated to the J amenities.

B.2.1 The EM algorithm

Consider an individual i, and a given job which lasts from t_{i0} to $t_{i1} - 1$. Unobserved match characteristics are assumed constant on $[t_{i0} + 1, t_{i1}]$. Moreover, they are realized after the individual has started to work in the new job. It is thus convenient to estimate the incomplete likelihood (Dempster *et al.*, 1977) of the individual observation between $t_{i0} + 1$ and t_{i1}, *conditional* on wage/amenity realizations at t_{i0}:

$$\sum_{\theta_1, \theta_2} \pi_{\theta_1, \theta_2}\left((y, \mathbf{a})_{it_{i0}}; \Theta_1\right) \prod_{t=t_{i0}}^{t_{i1}-1} f((y, \mathbf{a})_{it+1}, e_{it+1}, z_{it}^c, z_{it} | e_{it}, \theta_1, \theta_2, x_{it}, x_{it+1}; \Theta_2).$$

In this expression, Θ_1 and Θ_2 are sets of parameters. In the rest of this section we shall denote as $\boldsymbol{\Theta} = (\Theta_1, \Theta_2)$ the set of parameters with respect to which the incomplete likelihood is to be maximized. Then, π_{k_1,k_2} are the prior probabilities $\mathbb{P}(\theta_{1it} = k_1, \theta_{2it} = k_2 | (y, \mathbf{a})_{it_{i0}})$, conditional on the wage and amenities at t_{i0}. It is implicitly assumed that both θ_1 and θ_2 are independent of $x_{t_{i0}}$, conditional on the wage and amenities at t_{i0}.

We model θ_1 and θ_2 as two independent random variables (conditional on $(y, \mathbf{a})_{it_{i0}}$) following the Ordered PROBIT specification. Precisely, we assume that there exist two latent variables:

$$\widetilde{\theta}_{1i} = \alpha_{1y}^{init} y_{it_{i0}} + \beta_{1y}^{init} \boldsymbol{a}_{it_{i0}} + u_{1it_{i0}}^{init}, \tag{B.22}$$

$$\widetilde{\theta}_{2i} = \beta_{2y}^{init} \boldsymbol{a}_{it_{i0}} + u_{2it_{i0}}^{init}, \tag{B.23}$$

and $K_1 + K_2 - 2$ thresholds $s_{1,1}, ..., s_{1,K_1-1}, s_{2,1}, ..., s_{2,K_2-1}$, such that:

$$\theta_{1i} = 1 \text{ if } \widetilde{\theta}_{1i} < s_{1,1}, \quad \theta_{1i} = 2 \text{ if } s_{1,1} \leq \widetilde{\theta}_{1i} < s_{1,2}, \quad \theta_{1i} = K_1 \text{ if } s_{1,K_1-1} \leq \widetilde{\theta}_{1i},$$

and:

$$\theta_{2i} = 1 \text{ if } \widetilde{\theta}_{2i} < s_{2,1}, \quad \theta_{2i} = 2 \text{ if } s_{2,1} \leq \widetilde{\theta}_{2i} < s_{2,2}, \quad \theta_{2i} = K_2 \text{ if } s_{2,K_2-1} \leq \widetilde{\theta}_{2i}.$$

Residuals $u_{1it_{i0}}^{init}$ and $u_{2it_{i0}}^{init}$ are independent of each other and covariates, and follow standard normal distributions. We allow the parameters in (B.22)-(B.23) to be different, in the case where t_{i0} corresponds to the first date of observation for individual i (initial conditions).

Let us partition Θ_2 into subsets corresponding to different transitions. For instance, Θ_2^{yr} corresponds to the parameters in the hedonic equation (2.2) of 2.3.1, including the standard deviation of u. Similarly, Θ_2^{z} is defined as the set of parameters ruling voluntary mobility decisions, see equation (2.6) in the same section.

Then we can factorize $f((y,\mathbf{a})_{it+1}, e_{it+1}, z^c_{it}, z_{it}|e_{it}, \theta_1, \theta_2, x_{it}, x_{it+1}; \Theta_2)$ into:

$$
\begin{aligned}
& f\left(e_{it+1}|e_{it}=1,\theta_1,\theta_2,x_{it};\Theta_2^e\right)^{e_{it}=1} \cdot f\left(e_{it+1}|e_{it}=0,\theta_1,\theta_2,x_{it};\Theta_2^n\right)^{e_{it}=0} \\
\cdot & \left[f\left(y_{it+1}|\mathbf{a}_{it+1}, z^n_{it}=1,\theta_1,x_{it};\Theta_2^{yn}\right) \cdot \prod_{j=1}^J f\left(a_{j,it+1}|z^n_{it}=1,\theta_1,\theta_2,x_{it};\Theta_2^{jan}\right) \right]^{z^n_{it}=1} \\
\cdot & f\left(z^c_{it}|z^n_{it}=0,e_{it}=1,\theta_1,\theta_2,x_{it};\Theta_2^c\right)^{z^n_{it}=0} \\
\cdot & \Bigg[f\left(y_{it+1}|\mathbf{a}_{it+1}, z^c_{it}=1, z^n_{it}=0, e_{it}=1,\theta_1,x_{it};\Theta_2^{yc}\right) \\
& \cdot \prod_{j=1}^J f\left(a_{j,it+1}|z^c_{it}=1, z^n_{it}=0, e_{it}=1,\theta_1,\theta_2,x_{it};\Theta_2^{jac}\right) \Bigg]^{z^n_{it}=0, z^c_{it}=1} \\
\cdot & f\left(z_{it}=1,(y,\mathbf{a})_{it+1}|z^c_{it}=z^n_{it}=0,e_{it}=1,\theta_1,\theta_2,x_{it};\Theta_2^z,\Theta_2^{y*},\left\{\Theta_2^{ja*}\right\}_j\right)^{z^n_{it}=z^c_{it}=0,z_{it}=1} \\
\cdot & f\left(z_{it}=0|z^c_{it}=z^n_{it}=0,e_{it}=1,\theta_1,\theta_2,x_{it};\Theta_2^z,\Theta_2^{y*},\left\{\Theta_2^{ja*}\right\}_j\right)^{z^n_{it}=z^c_{it}=z_{it}=0} \\
\cdot & \Bigg[f\left(y_{it+1}|\mathbf{a}_{it+1}, z_{it}=z^c_{it}=z^n_{it}=0, e_{it}=1,\theta_1,x_{it+1};\Theta_2^{yr}\right) \\
& \cdot \prod_{j=1}^J f\left(a_{j,it+1}|z_{it}=z^c_{it}=z^n_{it}=0,e_{it}=1,\theta_1,\theta_2,x_{it+1};\Theta_2^{jar}\right) \Bigg]^{z^n_{it}=z^c_{it}=z_{it}=0} .
\end{aligned}
$$

We can thus rewrite:

$$
\begin{aligned}
& f\left((y,\mathbf{a})_{it+1}, e_{it+1}, z^c_{it}, z_{it}|e_{it},\theta_1,\theta_2,x_{it},x_{it+1};\Theta_2\right) \\
= \; & g\left((y,\mathbf{a})_{it+1}, e_{it+1}, z^c_{it}, z_{it}|e_{it},\theta_1,\theta_2,x_{it},x_{it+1};\Theta_2^{-z}\right) \\
& \cdot \Big[f\left(z_{it}=0|z^c_{it}=z^n_{it}=0,e_{it}=1,\theta_1,\theta_2,x_{it};\Theta_2^z,\Theta_2^{-z}\right)^{z^n_{it}=z^c_{it}=z_{it}=0} \\
& \cdot f\left(z_{it}=1,(y,\mathbf{a})_{it+1}|z^c_{it}=z^n_{it}=0,e_{it}=1,\theta_1,\theta_2,x_{it};\Theta_2^z,\Theta_2^{-z}\right)^{z^n_{it}=z^c_{it}=0,z_{it}=1} \Big] \\
= \; & g\left((y,\mathbf{a})_{it+1}, e_{it+1}, z^c_{it}, z_{it}|e_{it},\theta_1,\theta_2,x_{it},x_{it+1};\Theta_2^{-z}\right) \\
& \cdot h\left(z_{it},(y,\mathbf{a})_{it+1}|z^c_{it}=z^n_{it}=0,e_{it}=1,\theta_1,\theta_2,x_{it};\Theta_2^z,\Theta_2^{-z}\right)
\end{aligned}
$$

where Θ_2^{-z} is the subset of Θ_2 containing all the parameters in Θ_2 but those in Θ_2^z, and g is a product of conditional likelihoods.

The joint maximization of the likelihood being cumbersome, we take advantage of this factorization to estimate the model's parameters sequentially. We do so in each M-step of the EM algorithm, as we now explain. Given initial values for the parameters, $\Theta_1^{(s)}$ and $\Theta_2^{(s)}$, the two steps of the sequential EM write as follows.

E-Step: Compute the posterior probabilities of (θ_1, θ_2) given the data

$$X_{i.} = \{(y, \boldsymbol{a})_{it+1}, e_{it+1}, z_{it}^c, z_{it}\}_{t_{i0}+1 \le t \le t_{i1}},$$

and $X_{i0.} = (y, \boldsymbol{a})_{it_{i0}}$, and conditional on $x_{i.} = \{x_{it}\}_{t_{i0}+1 \le t \le t_{i1}}$, as:

$$p_{\theta_1,\theta_2}(X_{i.}|X_{i0.}; \boldsymbol{\Theta}^{(s)}) = \frac{\pi_{\theta_1,\theta_2}\left(X_{i0.}; \Theta_1^{(s)}\right) f\left(X_{i.}|\theta_1, \theta_2, x_{i.}; \Theta_2^{(s)}\right)}{\sum_{k_1,k_2} \pi_{k_1,k_2}\left(X_{i0.}; \Theta_1^{(s)}\right) f\left(X_{i.}|k_1, k_2, x_{i.}; \Theta_2^{(s)}\right)}.$$

M-step: Update the parameters as follows:

$$\Theta_1^{(s+1)} = \underset{\Theta_1}{Argmax} \sum_i \sum_{\theta_1,\theta_2} p_{\theta_1,\theta_2}(X_{i.}|X_{i0.}; \boldsymbol{\Theta}^{(s)}) \ln \pi_{\theta_1,\theta_2}(X_{i0.}; \Theta_1), \tag{B.24}$$

$$\begin{aligned}\Theta_2^{(s+1),-z} &= \underset{\Theta_2^{-z}}{Argmax} \sum_{i,J(i)} \sum_{\theta_1,\theta_2} \Big[p_{\theta_1,\theta_2}(X_{i.}|X_{i0.}; \boldsymbol{\Theta}^{(s)}) \quad \ldots \\ &\ldots \sum_{t=t_{i0}}^{t_{i1}-1} \ln g\left((y, \mathbf{a})_{it+1}, e_{it+1}, z_{it}^c, z_{it}|e_{it}, \theta_1, \theta_2, x_{it}, x_{it+1}; \Theta_2^{-z}\right)\Big],\end{aligned} \tag{B.25}$$

$$\begin{aligned}\Theta_2^{(s+1),z} &= \underset{\Theta_2^{z}}{Argmax} \sum_{i,J(i)} \sum_{\theta_1,\theta_2} \Big[p_{\theta_1,\theta_2}(X_{i.}|X_{i0.}; \boldsymbol{\Theta}^{(s)}) \quad \ldots \\ &\ldots \sum_{t=t_{i0}}^{t_{i1}-1} \ln h\left(z_{it}, (y, \mathbf{a})_{it+1}|z_{it}^c = z_{it}^n = 0, e_{it} = 1, \theta_1, \theta_2, x_{it}; \Theta_2^z, \Theta_2^{(s+1),-z}\right)\Big],\end{aligned} \tag{B.26}$$

where $J(i)$ indexes the jobs held by individual i.

The M-step of this algorithm differs from the maximization of the complete likelihood as maximization with respect to Θ_2 is achieved sequentially, maximizing the part of the complete likelihood corresponding to voluntary mobility given previous estimates of the offer parameters. This sequential EM algorithm is considered by Arcidiacono and Jones (2003), who prove that it yields consistent estimates of the parameters. They also show that, as this estimator differs from ML, it is not asymptotically efficient. It is to be noted that Arcidiacono and Jones (2003) focus on *unconditional* EM. The extension of their analysis to the case where prior probabilities are conditional on some covariates is however straightforward.

The calculations in the M-step are as follows. In (B.24), we estimate the two components of Θ_1 by ordered PROBIT. The dependent variable is the indicator of the sample number, where the modified dataset is the original one duplicated $K_1 K_2$ times. The maximization is weighted by the groups' posterior probabilities. In (B.25), we estimate all sets of parameters corresponding to wage equations, like Θ_2^{yc} for instance, by OLS. We estimate parameters corresponding to amenity equations, like Θ_2^{1ac}, by PROBIT. Lastly, we estimate the set of parameters corresponding to shocks or decisions, like Θ_2^c, also by PROBIT. All regressions are weighted by the posterior probabilities.

To update the parameters ruling voluntary mobility (Θ_2^z), we maximize a likelihood that we present in the next section. The maximization of this likelihood being somewhat longer than the rest of the E and M steps, we proceeded in two stages. In a first stage, we estimated the global model assuming no selection effects in the voluntary mobility process, *i.e.* we replaced (2.6) with:

$$z_t = \mathbf{1}\left\{\tilde{\alpha}_z x + \tilde{\beta}_{1z}\theta_1 + \tilde{\beta}_{2z}\theta_2 + \varepsilon_z > 0\right\},$$

where ε_z is normally distributed, *i.i.d.*, and independent of covariates. We then computed the posterior probabilities for every individual in the sample, and maximized the likelihood corresponding to the voluntary mobility rule, weighted by these probabilities. We finally runned a second sequential EM, allowing for selection in voluntary mobility, and taking the latter estimates as initial conditions for the maximization of the corresponding likelihood. Our experiments showed that the number of iterations necessary for EM to converge numerically was much reduced by proceeding this way.

B.2.2 Voluntary mobility

The likelihood corresponding to the voluntary mobility rule, for one transition $t/t+1$, conditional on $(x_{it}, z_{it}^c = z_{it}^n = 0, e_{it} = 1, \theta_{1it} = \theta_1, \theta_{2it} = \theta_2)$ writes:

$$\begin{aligned}\mathcal{L} &= \prod_{i,z_{it}=1} f(z_{it}=1,(y,\mathbf{a})_{it+1};\Theta_2^z,\Theta_2^{-z}) \times \prod_{i,z_{it}=0} f(z_{it}=0;\Theta_2^z,\Theta_2^{-z}),\\ &= \prod_{i,z_{it}=1} f((y,\mathbf{a})_{it+1};\Theta_2^{-z}) \times \prod_{i,z_{it}=1} f(z_{it}=1|(y,\mathbf{a})_{it+1};\Theta_2^z,\Theta_2^{-z}) \times \prod_{i,z_{it}=0} f(z_{it}=0;\Theta_2^z,\Theta_2^{-z}).\end{aligned} \tag{B.27}$$

Note that the first term in (B.27) depends on Θ_2^{-z} only, namely on parameters α_y^*, ρ^*... As the maximization in (B.26) is with respect to Θ_2^z only, it is equivalent to maximize only the two last terms in (B.27).

From equation (2.6) the latter read:

$$\begin{aligned}
\widetilde{\mathcal{L}} &= \prod_{i,z_{it}=1} f(z_{it}=1|(y,\mathbf{a})_{it+1};\Theta_2^z,\Theta_2^{-z}) \times \prod_{i,z_{it}=0} f(z_{it}=0;\Theta_2^z,\Theta_2^{-z}),\\
&= \prod_{i,z_{it}=1} \Phi\left(\frac{y_{it+1}+\boldsymbol{\delta}^*\mathbf{a}_{it+1}-\alpha_z x_{it}-\beta_{1z}\theta_1-\beta_{2z}\theta_2}{\sigma}\right)\\
&\quad \cdot \prod_{i,z_{it}=0}\Bigg[\sum_{\mathbf{b}\in\{0,1\}^J}\left(\prod_{j=1}^{J}\Phi\left((-1)^{b_j}(\alpha_{j,a}^* x_{it}+\beta_{j,1a}^*\theta_1+\beta_{j,2a}^*\theta_2)\right)\right)\\
&\qquad \cdot \ \Phi\left(\frac{\alpha_y^* x_{it}+\beta_y^*\theta_1+(\boldsymbol{\rho}^*+\boldsymbol{\delta}^*)\,\boldsymbol{b}-\alpha_z x_{it}-\beta_{1z}\theta_1-\beta_{2z}\theta_2}{\sigma_z}\right)\Bigg].
\end{aligned}$$

where we dropped the conditioning variables for simplicity. Recall that the parameters of wage and amenity offers are supposedly known at this stage of the EM algorithm. For instance, they are equal to the parameters of the wage and amenity equations posterior to constrained job change. Lastly, to find initial conditions in the first M-step of the algorithm we performed a PROBIT regression of z_t on x, weighted by the posterior probabilities.

B.2.3 Inference

Lastly, we turn to the estimation of asymptotic standard errors. Slightly modifying the moment conditions derived by Arcidiacono and Jones (2003) to account for the conditionning, we obtain, from (B.24)-(B.26):

$$\begin{aligned}
\mathbb{E}\left(\sum_{\theta_1,\theta_2} p_{\theta_1,\theta_2}(X|(y,\boldsymbol{a})_{t_0};\boldsymbol{\Theta})\frac{\partial}{\partial\Theta_1}\ln\pi_{\theta_1,\theta_2}((y,\boldsymbol{a})_{t_0};\Theta_1)\right) &= 0,\\
\mathbb{E}\left(\sum_{\theta_1,\theta_2} p_{\theta_1,\theta_2}(X|(y,\boldsymbol{a})_{t_0};\boldsymbol{\Theta})\frac{\partial}{\partial\Theta_2^{-z}}\ln g\left((y,\mathbf{a}),e,z^c,z|e,\theta_1,\theta_2,x;\Theta_2^{-z}\right)\right) &= 0,\\
\mathbb{E}\left(\sum_{\theta_1,\theta_2} p_{\theta_1,\theta_2}(X|(y,\boldsymbol{a})_{t_0};\boldsymbol{\Theta})\frac{\partial}{\partial\Theta_2^{z}}\ln h\left(z,(y,\mathbf{a})|z^c=z=0,e=1,\theta_1,\theta_2,x;\Theta_2^z,\Theta_2^{-z}\right)\right) &= 0.
\end{aligned}$$

Arcidiacono and Jones (2003) propose to estimate the asymptotic variance-covariance matrix by using the usual GMM formula. However, this formula involves first-order derivatives of the moment conditions; that is: second derivatives of the (sequential) likelihood. In models allowing for many parameters, the calculation of these quantities can be cumbersome. In our case, indeed, we found second derivatives of the likelihoods to be long to compute. With more than 200 parameters in the model, the computation of the matrix of second derivatives requires 20000 evaluations (numerical evaluations of the likelihood, or analytical computations of second derivatives).

To circumvent this problem, we use the *generalized information matrix equality* derived by Newey and McFadden (1994), p. 2163. To see how this works in the case of a mixture of sequential likelihoods, let us rewrite the above system of moment equations in compact form:

$$\mathbb{E}\left(\sum_k p_k \frac{\partial \ln \pi_k}{\partial \pi}\right) = 0, \tag{B.28}$$

$$\mathbb{E}\left(\sum_k p_k \frac{\partial \ln g_k}{\partial \alpha}\right) = 0, \tag{B.29}$$

$$\mathbb{E}\left(\sum_k p_k \frac{\partial \ln h_k}{\partial \beta}\right) = 0, \tag{B.210}$$

where:

$$f(y|x;\pi,\alpha,\beta) = \sum_k \pi_k(x;\pi) g_k(y|x;\alpha) h_k(y|x;\alpha,\beta),$$

is the incomplete likelihood, and

$$p_k = p(k|y,x;\pi,\alpha,\beta) = \frac{\pi_k(x;\pi) g_k(y|x;\alpha) h_k(y|x;\alpha,\beta)}{f(y|x;\pi,\alpha,\beta)},$$

is the posterior probability of group k given the data. Vector parameters α, β, π are mutually exclusive.

As f is a mixture of sequential, or *partial* likelihoods in the sense of Cox (1975), the following equality holds:

$$\int g_k(y|x;\alpha_1) h_k(y|x;\alpha_2,\beta) dy = 1, \tag{B.211}$$

for all k, all x and for *all* $(\alpha_1, \alpha_2, \beta)$. This point is crucial, as it is used in the consistency proof of the partial likelihood estimator. Moreover, it implies that the moment conditions (B.28)-(B.210) are satisfied over the whole parameter space, as noticed by Newey and McFadden (1994) in the general context of GMM estimators.

To see why, let us consider (B.29). Differentiating (B.211) with respect to α_1 yields, for *all* parameters, and under suitable regularity conditions:

$$\begin{aligned}
\int \frac{\partial \ln g_k(y|x;\alpha_1)}{\partial \alpha_1} g_k(y|x;\alpha_1) h_k(y|x;\alpha_2,\beta) dy &= \int \frac{\partial g_k(y|x;\alpha_1)}{\partial \alpha_1} h_k(y|x;\alpha_2,\beta) dy, \\
&= \frac{\partial}{\partial \alpha_1} \int g_k(y|x;\alpha_1) h_k(y|x;\alpha_2,\beta) dy, \\
&= 0.
\end{aligned}$$

It follows that:

$$\begin{aligned}
\int \sum_k p(k|y,x;\pi,\alpha_1,\alpha_2,\beta) \frac{\partial \ln g_k(y|x;\alpha_1)}{\partial \alpha_1} f(y|x;\pi,\alpha_1,\alpha_2,\beta) dy &= \\
\sum_k \pi_k(x;\pi) \int \frac{\partial \ln g_k(y|x;\alpha_1)}{\partial \alpha_1} g_k(y|x;\alpha_1) h_k(y|x;\alpha_2,\beta) dy &= 0,
\end{aligned}$$

for *all* parameters, at not only at true values.

The same argument applies to (B.210). Moreover, for all parameters:

$$\begin{aligned}
& \int \sum_k p(k|y,x;\pi,\alpha_1,\alpha_2,\beta)\frac{\partial \ln \pi_k(x;\pi)}{\partial \pi} f(y|x;\pi,\alpha_1,\alpha_2,\beta)dy \\
= & \sum_k \frac{\partial \pi_k(x;\pi)}{\partial \pi} \int g_k(y|x;\alpha_1)h_k(y|x;\alpha_2,\beta)dy \\
= & \frac{\partial \sum_k \pi_k(x;\pi)}{\partial \pi} = 0.
\end{aligned}$$

Let us now consider the first term in the matrix of second derivatives. From the previous argument we obtain:

$$\begin{aligned}
0 &= \frac{\partial}{\partial \pi}\left(\int \sum_k p_k \frac{\partial \ln \pi_k}{\partial \pi} f(y)dy\right), \\
&= \mathbb{E}\left(\frac{\partial}{\partial \pi}\left(\sum_k p_k \frac{\partial \ln \pi_k}{\partial \pi}\right)\right) + \int \sum_k p_k \frac{\partial \ln \pi_k}{\partial \pi}\frac{\partial f(y)'}{\partial \pi}dy, \\
&= \mathbb{E}\left(\frac{\partial}{\partial \pi}\left(\sum_k p_k \frac{\partial \ln \pi_k}{\partial \pi}\right)\right) + \mathbb{E}\left(\sum_k p_k \frac{\partial \ln \pi_k}{\partial \pi}\frac{\partial \ln f(y)'}{\partial \pi}\right),
\end{aligned}$$

where from now on all functions and derivatives are evaluated at true values. Now, notice that:

$$\begin{aligned}
\frac{\partial f(y)}{\partial \pi} &= \frac{\partial}{\partial \pi}\left(\sum_k \pi_k g_k h_k\right), \\
&= \sum_k \frac{\partial \ln \pi_k}{\partial \pi}\pi_k g_k h_k, \\
&= \left(\sum_k p_k \frac{\partial \ln \pi_k}{\partial \pi}\right) f(y).
\end{aligned}$$

It thus follows that:

$$\mathbb{E}\left(\frac{\partial}{\partial \pi}\left(\sum_k p_k \frac{\partial \ln \pi_k}{\partial \pi}\right)\right) = -\mathbb{E}\left(\left(\sum_k p_k \frac{\partial \ln \pi_k}{\partial \pi}\right)\left(\sum_k p_k \frac{\partial \ln \pi_k}{\partial \pi}\right)'\right).$$

We can derive similar expressions for the eight other terms by using the same methodology. Namely, we find after some calculation:

$$\mathbb{E}\left(\frac{\partial}{\partial \alpha}\left(\sum_k p_k \frac{\partial \ln \pi_k}{\partial \pi}\right)\right) = -\mathbb{E}\left(\left(\sum_k p_k \frac{\partial \ln \pi_k}{\partial \pi}\right)\left(\sum_k p_k \frac{\partial \ln (g_k h_k)}{\partial \alpha}\right)'\right),$$

$$\mathbb{E}\left(\frac{\partial}{\partial \beta}\left(\sum_k p_k \frac{\partial \ln \pi_k}{\partial \pi}\right)\right) = -\mathbb{E}\left(\left(\sum_k p_k \frac{\partial \ln \pi_k}{\partial \pi}\right)\left(\sum_k p_k \frac{\partial \ln h_k}{\partial \beta}\right)'\right),$$

$$\mathbb{E}\left(\frac{\partial}{\partial\pi}\left(\sum_k p_k\frac{\partial\ln g_k}{\partial\alpha}\right)\right) = -\mathbb{E}\left(\left(\sum_k p_k\frac{\partial\ln g_k}{\partial\alpha}\right)\left(\sum_k p_k\frac{\partial\ln\pi_k}{\partial\pi}\right)'\right),$$

$$\mathbb{E}\left(\frac{\partial}{\partial\alpha}\left(\sum_k p_k\frac{\partial\ln g_k}{\partial\alpha}\right)\right) = -\mathbb{E}\left(\left(\sum_k p_k\frac{\partial\ln g_k}{\partial\alpha}\right)\left(\sum_k p_k\frac{\partial\ln(g_k h_k)}{\partial\alpha}\right)'\right),$$

$$\mathbb{E}\left(\frac{\partial}{\partial\beta}\left(\sum_k p_k\frac{\partial\ln g_k}{\partial\alpha}\right)\right) = -\mathbb{E}\left(\left(\sum_k p_k\frac{\partial\ln g_k}{\partial\alpha}\right)\left(\sum_k p_k\frac{\partial\ln h_k}{\partial\beta}\right)'\right),$$

$$\mathbb{E}\left(\frac{\partial}{\partial\pi}\left(\sum_k p_k\frac{\partial\ln h_k}{\partial\beta}\right)\right) = -\mathbb{E}\left(\left(\sum_k p_k\frac{\partial\ln h_k}{\partial\beta}\right)\left(\sum_k p_k\frac{\partial\ln\pi_k}{\partial\pi}\right)'\right),$$

$$\mathbb{E}\left(\frac{\partial}{\partial\alpha}\left(\sum_k p_k\frac{\partial\ln h_k}{\partial\beta}\right)\right) = -\mathbb{E}\left(\left(\sum_k p_k\frac{\partial\ln h_k}{\partial\beta}\right)\left(\sum_k p_k\frac{\partial\ln(g_k h_k)}{\partial\alpha}\right)'\right),$$

$$\mathbb{E}\left(\frac{\partial}{\partial\beta}\left(\sum_k p_k\frac{\partial\ln h_k}{\partial\beta}\right)\right) = -\mathbb{E}\left(\left(\sum_k p_k\frac{\partial\ln h_k}{\partial\beta}\right)\left(\sum_k p_k\frac{\partial\ln h_k}{\partial\beta}\right)'\right).$$

It is clear from these expressions that the information matrix equality is not satisfied if h_k depends in a non-trivial manner on α, *i.e.* if the sequential EM is not equivalent to ML. However, the computation of the matrix of cross-products of first derivatives involves roughly the same number of calculations as the computation of the matrix composed of the above terms.

In problems where the number of parameters is large and the (sequential) likelihood is not straightforward to compute, this idea provides a fast alternative way to compute standard errors.

Appendix C

Appendix of Chapter 3

C.1 Mathematical proofs

C.1.1 Proof of Theorem 3

The proof of proposition (i) is a straightforward consequence of Theorem 10.3.1 in Kagan, Linnik and Rao (1973).

Theorem 17 ***(Theorem 10.3.1, Kagan, Linnik and Rao, 1973)*** *Let A and B be two non-stochastic matrices and let $S = (s_1, ..., s_m)^T$ and $R = (r_1, ..., r_n)^T$ be two random vectors with independent components. Assume that AS and BR have the same distribution. If s_i, for some $i \leq m$, is not normal, then the ith column of A is the multiple of a column of B.*

Assume that $\Lambda X + U$ and $\widetilde{\Lambda}\widetilde{X} + \widetilde{U}$ have the same distribution. The components of vectors (X^T, U^T) and $(\widetilde{X}^T, \widetilde{U}^T)$, respectively, are independent. Let $k \leq K$. Since X_k is not normal, Kagan *et al.*'s result applies to show that the k's column of Λ, say Λ_k, is the multiple of a column of the $L \times (K + L)$ matrix $(\widetilde{\Lambda}, I_L)$, where I_L is the $L \times L$ identity matrix. Since every column of matrices Λ and $\widetilde{\Lambda}$ has at least two non-zero coefficients, it must be that Λ_k is the multiple of a column of $\widetilde{\Lambda}$. This shows proposition (i).

To show proposition (ii) let $\boldsymbol{\kappa}''_Y(t) = \left(\kappa_Y^{(\alpha)}(t), \alpha \in \Delta_{L,2}\right)$, for $t \in \mathbb{R}^L$, be the $\#\Delta_{L,2} \times 1$ vector of second-order partial cross-derivatives of $\kappa_Y(t)$. Let also

$$\begin{aligned}
\boldsymbol{\kappa}_X(t) &= (\kappa_{X_1}(t_1), ..., \kappa_{X_K}(t_K))^T \\
\boldsymbol{\kappa}'_X(t) &= \left(\kappa'_{X_1}(t_1), ..., \kappa'_{X_K}(t_K)\right)^T, \\
\boldsymbol{\kappa}''_X(t) &= \left(\kappa''_{X_1}(t_1), ..., \kappa''_{X_K}(t_K)\right)^T, \quad t = (t_1, ..., t_K) \in \mathbb{R}^K.
\end{aligned}$$

Equation (4.11) implies the following restrictions on factor cumulant generating functions:

$$\boldsymbol{\kappa}''_Y(t) = Q(\Lambda)\boldsymbol{\kappa}''_X\left(\Lambda^T t\right). \tag{C.11}$$

To show the second proposition, remark that it must be that

$$Q(\Lambda)\boldsymbol{\kappa}''_X\left(t^T\Lambda\right) = Q(\widetilde{\Lambda})\boldsymbol{\kappa}''_{\widetilde{X}}\left(t^T\widetilde{\Lambda}\right). \tag{C.12}$$

By proposition (i), every column of Λ is a scalar multiple of a column of $\widetilde{\Lambda}$. Since $\text{rank}(Q(\Lambda)) = K$, it follows that there exists no couple of columns of Λ which are proportional. Therefore, there exist a permutation matrix P and a diagonal matrix D with non zero entries in the diagonal such that $\widetilde{\Lambda} = \Lambda DP$. Now, since $\ker(Q(\Lambda)) = 0$, (C.12) implies:

$$\boldsymbol{\kappa}''_X\left(t^T\Lambda\right) = \boldsymbol{\kappa}''_{DP\widetilde{X}}\left(t^T\Lambda\right). \tag{C.13}$$

Taking this equation at $t = 0$ and using the normalization assumption A2 yields:

$$D^2 = \text{Var}(DP\widetilde{X}) = \text{Var}(X) = I_K.$$

Moreover, integrating the differential equation (C.13) shows that $\boldsymbol{\kappa}_X$ and $\boldsymbol{\kappa}_{DP\widetilde{X}}$ differ by an affine function. By definition, $\boldsymbol{\kappa}_X(0) = \boldsymbol{\kappa}_{DP\widetilde{X}}(0) = 0$. By assumption, since the factor distributions have zero mean: $\boldsymbol{\kappa}'_X(0) = \boldsymbol{\kappa}'_{DP\widetilde{X}}(0) = 0$. This shows that the d.f. of X and $DP\widetilde{X}$ are equal. Lastly, the d.f. of U and $\widetilde{U}$ are equal by deconvolution, since the characteristic functions of the factors are nonvanishing everywhere.

This ends the proof.

C.1.2 Proof of Theorem 5

Let $x = (x_1, ..., x_K)^T \in \ker(Q(\Lambda))$ such that $x \neq 0$. For all $k = 1, ..., K$, and all $\ell = 1, ..., L$, define

$$\begin{aligned}\psi_k(\tau_k) &= \kappa_{X_k}(\tau_k) - x_k\frac{\tau_k^2}{2},\ \forall\tau_k \in \mathbb{R},\\ \zeta_\ell(t_\ell) &= \kappa_{U_\ell}(t_\ell) + (\Lambda_\ell \otimes \Lambda_\ell)\,x\frac{t_\ell^2}{2},\ \forall t_\ell \in \mathbb{R},\end{aligned}$$

where $\Lambda_\ell = (\lambda_{\ell 1}, ..., \lambda_{\ell K})$ is the ℓth row of Λ and $\otimes$ is the Kronecker product ($(\Lambda_\ell \otimes \Lambda_\ell)\,x = \sum_{k=1}^K x_k\lambda_{\ell k}^2$). If $x_k \geq 0$, $\psi_k$is the the cumulant generating function (c.g.f.) of the convolution of the distribution of X_k and the normal distribution $\mathcal{N}\left(0, \sqrt{x_k}\right)$. Now, suppose that $x_k < 0$. The distribution of X_k is divisible by a normal distribution, say $\mathcal{N}\left(0, \sigma_k^2\right)$. If $\sigma_k^2 + x_k > 0$, then ψ_k is the c.g.f. of some random variable that is the sum of the random variable with c.g.f. $\kappa_{X_k}(\tau_k) + \frac{\sigma_k^2\tau_k^2}{2}$ and of the normal variable $\mathcal{N}\left(0, \sigma_k^2 + x_k\right)$. The same argument applies to ζ_ℓ. If $(\Lambda_\ell \otimes \Lambda_\ell)\,x \leq 0$, then ζ_ℓ is the c.g.f. of $U_\ell + \mathcal{N}\left(0, -(\Lambda_\ell \otimes \Lambda_\ell)\,x\right)$. Otherwise, U_ℓ is divisible by a normal distribution, say $\mathcal{N}\left(0, \omega_\ell^2\right)$. If $\omega_\ell^2 - (\Lambda_\ell \otimes \Lambda_\ell)\,x > 0$, then ζ_ℓ is the c.g.f. of some random variable that is the sum of the random variable whose c.g.f. is $\kappa_{U_\ell}(t_\ell) + \frac{\omega_\ell^2 t_\ell^2}{2}$ and the normal variable $N\left(0, \omega_\ell^2 - (\Lambda_\ell \otimes \Lambda_\ell)\,x\right)$. Rescale x if necessary so that $x_k > -\sigma_k^2$, for all $k = 1, ..., K$, and $\omega_\ell^2 > (\Lambda_\ell \otimes \Lambda_\ell)\,x$, for all $\ell = 1, ..., L$. One can thus construct $K + L$ non degenerate, independent random variables with zero mean and finite variance: $Z_1, ..., Z_K, \widetilde{U}_1, ..., \widetilde{U}_L$, with given c.g.f.'s $\psi_1, ..., \psi_K, \zeta_1, ..., \zeta_L$.

Next, for all $t = (t_1, ..., t_L)^T$, define

$$\begin{aligned}\kappa(t) &\equiv \sum_{k=1}^{K} \psi_k(\boldsymbol{\lambda}_k^T t) + \sum_{\ell=1}^{L} \zeta_\ell(t_\ell) \\ &= \sum_{k=1}^{K} \kappa_{X_k}(\boldsymbol{\lambda}_k^T t) - \sum_{k=1}^{K} x_k \frac{\left(\boldsymbol{\lambda}_k^T t\right)^2}{2} + \sum_{\ell=1}^{L} \kappa_{U_\ell}(t_\ell) + \sum_{\ell=1}^{L} (\Lambda_\ell \otimes \Lambda_\ell)\, x \frac{t_\ell^2}{2}.\end{aligned}$$

As $Q(\Lambda)x = 0$, $\sum_{k=1}^{K} x_k \lambda_{\ell k} \lambda_{mk} = 0$ for all $\ell \neq m$ in $\{1, ..., L\}$. Hence,

$$\begin{aligned}\sum_{k=1}^{K} x_k \left(\boldsymbol{\lambda}_k^T t\right)^2 &= \sum_{k=1}^{K} x_k \sum_{\ell=1}^{L} \lambda_{\ell k}^2 t_\ell^2 = \sum_{\ell=1}^{L} \sum_{k=1}^{K} x_k \lambda_{\ell k}^2 t_\ell^2 \\ &= \sum_{\ell=1}^{L} (\Lambda_\ell \otimes \Lambda_\ell)\, x t_\ell^2;\end{aligned}$$

and, therefore,

$$\kappa(t) = \sum_{k=1}^{K} \kappa_{X_k}(\boldsymbol{\lambda}_k^T t) + \sum_{\ell=1}^{L} \kappa_{U_\ell}(t_\ell) = \kappa_Y(t).$$

Now, define D as the diagonal of order K with diagonal entries: $d_k = \sqrt{1 - x_k}$. Rescale x if necessary such that D is invertible. Then:

$$\text{Var}(D^{-1}Z) = D^{-1}\, \text{diag}(1 - x_k) D^{-1} = I_K.$$

It follows that $\left(\Lambda D, D^{-1}Z, \widetilde{U}\right)$ is an alternative representation to (Λ, X, U).

Lastly, we have to show that these two representations are different. Note that, by the above construction, one can find an *infinity* of alternative representations $\left(\widetilde{\Lambda}, \widetilde{X}, \widetilde{U}\right)$ by appropriately rescaling x. Since the cardinal of $\mathcal{S}_K$ is *finite*, it follows that (Λ, X, U) is not identified. This ends the proof.

C.1.3 Proof of Theorem 6

To prove Theorem 6, we first prove the following lemma giving conditions under which the joint eigenvectors of a set of matrices is uniquely defined (up to sign and permutation).

Lemma 18 *Let K and L be any integers. Let $A_1, ..., A_L$ be matrices of $\mathbb{R}^{K \times K}$. Suppose that there exist $x^k = \left(x_1^k, ..., x_L^k\right)^T \in \mathbb{R}^L$ and $v^k \in \mathbb{R}^K$, $v_k \neq 0$, $k = 1, ..., K+1$, solutions to the joint diagonalization problem:*

$$x_\ell^k v^k = A_\ell v^k, \quad \forall \ell = 1, ..., L.$$

Assume that the set $\left\{v^1, ..., v^K\right\}$ is linearly independent, that all v_k, $k = 1, ..., K+1$, have norm one, and that $x^k \neq x^{k'}$ for all $(k, k') \in \{1, ..., K\}^2$. Then there exists $k \in \{1, ..., K\}$ such that $v^{K+1} = \pm v^k$.

Proof. Since $\{v^1, ..., v^K\}$ is a basis of $\mathbb{R}^K$, there exists $c = (c_1, ..., c_K) \neq 0$ such that $v^{K+1} = c_1 v^1 + ... + c_K v^K$. Then, for all $\ell = 1, ..., L$,

$$\begin{aligned}
\sum_{k=1}^{K} c_k x_\ell^k v^k &= \sum_{\ell=1}^{K} c_k A_\ell v^k \\
&= A_\ell \sum_{\ell=1}^{K} c_k v^k \\
&= A_\ell v^{K+1} \\
&= x_\ell^{K+1} v^{K+1} \\
&= x_\ell^{K+1} \left(\sum_{k=1}^{K} c_k v^k \right).
\end{aligned}$$

As $(v^1, ..., v^K)$ is linearly independent, it follows from the last equality that:

$$c_k x_\ell^k = c_k x_\ell^{K+1},$$

for all (k, ℓ). Hence, for all k:

$$c_k x^k = c_k x^{K+1}.$$

As $c \neq 0$, there exists k such that $c_k \neq 0$. For this k: $x^k = x^{K+1}$. Moreover, as $x^k \neq x^{k'}$ for all $k' \neq k$ in $\{1, ..., K\}$, it follows that $c_{k'} = 0$ for all $k' \neq k$. Hence

$$v^{K+1} = c_k v^k.$$

As both v^k and v^{K+1} have norm one, $c_k = \pm 1$. The result follows. ■

The proof of Theorem 6 easily follows.

Fourth-order moments. In the case where $U = 0$, second and fourth-order cumulant restrictions (3.8)-(3.12) yield:

$$\begin{aligned}
\Omega_Y(\ell, m) &= \Lambda D_4 \operatorname{diag}(\Lambda_\ell \odot \Lambda_m) \Lambda^T, \quad (\ell, m) \in \overline{\Delta}_{L,2}, \\
\Sigma_Y &= \Lambda \Lambda^T.
\end{aligned}$$

To show that Λ is identified from this system, let P be the Cholesky decomposition of Σ_Y, such that $PP^T = \Sigma_Y - \Sigma_U$, and P is a lower triangular $L \times K$ full-column rank matrix.

Then $P^- \Lambda$, where P^- is a generalized inverse of P (*e.g.* $P^- = \left[P^T P\right]^{-1} P^T$), is a matrix of joint orthonormal eigenvectors of:

$$P^- \Omega_Y(\ell, m) P^{-T} = P^- \Lambda D_4 \operatorname{diag}(\Lambda_\ell \odot \Lambda_m) \Lambda^T P^{-T}, \quad (\ell, m) \in \overline{\Delta}_{L,2}.$$

In general, there can be infinitely many joint eigenvectors to a set of matrices if all matrices have multiple roots. Lemma 18 shows that the problem of diagonalizing matrices $P^- \Omega_Y(\ell, m) P^{-T}$, $(\ell, m) \in \overline{\Delta}_{L,2}$, has a unique solution up to column sign and permutation if for all $(k, k') \in \{1...K\}^2$, $k \neq k'$, there exists $(\ell, m) \in \overline{\Delta}_{L,2}$ such that

$$\lambda_{\ell k} \lambda_{mk} \kappa_4(X_k) \neq \lambda_{\ell k'} \lambda_{mk'} \kappa_4(X_{k'}).$$

As either $\kappa_4(X_k) \neq 0$ or $\kappa_4(X_{k'}) \neq 0$, and as any two columns of Λ are linearly independent, this condition is always satisfied. It follows that V, and thus $\Lambda = PV$, are identified (up to column sign and permutation).

Third-order moments. The same argument applies to third-order cumulant matrices $\Gamma_Y(\ell)$. Indeed, in the noise-free case third-order restrictions (3.10) become

$$\Gamma_Y(\ell) = \Lambda D_3 \operatorname{diag}(\Lambda_\ell)\Lambda^T, \quad \ell \in \{1...L\},$$

where $\Gamma_Y(\ell)$, for all $\ell \in \{1...L\}$, is a $L \times L$ matrix of third-order cumulants of the data, and D_3 is the diagonal matrix of factor cumulants.

In this case, Lemma 18 shows that the problem of diagonalizing matrices $P^- \Gamma_Y(\ell) P^{-T}$, $\ell \in \{1...L\}$, has a unique solution up to column sign and permutation if for all $(k,k') \in \{1...K\}^2$, $k \neq k'$, there exists $\ell \in \{1...L\}$ such that

$$\lambda_{\ell k}\kappa_3(X_k) \neq \lambda_{\ell k'}\kappa_3(X_{k'}).$$

As before, this condition is always satisfied.

Third and fourth-order moments. The proof is almost identical to the two previous ones. Lemma 18 shows that the problem of diagonalizing matrices $P^- \Omega_Y(\ell,m) P^{-T}$, $(\ell,m) \in \overline{\Delta}_{L,2}$, and $P^- \Gamma_Y(\ell) P^{-T}$, $\ell \in \{1...L\}$, has a unique solution up to column sign and permutation if for all $(k,k') \in \{1...K\}^2$, $k \neq k'$, there exists $(\ell,m) \in \overline{\Delta}_{L,2}$ such that

$$\lambda_{\ell k}\lambda_{mk}\kappa_4(X_k) \neq \lambda_{\ell k'}\lambda_{mk'}\kappa_4(X_{k'}),$$

or there exists $\ell \in \{1...L\}$ such that

$$\lambda_{\ell k}\kappa_3(X_k) \neq \lambda_{\ell k'}\kappa_3(X_{k'}).$$

As one of the four moments $\kappa_3(X_k)$, $\kappa_3(X_{k'})$, $\kappa_4(X_k)$ and $\kappa_4(X_{k'})$ is non zero, it follows from the assumptions on Λ that this condition is always satisfied.

C.1.4 Proof of Lemma 7

1. Let Ω_Y be defined by (3.16). As Q has rank K and D_4 is non singular, restrictions (3.17) imply that

$$\Omega_Y = \overline{Q} D_4 Q^T,$$

has rank K. It follows that there exists $\overline{C} \in \mathbb{R}^{\#\overline{\Delta}_{L,2} \times (\#\overline{\Delta}_{L,2} - K)}$, full column rank, such that $\overline{C}^T \Omega_Y = 0$. Since $D_4 Q^T$ has rank K, it must also be that $\overline{C}^T \overline{Q} = 0$.

2. Let vech be the operator stacking all elements on and below the main diagonal of a $L \times L$ symmetric matrix column by column into a $\frac{L(L+1)}{2}$-vector. Then,

$$\begin{aligned} \operatorname{vech}(\Omega_Y(\ell,m)) &= \operatorname{vech}\left(\Lambda D_4 \operatorname{diag}(\Lambda_\ell \odot \Lambda_m)\Lambda^T + \delta_{\ell m}\kappa_4(U_\ell)\operatorname{Sp}_{L,\ell}\right), \\ &= \overline{Q} D_4 (\Lambda_\ell \odot \Lambda_m) + \delta_{\ell m}\kappa_4(U_\ell)\operatorname{vech}\left(\operatorname{Sp}_{L,\ell}\right), \end{aligned}$$

where $\mathrm{Sp}_{L,\ell}$ is the sparse matrix of dimension (L,L) with only one 1 in position (ℓ,ℓ). It follows that

$$\overline{C}^T \operatorname{vech}\left(\Omega_Y(\ell,m)\right)=\delta_{\ell m}\kappa_4\left(U_\ell\right)\overline{C}_{(\ell,\ell)},$$

where $\overline{C}_{(\ell,\ell)}$ is the (ℓ,ℓ)th column of $\overline{C}^T$, and the columns of $\overline{C}^T$ (the rows of $\overline{C}$) are indexed by $(i,j)\in\overline{\Delta}_{L,2}$.

Moreover, the second-order restrictions are equivalently written as

$$\begin{aligned}\operatorname{vech}\left(\Sigma_Y\right) &= \operatorname{vech}\left(\Lambda\Lambda^T+\Sigma_U\right),\\ &= \overline{Q}\mathbf{1}_K+\operatorname{vech}\left(\Sigma_U\right),\end{aligned}$$

where $\mathbf{1}_K$ is a K-dimensional vector of ones. Hence,

$$\overline{C}^T \operatorname{vech}\left(\Sigma_Y\right)=\overline{C}^T \operatorname{vech}\left(\Sigma_U\right)=\sum_{\ell=1}^{L}\operatorname{Var}\left(U_\ell\right)\overline{C}_{(\ell,\ell)}.$$

Lastly, consider

$$\begin{aligned}\operatorname{vech}\left(\Gamma_Y(\ell)\right) &= \left[\operatorname{Cum}\left(Y_\ell,Y_i,Y_j\right),(i,j)\in\overline{\Delta}_{L,2}\right]\\ &= \operatorname{vech}\left(\Lambda D_3\operatorname{diag}\left(\Lambda_\ell\right)\Lambda^T+\kappa_3\left(U_\ell\right)\mathrm{Sp}_{L,\ell}\right).\end{aligned}$$

This vector of third-order moments of Y satisfies the equality

$$\operatorname{vech}\left(\Gamma_Y(\ell)\right)=\overline{Q}D_3\Lambda_\ell+\kappa_3\left(U_\ell\right)\operatorname{vech}\left(\mathrm{Sp}_{L,\ell}\right).$$

It follows that

$$\overline{C}^T \operatorname{vech}\left(\Gamma_Y(\ell)\right)=\kappa_3\left(U_\ell\right)\overline{C}_{(\ell,\ell)}.$$

3. Lastly, we show that the submatrix $\left[\overline{C}_{(1,1)},...,\overline{C}_{(L,L)}\right]^T\in\mathbb{R}^{L\times\left(\#\overline{\Delta}_{L,2}-K\right)}$ of $\overline{C}$ is full-row rank. To show this assertion, partition $\overline{C}$ as

$$\overline{C}=\left[\begin{array}{cc}\overline{C}_{11} & \overline{C}_{12}\\ \overline{C}_{21} & \overline{C}_{22}\end{array}\right],$$

with $\overline{C}_{11}\in\mathbb{R}^{\#\Delta_{L,2}\times\left(\#\Delta_{L,2}-K\right)}$, $\overline{C}_{12}\in\mathbb{R}^{\#\Delta_{L,2}\times L}$, $\overline{C}_{21}\in\mathbb{R}^{L\times\left(\#\Delta_{L,2}-K\right)}$ and $\overline{C}_{22}\in\mathbb{R}^{L\times L}$. To simplify the notations, suppose that rows $\overline{C}_{(1,1)}^T,...,\overline{C}_{(L,L)}^T$ are located at the bottom of $\overline{C}$, so that $\left[\overline{C}_{21},\overline{C}_{22}\right]=\left[\overline{C}_{(1,1)},...,\overline{C}_{(L,L)}\right]^T$. Without loss of generality, one can assume that $\overline{C}_{21}=0$ and that $\overline{C}_{11}$ is a basis of the null space of Q^T. Now, suppose that $\overline{C}_{22}$ is singular. Then there exists a linear combination of the columns of $\overline{C}_{22}$ that is equal to zero. The same linear combination of the columns of $\overline{C}_{12}$ is both linearly independent of $\overline{C}_{11}$, as $\overline{C}$ is full-column rank, and orthogonal to the columns of Q. This contradicts the assumption that Q has rank K. Consequently, $\overline{C}_{22}$ is non singular and $\left[\overline{C}_{21},\overline{C}_{22}\right]$ is full-row rank.

As matrix $\left[\overline{C}_{(1,1)},...,\overline{C}_{(L,L)}\right]^T$ is full-row rank, it follows that error variances are identified. Moreover, it also follows that $\overline{C}_{(\ell,\ell)}\neq 0$. So, $\kappa_3\left(U_\ell\right)$ and $\kappa_4\left(U_\ell\right)$ are identified.

This ends the proof of Lemma 7.

C.1.5 Proof of Lemma 9

1. The factor structure implies that

$$\begin{aligned}\Xi_Y &= [\Gamma_Y, \Omega_Y(1), ..., \Omega_Y(L)], \\ &= \Lambda\left[D_3 Q^T, D_4 \operatorname{diag}(\Lambda_1) Q^T, ..., D_4 \operatorname{diag}(\Lambda_L) Q^T\right].\end{aligned}$$

Let $\gamma \in \mathbb{R}^K$ such that

$$\gamma^T\left[D_3 Q^T, D_4 \operatorname{diag}(\Lambda_1) Q^T, ..., D_4 \operatorname{diag}(\Lambda_L) Q^T\right] = 0.$$

As Q has rank K, it follows that $\gamma^T D_3 = 0$ and $\gamma^T D_4 \operatorname{diag}(\Lambda_\ell) = 0$ for all $\ell \in \{1...L\}$. Then, as Λ is full column rank, this implies that $\gamma^T D_4 = 0$. Lastly, as for all k either $\kappa_3(X_k) \neq 0$ or $\kappa_4(X_k) \neq 0$, it follows that $\gamma = 0$.

Therefore: $\left[D_3 Q^T, D_4 \operatorname{diag}(\Lambda_1) Q^T, ..., D_4 \operatorname{diag}(\Lambda_L) Q^T\right]$ as rank K. As Λ has rank K by assumption, Ξ_Y has also rank K.

Then, let $C \in \mathbb{R}^{L\times(L-K)}$ such that

$$C^T \Xi_Y = 0.$$

As $\left[D_3 Q^T, D_4 \operatorname{diag}(\Lambda_1) Q^T, ..., D_4 \operatorname{diag}(\Lambda_L) Q^T\right]$ is full row rank, it must also be that $C^T\Lambda = 0$.

2. One thus has

$$\begin{aligned}C^T\Sigma_Y &= C^T\Lambda\Lambda^T + C^T\Sigma_U, \\ &= C^T\Sigma_U \\ &= [\operatorname{Var}(U_1) C_1, ..., \operatorname{Var}(U_L) C_L]\end{aligned}$$

or

$$C^T\begin{pmatrix}\operatorname{Cov}(Y_1, Y_\ell) \\ \vdots \\ \operatorname{Cov}(Y_L, Y_\ell)\end{pmatrix} = \operatorname{Var}(U_\ell) C_\ell, \quad \ell = 1, ..., L,$$

where C_ℓ^T is the ℓth row of C.

Moreover, matrices $\Gamma_Y(\ell)$ defined by (3.9) satisfy the equality:

$$\Gamma_Y(\ell) = \Lambda D_3 \operatorname{diag}(\Lambda_\ell)\Lambda^T + \kappa_3(U_\ell)\operatorname{Sp}_{L,\ell}.$$

Hence

$$\begin{aligned}C^T\Gamma_Y(\ell) &= C^T\Lambda D_3 \operatorname{diag}(\Lambda_\ell)\Lambda^T + \kappa_3(U_\ell) C^T \operatorname{Sp}_{L,\ell}, \\ &= \kappa_3(U_\ell) C^T \operatorname{Sp}_{L,\ell},\end{aligned}$$

or

$$C^T\begin{pmatrix}\operatorname{Cum}(Y_1, Y_\ell, Y_\ell) \\ \vdots \\ \operatorname{Cum}(Y_L, Y_\ell, Y_\ell)\end{pmatrix} = \kappa_3(U_\ell) C_\ell.$$

Lastly,

$$\Omega_Y(\ell,\ell) = \Lambda D_4 \operatorname{diag}(\Lambda_\ell \odot \Lambda_\ell)\Lambda^T + \kappa_4(U_\ell)\operatorname{Sp}_{L,\ell}$$

implies that

$$C^T\Omega_Y(\ell,\ell) = \kappa_4(U_\ell)\, C^T \operatorname{Sp}_{L,\ell}$$

and

$$C^T \begin{pmatrix} \operatorname{Cum}(Y_1, Y_\ell, Y_\ell, Y_\ell) \\ \vdots \\ \operatorname{Cum}(Y_L, Y_\ell, Y_\ell, Y_\ell) \end{pmatrix} = \kappa_4(U_\ell)\, C_\ell.$$

3. Let $\Lambda_{-\ell}$ be matrix Λ without its ℓth row. As $\Lambda_{-\ell}$ has rank K by assumption, it follows from equality $C^T\Lambda = 0$ that $C_\ell \neq 0$. Otherwise, one would have $C_{-\ell}^T\Lambda_{-\ell} = 0$ for a full $(L-1)\times(L-K)$ matrix $C_{-\ell}$, contradicting the assumption that $\operatorname{rank}(\Lambda_{-\ell}) = K$. Hence $\operatorname{Var}(U_\ell)$, $\kappa_3(U_\ell)$ and $\kappa_4(U_\ell)$ are identified.

This ends the proof of Lemmas 9 and 10.

C.2 The JADE algorithm

Let $\mathcal{A} = \{A_k, k = 1...K\}$ a set of real symmetric $L \times L$ matrices. Let us define the function:

$$\operatorname{off}(A) = \sum_{i\neq j} a_{ij}^2,$$

for all $A = [a_{ij}]$. Then joint diagonalization of $\mathcal{A}$ is achieved by minimizing

$$\sum_{k=1}^{K} \operatorname{off}(UA_kU^T), \tag{C.24}$$

with respect to U orthonormal.

Let $\theta \in [-\pi,\pi]$, let $(i,j) \in \{1...L\}^2$ and let $R_{ij}(\theta)$ be the $L\times L$ matrix equal to zero everywhere except at the (i,i), (i,j), (j,i) and (j,j) entries where it is equal to:

$$\begin{pmatrix} \cos(\theta) & \sin(\theta) \\ -\sin(\theta) & \cos(\theta) \end{pmatrix}.$$

Let $i \neq j$, and let us define:

$$O_{i,j}(\theta) = \sum_{k=1}^{K} \operatorname{off}\left(R_{ij}(\theta)A_kR_{ij}(\theta)^T\right).$$

Lastly, let $h_{i,j}(A) = (a_{ii} - aij, a_{ij} + a_{ji})$, and let:

$$G_{i,j} = \sum_{k=1}^{K} h_{i,j}^T(A_k)h_{i,j}(A_k) = (g_{ij})_{i,j=1,2}.$$

Then, Cardoso and Souloumiac (1996) show that θ_0 such that:

$$\cos(\theta_0) = \sqrt{\frac{x+r}{2r}}, \quad \sin(\theta_0) = \sqrt{\frac{y}{2r(x+r)}},$$

where $x = g_{11} - g_{22}$, $y = g_{12} + g_{21}$ and $r = \sqrt{x^2 + y^2}$, minimizes $O_{i,j}(\theta)$.

This closed-form expression for θ_0 allows to minimize (C.24) by the following algorithm:

1. Start with $U(0) = I_L$.
2. Begin loop on step s.
3. Begin loop on (i, j).
4. Compute $G_{i,j}$.
5. Compute θ_0.
6. If θ_0 is different enough from zero, continue. Else stop.
7. Compute $R_{ij}(\theta_0) A_k R_{ij}(\theta_0)^T$ and modify $\mathcal{A}$ consequently.
8. Update $U(s)$ as $U(s+1) = R_{ij}(\theta_0)U(s)$.
9. End loop on (i, j).
10. End loop on s.

C.3 Asymptotic theory of the JADE estimator

First-order conditions. The JADE estimator solves

$$\widehat{V} = \arg \min_{V \in \mathcal{O}_K} \sum_{j=1}^{J} \text{off}(V^T \widehat{A}_j V).$$

The Lagrangian associated with the minimization problem is:

$$\begin{aligned} \mathcal{L}(V, \gamma) &= \sum_{j=1}^{J} \text{off}(V^T \widehat{A}_j V) + \gamma^T \text{vec}\left(V^T V - I_K\right), \\ &= \sum_{j} \sum_{m \neq k} (v_k^T \widehat{A}_j v_m)^2 + \sum_{k} \gamma_{kk}(v_k^T v_k - 1) + \sum_{m \neq k} \gamma_{mk} v_k^T v_m, \end{aligned}$$

where γ is a vector of K^2 Lagrange multipliers γ_{mk}, and v_k is the kth column of matrix V.

Differentiating the Lagrangian with respect to v_ℓ, for $\ell = 1...K$, yields:

$$\frac{\partial \mathcal{L}\left(\widehat{V}, \widehat{\gamma}\right)}{\partial v_\ell} = 2 \sum_{j} \sum_{k \neq \ell} (\widehat{v}_k^T \widehat{A}_j \widehat{v}_\ell) \widehat{A}_j \widehat{v}_k + 2\widehat{\gamma}_{\ell\ell} \widehat{v}_\ell + \sum_{k \neq \ell} \widehat{\gamma}_{k\ell} \widehat{v}_k = 0.$$

Then, multiplying this equation by $\widehat{v}_m^T$, for $m \neq \ell$, gives:

$$2\sum_j \sum_{k\neq \ell} (\widehat{v}_k^T \widehat{A}_j \widehat{v}_\ell)\widehat{v}_m^T \widehat{A}_j \widehat{v}_k + \widehat{\gamma}_{m\ell} = 0.$$

Using that $\widehat{\gamma}_{m\ell} = \widehat{\gamma}_{\ell m}$ by symmetry, it follows that

$$\sum_j \sum_{k\neq \ell} (\widehat{v}_k^T \widehat{A}_j \widehat{v}_\ell)\widehat{v}_m^T \widehat{A}_j \widehat{v}_k = \sum_j \sum_{k\neq m} (\widehat{v}_k^T \widehat{A}_j \widehat{v}_m)\widehat{v}_\ell^T \widehat{A}_j \widehat{v}_k,$$

or, equivalently, as $\widehat{A}_j$ is symmetric for all j:

$$\sum_j \widehat{v}_\ell^T \widehat{A}_j \left(\sum_{k\neq \ell} \widehat{v}_k \widehat{v}_k^T\right) \widehat{A}_j \widehat{v}_m = \sum_j \widehat{v}_m^T \widehat{A}_j \left(\sum_{k\neq m} \widehat{v}_k \widehat{v}_k^T\right) \widehat{A}_j \widehat{v}_\ell.$$

Then, as $\sum_{k=1}^K \widehat{v}_k \widehat{v}_k^T = \widehat{V}\widehat{V}^T = I_K$ we obtain

$$\sum_j \widehat{v}_\ell^T \widehat{A}_j \left(I_K - \widehat{v}_\ell \widehat{v}_\ell^T\right) \widehat{A}_j \widehat{v}_m = \sum_j \widehat{v}_m^T \widehat{A}_j \left(I_K - \widehat{v}_m \widehat{v}_m^T\right) \widehat{A}_j \widehat{v}_\ell,$$

which we write after rearranging:

$$\sum_j \widehat{v}_\ell^T \widehat{A}_j \widehat{v}_m \left(\widehat{v}_m^T \widehat{A}_j \widehat{v}_m - \widehat{v}_\ell^T \widehat{A}_j \widehat{v}_\ell\right) = 0. \tag{C.35}$$

Equation (C.35) holds for all $\ell < m$. The JADE estimator $\widehat{V}$ solves these $K(K-1)/2$ non redundant equations, together with the $K(K+1)/2$ orthogonality constraints:

$$\widehat{v}_\ell^T \widehat{v}_m = \delta_{\ell m}, \text{ for all } \ell \leq m.$$

Identification and consistency. Let $\widetilde{V} = (\widetilde{v}_1, ..., \widetilde{v}_K) \in \mathcal{O}_K$ be such that

$$\widetilde{V} = \arg\min_{V\in\mathcal{O}_K} \sum_{j=1}^J \mathrm{off}(V^T A_j V).$$

Then, as: $\min_{V\in\mathcal{O}_K} \sum_{j=1}^J \mathrm{off}(V^T A_j V) = 0$ at the true value, it follows that $\widetilde{V}^T A_j \widetilde{V} = \widetilde{D}_j$ is diagonal for all j. As for all $k \neq m$ there exists $j \in \{1...J\}$ such that $d_{jk} \neq d_{jm}$, one can apply Lemma 18 to show that $\widetilde{V}$ is equal to the true V, up to column sign and permutation. This shows the identification of V. Consistency follows from classical arguments, as the parameter space $\mathcal{O}_K$ is compact.

Asymptotic distribution. A first-order Taylor expansion of (C.35) around the true value V yields:

$$\sum_j^J v_m^T \widehat{A}_j v_k \left(v_k^T \widehat{A}_j v_k - v_m^T \widehat{A}_j v_m \right) + \sum_j^J \left(v_k^T \widehat{A}_j v_k - v_m^T \widehat{A}_j v_m \right) \left(v_m^T \widehat{A}_j (\widehat{v}_k - v_k) + v_k^T \widehat{A}_j (\widehat{v}_m - v_m) \right)$$
$$+ \sum_j^J v_m^T \widehat{A}_j v_k \left(v_k^T \widehat{A}_j (\widehat{v}_k - v_k) - v_m^T \widehat{A}_j (\widehat{v}_m - v_m) \right) = o_p\left(N^{-1/2} \right).$$

As $\operatorname*{plim}_{N\to\infty} \widehat{A}_j = A_j$ for all j, and as $v_k^T A_j v_m = 0$ for all $k \neq m$, this yields:

$$\sum_j^J (d_{jk} - d_{jm})\, v_m^T \left(\widehat{A}_j - A_j \right) v_k + \sum_j^J (d_{jk} - d_{jm}) \left(v_m^T A_j (\widehat{v}_k - v_k) + v_k^T A_j (\widehat{v}_m - v_m) \right) = o_p\left(N^{-1/2} \right),$$

where $d_{jk} = v_k^T A_j v_k$ are the diagonal elements of $V^T A_j V$.

At this stage, it is convenient to define $\widehat{x}_{mk} \equiv v_m^T (\widehat{v}_k - v_k)$. As $v_m^T A_j = d_{jm} v_m^T$, one has:

$$\sum_j^J (d_{jk} - d_{jm})\, v_m^T \left(\widehat{A}_j - A_j \right) v_k + \sum_j^J (d_{jk} - d_{jm}) \left(d_{jm} \widehat{x}_{mk} + d_{jk} \widehat{x}_{km} \right) = o_p\left(N^{-1/2} \right).$$

Now, a Taylor expansion of the orthogonality constraints yields:

$$\widehat{x}_{mk} + \widehat{x}_{km} = v_m^T (\widehat{v}_k - v_k) + v_k^T (\widehat{v}_m - v_m) = 0, \text{ for all } m, k.$$

Thus we have:

$$\sum_j^J (d_{jk} - d_{jm})^2\, \widehat{x}_{mk} = -\sum_j^J (d_{jk} - d_{jm})\, v_m^T \left(\widehat{A}_j - A_j \right) v_k + o_p\left(N^{-1/2} \right). \qquad \text{(C.36)}$$

Let $\widehat{X} = V^T \left(\widehat{V} - V \right)$. Then equation (C.36) is equivalently written, in matrix form, as:

$$\operatorname{vec}\left(\widehat{X} \right) = -W \left(I_J \otimes V^T \otimes V^T \right) \left(\operatorname{vec}\left(\widehat{A} \right) - \operatorname{vec}(A) \right) + o_p\left(N^{-1/2} \right),$$

where W, A and $\widehat{A}$ have been defined in the text. Note that W is provided that $\sum_j^J (d_{jk} - d_{jm})^2 \neq 0$ for all $k \neq m$.

Then, as:

$$\operatorname{vec}\left(\widehat{X} \right) = \left(I_K \otimes V^T \right) \left(\operatorname{vec}\left(\widehat{V} \right) - \operatorname{vec}(V) \right),$$

it follows that

$$N^{\frac{1}{2}} \left(\operatorname{vec}\left(\widehat{V} \right) - \operatorname{vec}(V) \right) = -\left(I_K \otimes V \right) W \left(I_J \otimes V^T \otimes V^T \right) N^{\frac{1}{2}} \left(\operatorname{vec}\left(\widehat{A} \right) - \operatorname{vec}(A) \right) + o_p(1),$$

from which

$$N^{\frac{1}{2}} \left(\operatorname{vec}(\widehat{V}) - \operatorname{vec}(V) \right) \underset{d}{\to} \mathcal{N}\left(0, \mathbb{V}_V \right),$$

where the expression of $\mathbb{V}_V$ is given by (3.36).

C.4 Robin and Smith's (2000) rank test

Let $\widehat{B}$ be a root-N consistent estimator of a (p,q), $p \geq q$, matrix B, such that

$$N^{1/2}\,\mathrm{vec}\left(\widehat{B}-B\right) \xrightarrow{d} \mathcal{N}\left(0, \Sigma_{\mathrm{vec}\left(\widehat{B}\right)}\right),$$

where $\Sigma_{\mathrm{vec}\left(\widehat{B}\right)}$ is definite and $\mathrm{rank}\left(\Sigma_{\mathrm{vec}\left(\widehat{B}\right)}\right) = s$, $0 < s \leq pq$.[1] Let $\widehat{\Sigma}_{\mathrm{vec}\left(\widehat{B}\right)}$ be a consistent estimate of $\Sigma_{\mathrm{vec}\left(\widehat{B}\right)}$. Let $\widehat{B} = \widehat{C}\widehat{D}\widehat{E}^T$ be the singular value decomposition of $\widehat{B}$, where $\widehat{C}$ and $\widehat{E}$ are (p,p) and (q,q) orthogonal matrices and $\widehat{D}$ is a (q,p) diagonal matrix. Let $\widehat{d}_1 \geq ... \geq \widehat{d}_K$ denote the diagonal entries of $\widehat{D}^2$ (the eigenvalues of $\widehat{B}^T\widehat{B}$). For a given null hypothesis: $H_0^r : K = r$, the statistics

$$\mathcal{CRT}_r \equiv N \sum_{i=r+1}^{q} \widehat{d}_i$$

has the same limiting distribution as $\sum_{i=1}^{t} d_i^r Z_i^2$, where $d_1^r \geq ... \geq d_t^r$, $t \leq \min\{s, (p-r)(q-r)\}$, are the non-zero ordered eigenvalues of the matrix

$$(\widehat{E}_{q-r} \otimes \widehat{C}_{p-r})^T \widehat{\Sigma}_{\mathrm{vec}\left(\widehat{B}\right)} (\widehat{E}_{q-r} \otimes \widehat{C}_{p-r}),$$

where $\widehat{E}_{q-r}$ and $\widehat{C}_{p-r}$ are the last $q-r$ and $p-r$ columns of $\widehat{E}$ and $\widehat{C}$, respectively, and $\{Z_i\}_{i=1}^t$ are independent standard normal variates.

To estimate K, we apply the following procedure. Start with $r = 0$. Test H_0^1 against $\widetilde{H}_0^1 : K > 0$. If H_0^1 is rejected, test H_0^2 against $\widetilde{H}_0^2 : K > 1$. And so on until one accepts H_0^r against $\widetilde{H}_0^r : K > r$. The test p-values can be approximated by drawing many independent values of the limiting statistics $\sum_{i=1}^{t} d_i^r Z_i^2$. This procedure delivers a consistent estimate of K if the asymptotic sizes α_N^r used for the sequential tests are such that $\alpha_N^r = o(1)$ and $-N^{-1}\ln\alpha_N^r = o(1)$.

[1] Note that $s < \dim(V)$ because of the symmetry properties of Γ_Y and Ω_Y.

Appendix D

Appendix of Chapter 4

D.1 Proof of Lemma 14

1. First, remark that

$$\mathbb{E}_N f_t - \mathbb{E} f_t = \mathbb{E}_N \operatorname{Re}(f_t) - \mathbb{E}\operatorname{Re}(f_t) + i\left[\mathbb{E}_N \operatorname{Im}(f_t) - \mathbb{E}\operatorname{Im}(f_t)\right]$$

and, for any $T > 0$,

$$\sup_{|t|\leq T} |\mathbb{E}_N f_t - \mathbb{E} f_t| \leq \sup_{|t|\leq T} |\mathbb{E}_N \operatorname{Re}(f_t) - \mathbb{E}\operatorname{Re}(f_t)| + \sup_{|t|\leq T} |\mathbb{E}_N \operatorname{Im}(f_t) - \mathbb{E}\operatorname{Im}(f_t)| .$$

It will thus suffice to show that the proposition is true for the family of functions $\operatorname{Re}(f_t)(x,y) = x\cos(t^T y)$, $t \in \mathbb{R}$, for it to be true for functions $\operatorname{Im}(f_t)$ and f_t. So, without loss of generality, we prove the result for real functions $f_t(x,y) = x\cos(t^T y)$, using the same notation for f_t and its real part. The proof uses the techniques exposed in Chapter II of Pollard (1984).

2. Firstly, the integrability of X allows us to choose a constant K, for any $\varepsilon > 0$, such that $\mathbb{E}[|X|\,\mathbf{1}\{|X| > K\}] \leq \varepsilon$. Then, writing $\mathbb{E}_N f_t$ for the sample mean $\frac{1}{N}\sum_{n=1}^N f_t(X_n, Y_n)$,

$$\begin{aligned}
\sup_{|t|\leq T} |\mathbb{E}_N f_t - \mathbb{E} f_t| &\leq \sup_{|t|\leq T} |\mathbb{E}_N\left[f_t \mathbf{1}\{|X| \leq K\}\right] - \mathbb{E}[f_t \mathbf{1}\{|X| \leq K\}]| \\
&\quad + \sup_{|t|\leq T} \mathbb{E}_N\left[|f_t|\,\mathbf{1}\{|X| > K\}\right] + \sup_{|t|\leq T} \mathbb{E}[|f_t|\,\mathbf{1}\{|X| > K\}] \\
&\leq \sup_{|t|\leq T} |\mathbb{E}_N\left[f_t \mathbf{1}\{|X| \leq K\}\right] - \mathbb{E}[f_t \mathbf{1}\{|X| \leq K\}]| \\
&\quad + \mathbb{E}_N\left[|X|\,\mathbf{1}\{|X| > K\}\right] + \mathbb{E}[|X|\,\mathbf{1}\{|X| > K\}] .
\end{aligned}$$

The last two terms converge almost surely to $2\mathbb{E}[|X|\,\mathbf{1}\{|X| > K\}]$, which is less than 2ε.

3. From now on, one may as well consider that the support of X is absolutely bounded by K (i.e. $|X| \leq K$ almost surely). Let $\mathbf{Z}_N = (Z_1, ..., Z_N)$ be an i.i.d. sample of random variables with distribution F. The two symmetrisation steps of the proof of the Glivenko-Cantelli Theorem provide a first bound. The first symmetrisation step replaces $F_N - F$

by $F_N - F'_N$, where F'_N (resp. $\mathbb{E}'_N$) is the empirical distribution (resp. the empirical mean operator) of another *i.i.d.* sample $\mathbf{Z}'_N = (Z'_1, ..., Z'_N)$ of random variables with distribution F, independent of $\mathbf{Z}_N$. Specifically, the symmetrisation lemma in section 3 of chapter II of Pollard (1984) shows that

$$\Pr\left\{\sup_{|t|\leq T} |\mathbb{E}_N f_t - \mathbb{E}f_t| \geq \varepsilon\right\} \leq 2\Pr\left\{\sup_{|t|\leq T} \left|\mathbb{E}_N f_t - \mathbb{E}'_N f_t\right| \geq \frac{1}{2}\varepsilon\right\}, \tag{D.11}$$

if $\Pr\left\{|\mathbb{E}_N f_t - \mathbb{E}f_t| \leq \frac{1}{2}\varepsilon\right\} \geq \frac{1}{2}$ for all $|t| \leq T$. Chebyshev inequality shows that the latter inequality holds whenever $N \geq \frac{8\,\mathrm{Var}\, f_t(X_n,Y_n)}{\varepsilon^2}$. As

$$\mathrm{Var}\, f_t(X_n, Y_n) = \mathrm{Var}\left[X_n \cos\left(t^T Y_N\right)\right] \leq \mathbb{E}X_n^2 \leq M_1,$$

inequality (D.11) is true for $N \geq \frac{8M_1}{\varepsilon^2}$.

4. The second symmetrisation step uses an i.i.d. sample of Rademacher random variables $\boldsymbol{\sigma}_N = (\sigma_1, ..., \sigma_N)$, where $\sigma_n = 1$ or -1 with the same probability $\frac{1}{2}$, $n = 1, ..., N$, independent of $\mathbf{Z}_N$ and $\mathbf{Z}'_N$. The sequence of random variables $\sigma_n\left[f_t(Z_n) - f_t(Z'_n)\right]$ then has the same joint distribution as the original sequence $f_t(Z_n) - f_t(Z'_n)$. It follows that

$$\begin{aligned}
\Pr\left\{\sup_{|t|\leq T} \left|\mathbb{E}_N f_t - \mathbb{E}'_N f_t\right| \geq \frac{1}{2}\varepsilon\right\} &= \Pr\left\{\sup_{|t|\leq T} \left|\frac{1}{N}\sum_{n=1}^N \sigma_n f_t(Z_n) - \frac{1}{N}\sum_{n=1}^N \sigma_n f_t(Z'_n)\right| \geq \frac{1}{2}\varepsilon\right\}, \\
&\leq \Pr\left\{\sup_{|t|\leq T} \left|\frac{1}{N}\sum_{n=1}^N \sigma_n f_t(Z_n)\right| \geq \frac{1}{4}\varepsilon\right\} \\
&\quad + \Pr\left\{\sup_{|t|\leq T} \left|\frac{1}{N}\sum_{n=1}^N \sigma_n f_t(Z'_n)\right| \geq \frac{1}{4}\varepsilon\right\}, \\
&= 2\Pr\left\{\sup_{|t|\leq T} \left|\frac{1}{N}\sum_{n=1}^N \sigma_n f_t(Z_n)\right| \geq \frac{1}{4}\varepsilon\right\}, \\
&= 2\mathbb{E}\Pr\left\{\sup_{|t|\leq T} \left|\frac{1}{N}\sum_{n=1}^N \sigma_n f_t(Z_n)\right| \geq \frac{1}{4}\varepsilon \,\middle|\, \mathbf{Z}_N\right\}.
\end{aligned} \tag{D.12}$$

5. A maximal inequality then follows from the following finite-covering argument. For any couple (t_1, t_2),

$$\begin{aligned}
\left|x\cos(t_1^T y) - x\cos(t_2^T y)\right| &\leq \left|x(t_1^T y - t_2^T y)\right| \\
&\leq \sum_\ell |x y_\ell\,(t_{1\ell} - t_{2\ell})| \\
&\leq \sum_\ell |x y_\ell| \cdot |t_1 - t_2| \\
&\leq L\,|x|\,|y| \cdot |t_1 - t_2|\,.
\end{aligned}$$

Fix $\mathbf{Z}_N$, and define $M_{2,N} = \frac{1}{N}\sum_n |Y_n|$. Partition $[-T,T]^L$ into $r_N = \left(2T\frac{8LKM_{2,N}}{\varepsilon}\right)^L$ adjacent hypercubes of side length $\frac{\varepsilon}{8KM_{2,N}}$. Lastly, let $\{t_k; k=1,...,r_N\}$ be the set of all cube corners. Then, for any $t \in [-T,T]^L$ there exists k such that

$$\begin{aligned}\left|\frac{1}{N}\sum_{n=1}^{N}\sigma_n\left(f_{t_k}-f_t\right)(Z_n)\right| &\leq \frac{1}{N}\sum_{n=1}^{N}\left|\left(f_{t_k}-f_t\right)(Z_n)\right| \\ &\leq LKM_{2,N}\left|t_k-t\right| \\ &\leq LKM_{2,N}\frac{\varepsilon}{8LKM_{2,N}} = \frac{1}{8}\varepsilon,\end{aligned}$$

as $|X_n| \leq K$. Hence, for all $t \in [-T,T]^L$ such that $\left|\frac{1}{N}\sum_{n=1}^{N}\sigma_n f_t(Z_n)\right| \geq \frac{1}{4}\varepsilon$,

$$\begin{aligned}\left|\frac{1}{N}\sum_{n=1}^{N}\sigma_n f_{t_k}(Z_n)\right| &\geq \left|\frac{1}{N}\sum_{n=1}^{N}\sigma_n f_t(Z_n)\right| - \left|\frac{1}{N}\sum_{n=1}^{N}\sigma_n\left(f_{t_k}-f_t\right)(Z_n)\right|, \\ &\geq \frac{1}{8}\varepsilon.\end{aligned}$$

One can thus refine the bound further as:

$$\begin{aligned}\Pr\left\{\sup_{|t|\leq T}\left|\frac{1}{N}\sum_{n=1}^{N}\sigma_n f_t(Z_n)\right| \geq \frac{1}{4}\varepsilon\,\middle|\,\mathbf{Z}_N\right\} &\leq \Pr\left\{\max_k\left|\frac{1}{N}\sum_{n=1}^{N}\sigma_n f_{t_k}(Z_n)\right| \geq \frac{1}{8}\varepsilon\,\middle|\,\mathbf{Z}_N\right\} \\ &\leq \sum_{k=1}^{r_N}\Pr\left\{\left|\frac{1}{N}\sum_{n=1}^{N}\sigma_n f_{t_k}(Z_n)\right| \geq \frac{1}{8}\varepsilon\,\middle|\,\mathbf{Z}_N\right\}, \\ &\leq r_N \max_k \Pr\left\{\left|\frac{1}{N}\sum_{n=1}^{N}\sigma_n f_{t_k}(Z_n)\right| \geq \frac{1}{8}\varepsilon\,\middle|\,\mathbf{Z}_N\right\}. \end{aligned} \tag{D.13}$$

6. Lastly, applying Hoeffding's inequality to the sequence $[\sigma_n f_{t_k}(Z_n)]$, bounded by K, yields:

$$\Pr\left\{\sup_k\left|\frac{1}{N}\sum_{n=1}^{N}\sigma_n f_{t_k}(Z_n)\right| \geq \frac{1}{8}\varepsilon\,\middle|\,\mathbf{Z}_N\right\} \leq 2\exp\left[-\frac{N\varepsilon^2}{128K^2}\right]. \tag{D.14}$$

7. At this stage, we have thus shown that, for $N \geq \frac{8M_1}{\varepsilon^2}$,

$$\begin{aligned}\Pr\left\{\sup_{|t|\leq T}\left|\mathbb{E}_N f_t - \mathbb{E}f_t\right| \geq \varepsilon\right\} &\leq 2^{4L+3}\mathbb{E}\left(M_{2,N}^L\right)\left(\frac{LKT}{\varepsilon}\right)^L \exp\left[-\frac{N\varepsilon^2}{128K^2}\right], \\ &= 2^{4L+3}L^L\mathbb{E}\left(M_{2,N}^L\right)\exp\left[L\ln\left(\frac{KT}{\varepsilon}\right) - \frac{N\varepsilon^2}{128K^2}\right].\end{aligned}$$

Assuming that $\mathbb{E}|Y_N|^i < +\infty$ for all $i \leq \{1...L\}$, then one can find $M_2 < +\infty$ such that $\mathbb{E}|Y_N|^i < M_2^i$, $\forall i \leq L$. Since (Y_N) is an i.i.d. sequence we thus have

$$\begin{aligned}\mathbb{E}\left(M_{2,N}^L\right) &= \mathbb{E}\left[\left(\frac{1}{N}\sum_{n=1}^{N}|Y_n|\right)^L\right] \\ &\leq (M_2)^L .\end{aligned}$$

Let (ε_N) be a sequence of positive numbers converging to zero and let (T_N) be a diverging sequence of positive numbers. If ε_N tends to zero slowly enough for $\varepsilon_N^2 \geq \frac{8M_1}{N}$ and so that $\sum_N \exp\left[L\ln\left(\frac{K_N T_N}{\varepsilon_N}\right) - \frac{N\varepsilon_N^2}{128K_N^2}\right] < \infty$, then

$$\sum_N \Pr\left\{\sup_{|t|\leq T_N} |\mathbb{E}_N f_t - \mathbb{E}f_t| \geq \varepsilon_N\right\} < \infty.$$

The Borel-Cantelli Lemma then implies that only a finite number of events are such that $\sup_{|t|\leq T_N} |\mathbb{E}_N f_t - \mathbb{E}f_t| \geq \varepsilon_N$. Hence,

$$\sup_{|t|\leq T_N} |\mathbb{E}_N f_t - \mathbb{E}f_t| = O(\varepsilon_N), \qquad \text{a.s..}$$

8. The last step of the proof characterises T_N, K_N and ε_N further. Let $K_{|X|}(\varepsilon)$ be implicitly defined by the equality:

$$\mathbb{E}[|X|\,\mathbf{1}\{|X| > K\}] = \int_K^\infty u f_{|X|}(u)\,du = \varepsilon.$$

Let $M > 0$ and $k(\varepsilon)$ positive and decreasing such that

$$\mathbb{E}[|X|\,\mathbf{1}\{|X| > Mk(\varepsilon)\}] \leq \varepsilon.$$

It follows from the monotonicity of $\mathbb{E}[|X|\,\mathbf{1}\{|X| > x\}]$ as a function of x that

$$K_{|X|}(\varepsilon) \leq Mk(\varepsilon).$$

Let now $0 < \gamma < 1/2$ and $r > 0$, and let ε_N, T_N such that

$$\frac{\varepsilon_N}{k(\varepsilon_N)} = \left(\frac{N}{\ln N}\right)^{\gamma-\frac{1}{2}}, \quad T_N = O(N^r).$$

Then $\varepsilon_N^2 \geq \frac{8M_1}{N}$ for N large enough. Moreover:

$$\begin{aligned}L\ln\left(\frac{K_N T_N}{\varepsilon_N}\right) - \frac{N\varepsilon_N^2}{128K_N^2} &\leq L\ln\left(\frac{Mk(\varepsilon_N)T_N}{\varepsilon_N}\right) - \frac{N\varepsilon_N^2}{128M^2 k(\varepsilon_N)^2}, \\ &\leq L\ln\left(M\left(\frac{\ln N}{N}\right)^{\gamma-\frac{1}{2}} O(N^r)\right) - \frac{1}{128M^2}\frac{N^{2\gamma}}{(\ln N)^{2\gamma-1}}, \\ &\leq -N^\gamma + O(\ln N),\end{aligned}$$

for N large enough. It follows that the series $\sum_N \exp\left[L\ln\left(\frac{K_N T_N}{\varepsilon_N}\right) - \frac{N\varepsilon_N^2}{128K_N^2}\right]$ converges.

This achieves to prove Lemma 14.

D.2 Proof of Theorem 15

(i) Fix any $t \in \mathbb{R}^L$, let $\varphi(t) \equiv \varphi_Y(t) = \mathbb{E}\left[e^{it^TY}\right]$, $\psi_\ell(t) = \mathbb{E}\left[Y_\ell e^{it^TY}\right]$ and $\xi_{\ell m}(t) = \mathbb{E}\left[Y_\ell Y_m e^{it^TY}\right]$, for any $\ell, m = 1, ..., L$. Then, Lemma 14 defines $\varepsilon_N \downarrow 0$ and $T_N \to \infty$ such that (all convergence statements are implicitly holding almost surely)

$$\sup_{|t|\leq T_N} |\widehat{\varphi}(t) - \varphi(t)| = O(\varepsilon_N),$$

$$\sup_{|t|\leq T_N} \left|\widehat{\psi}_\ell(t) - \psi_\ell(t)\right| = O(\varepsilon_N),$$

$$\sup_{|t|\leq T_N} \left|\widehat{\xi}_{\ell m}(t) - \xi_{\ell m}(t)\right| = O(\varepsilon_N),$$

hold simultaneously. One can take the largest ε_N and the smallest T_N.

In addition, we shall require below that $\frac{T_N^2 \varepsilon_N}{g(T_N)^3} = o(1)$. As $h(t) = \frac{t^2}{g(t)^3}$ is an increasing function, one can redefine T_N –if necessary– as $T_N = h^{-1}\left(\varepsilon_N^{\gamma-1}\right)$, with $0 < \gamma < 1$.

(ii) Removing the subscript Y from φ_Y and g_Y to simplify the notations, as $|\varphi(t)| \geq g(|t|)$ when $|t| \to \infty$, and as φ is nonvanishing everywhere, then for T_N large enough

$$\inf_{|t|\leq T_N} |\varphi(t)| \geq g(T_N),$$

and

$$\sup_{|t|\leq T_N} \left|\frac{\widehat{\varphi}(t) - \varphi(t)}{\varphi(t)}\right| = \frac{O(\varepsilon_N)}{g(T_N)} = o(1).$$

The last equality follows from the fact that $\frac{T_N^2 \varepsilon_N}{g(T_N)^3} \geq \frac{\varepsilon_N}{g(T_N)}$ for N large enough.

(iii) We have

$$\frac{\partial \kappa_Y(t)}{\partial t_\ell} = i\frac{\psi_\ell(t)}{\varphi(t)} = i\frac{\mathbb{E}\left[Y_\ell e^{it^TY}\right]}{\mathbb{E}\left[e^{it^TY}\right]},$$

and

$$\begin{aligned}
\frac{\widehat{\psi}_\ell(t)}{\widehat{\varphi}(t)} - \frac{\psi_\ell(t)}{\varphi(t)} &= \frac{\widehat{\psi}_\ell(t)}{\widehat{\varphi}(t)} - \frac{\widehat{\psi}_\ell(t)}{\varphi(t)} + \frac{\widehat{\psi}_\ell(t)}{\varphi(t)} - \frac{\psi_\ell(t)}{\varphi(t)} \\
&= -\frac{\widehat{\psi}_\ell(t)}{\varphi(t)} \frac{\frac{\widehat{\varphi}(t)-\varphi(t)}{\varphi(t)}}{\frac{\widehat{\varphi}(t)-\varphi(t)}{\varphi(t)} + 1} + \frac{1}{\varphi(t)}\left[\widehat{\psi}_\ell(t) - \psi_\ell(t)\right].
\end{aligned}$$

One can bound $\widehat{\psi}_\ell(t)$ as follows:

$$\begin{aligned}
\sup_{|t|\leq T_N} \left|\widehat{\psi}_\ell(t)\right| &\leq \sup_{|t|\leq T_N} \left|\widehat{\psi}_\ell(t) - \psi_\ell(t)\right| + \sup_{t\in[-T_N,T_N]} |\psi_\ell(t)| \\
&\leq \sup_{|t|\leq T_N} \left|\widehat{\psi}_\ell(t) - \psi_\ell(t)\right| + \mathbb{E}|Y_\ell| = O(1),
\end{aligned}$$

as $\mathbb{E}|Y_\ell| < \infty$ if $\mathbb{E}Y_\ell^2 \leq M_1 < \infty$.

It follows that

$$\sup_{|t|\leq T_N}\left|\frac{\widehat{\psi}_\ell(t)}{\widehat{\varphi}(t)} - \frac{\psi_\ell(t)}{\varphi(t)}\right| = \frac{O(\varepsilon_N)}{g(T_N)^2} = o\,(1)\,.$$

The same argument applies to show that

$$\sup_{|t|\leq T_N}\left|\frac{\widehat{\xi}_{\ell m}(t)}{\widehat{\varphi}(t)} - \frac{\xi_{\ell m}(t)}{\varphi(t)}\right| = \frac{O(\varepsilon_N)}{g(T_N)^2} = o\,(1)$$

for all ℓ, m.

(iv) It is easy to extend these results to second derivatives of cumulant generating functions:

$$\begin{aligned}\zeta_{\ell m}(t) &\equiv \frac{\partial^2\kappa_Y}{\partial t_\ell \partial t_m}(t)\\ &= -\frac{\mathbb{E}\left[Y_\ell Y_m e^{it^T Y}\right]}{\mathbb{E}\left[e^{it^T Y}\right]} + \frac{\mathbb{E}\left[Y_\ell e^{it^T Y}\right]}{\mathbb{E}\left[e^{it^T Y}\right]}\frac{\mathbb{E}\left[Y_m e^{it^T Y}\right]}{\mathbb{E}\left[e^{it^T Y}\right]}\\ &= -\frac{\xi_{\ell m}(t)}{\varphi(t)} + \frac{\psi_\ell(t)}{\varphi(t)}\frac{\psi_m(t)}{\varphi(t)}.\end{aligned}$$

Let $\widehat{\zeta}_{\ell m}(t) = -\frac{\widehat{\xi}_{\ell m}(t)}{\widehat{\varphi}(t)} + \frac{\widehat{\psi}_\ell(t)}{\widehat{\varphi}(t)}\frac{\widehat{\psi}_m(t)}{\widehat{\varphi}(t)}$. Then,

$$\begin{aligned}\widehat{\zeta}_{\ell m}(t) - \zeta_{\ell m}(t) &= -\left[\frac{\widehat{\xi}_{\ell m}(t)}{\widehat{\varphi}(t)} - \frac{\xi_{\ell m}(t)}{\varphi(t)}\right]\\ &\quad + \left[\frac{\widehat{\psi}_\ell(t)}{\widehat{\varphi}(t)} - \frac{\psi_\ell(t)}{\varphi(t)}\right]\frac{\psi_m(t)}{\varphi(t)} + \left[\frac{\widehat{\psi}_m(t)}{\widehat{\varphi}(t)} - \frac{\psi_m(t)}{\varphi(t)}\right]\frac{\psi_\ell(t)}{\varphi(t)}\\ &\quad + \left[\frac{\widehat{\psi}_\ell(t)}{\widehat{\varphi}(t)} - \frac{\psi_\ell(t)}{\varphi(t)}\right]\left[\frac{\widehat{\psi}_m(t)}{\widehat{\varphi}(t)} - \frac{\psi_m(t)}{\varphi(t)}\right].\end{aligned}$$

Since

$$\sup_{|t|\leq T_N}\left|\frac{\psi_\ell(t)}{\varphi(t)}\right| \leq \frac{\mathbb{E}|Y_\ell|}{g(T_N)}$$

for all ℓ, it follows that

$$\sup_{|t|\leq T_N}\left|\widehat{\zeta}_{\ell m}(t) - \zeta_{\ell m}(t)\right| = \frac{O(\varepsilon_N)}{g(T_N)^2} + \frac{O(\varepsilon_N)}{g(T_N)^3} + \left(\frac{O(\varepsilon_N)}{g(T_N)^2}\right)^2 = \frac{O(\varepsilon_N)}{g(T_N)^3}$$

because

$$\frac{\varepsilon_N}{g(T_N)^3} > \frac{\varepsilon_N^2}{g(T_N)^4} \Leftrightarrow 1 > \frac{\varepsilon_N}{g(T_N)}$$

for N large enough.

(v) For any vector $t = (t_1, ..., t_L)^T \in \mathbb{R}^L$ and $\tau \in \mathbb{R}$, then

$$\begin{aligned}
B_\ell(t) &= \sup_{\tau \in [-T_N, T_N]} \left| \int_0^\tau \frac{\widehat{\psi}_\ell(ut)}{\widehat{\varphi}(ut)} du - \int_0^{t_\ell} \frac{\psi_\ell(ut)}{\varphi(ut)} du \right| \\
&\leq \sup_{\tau \in [-T_N, T_N]} \left(\tau \sup_{|t| \leq T_N} \left| \frac{\widehat{\psi}_\ell(t)}{\widehat{\varphi}(t)} - \frac{\psi_\ell(t)}{\varphi(t)} \right| \right) \\
&\leq T_N \sup_{|t| \leq T_N} \left| \frac{\widehat{\psi}_\ell(t)}{\widehat{\varphi}(t)} - \frac{\psi_\ell(t)}{\varphi(t)} \right| \\
&= \frac{T_N}{g(T_N)^2} O(\varepsilon_N).
\end{aligned}$$

Similarly,

$$\begin{aligned}
C_{\ell m}(t) &= \sup_{\tau \in [-T_N, T_N]} \left| \int_0^\tau \int_0^u \widehat{\zeta}_{\ell m}(vt) dv du - \int_0^\tau \int_0^u \zeta_{\ell m}(vt) dv du \right| \\
&\leq \sup_{\tau \in [-T_N, T_N]} \left(\frac{\tau^2}{2} \sup_{|t| \leq T_N} \left| \widehat{\zeta}_{\ell m}(t) - \zeta_{\ell m}(t) \right| \right) \\
&\leq T_N^2 \sup_{|t| \leq T_N} \left| \widehat{\zeta}_{\ell m}(t) - \zeta_{\ell m}(t) \right| \\
&= \frac{T_N^2}{g(T_N)^3} O(\varepsilon_N).
\end{aligned}$$

Moreover, for any distribution W on $\mathcal{T}_k$,

$$\int B_\ell(t)\, dW(t) \leq \sup_{|t| \leq T_N} B_\ell(t) \cdot \int dW(t) = \frac{T_N}{g(T_N)^2} O(\varepsilon_N)$$

and

$$\int C_{\ell m}(t)\, dW(t) \leq \sup_{|t| \leq T_N} C_{\ell m}(t) \cdot \int dW(t) = \frac{T_N^2}{g(T_N)^3} O(\varepsilon_N).$$

(vi) It easily follows from the previous step that:

$$\sup_{\tau \in [-T_N, T_N]} \left| \widehat{\kappa}_{X_k}(\tau) - \kappa_{X_k}(\tau) \right| = \frac{T_N^2}{g(T_N)^3} O(\varepsilon_N) = o(1).$$

In particular, $\sup_{\tau \in [-T_N, T_N]} \left| \widehat{\kappa}_{X_k}(\tau) - \kappa_{X_k}(\tau) \right| < 1$ for N large enough. Therefore, for N large enough

$$\begin{aligned}
\sup_{\tau \in [-T_N, T_N]} \left| \widehat{\varphi}_{X_k}(\tau) - \varphi_{X_k}(\tau) \right| &= \sup_{\tau \in [-T_N, T_N]} \left| \exp\left(\widehat{\kappa}_{X_k}(\tau) \right) - \exp\left(\widehat{\kappa}_{X_k}(\tau) \right) \right|, \\
&\leq \sup_{\tau \in [-T_N, T_N]} \left| \widehat{\kappa}_{X_k}(\tau) - \kappa_{X_k}(\tau) \right|,
\end{aligned}$$

from which it follows that

$$\sup_{\tau\in[-T_N,T_N]}\left|\widehat{\varphi}_{X_k}(\tau)-\varphi_{X_k}(\tau)\right|=\frac{T_N^2}{g(T_N)^3}O(\varepsilon_N).$$

This ends the proof of Theorem 15.

D.3 Proof of Theorem 16

For all x in the support of X_k:

$$\begin{aligned}\left|\widehat{f}_{X_k}(x)-f_{X_k}(x)\right| &\leq \frac{1}{2\pi}\left(\int_{-T_N}^{T_N}\left|\widehat{\varphi}_{X_k}(v)-\varphi_{X_k}(v)\right|dv+\int_{-\infty}^{-T_N}\left|\varphi_{X_k}(v)\right|dv+\int_{T_N}^{+\infty}\left|\varphi_{X_k}(v)\right|dv\right)\\ &\leq \frac{T_N}{\pi}\sup_{|\tau|\leq T_N}\left|\widehat{\varphi}_{X_k}(\tau)-\varphi_{X_k}(\tau)\right|+\frac{1}{\pi}\int_{T_N}^{+\infty}h_{X_k}(v)dv.\end{aligned}$$

Hence, Theorem 15 implies:

$$\sup_x\left|\widehat{f}_{X_k}(x)-f_{X_k}(x)\right|=\frac{T_N^3}{g(T_N)^3}O(\varepsilon_N)+O\left(\int_{T_N}^{+\infty}h_{X_k}(v)dv\right).$$

where $g(|t|)=g_X(L|A||t|)$ and $\int_{T_N}^{+\infty}h_X(v)dv=o(T_N)$, as h_{X_k} is integrable. This ends the proof of Theorem 16.

Bibliography

[1] ABOWD, J. and D. CARD (1989), "On the Covariance Structure of Earnings and Hours Changes," *Econometrica*, 57, 411-445.

[2] AIGNER, D. J., C. HSIAO, A. KAPTEYN, and T. WANSBEEK (1984): "Latent Variable Models in Econometrics," in *Handbook of Econometrics*, Vol. II, ed. by Z. Griliches and M. D. Intriligator. Amsterdam: North Holland.

[3] ATTIAS, H. (1999), "Independent Factor Analysis," *Neural Computation*, ;11:803-851.

[4] AKAIKE, H. (1969): "Fitting Autoregressive Models for Prediction," *Annals of the Institute of Statistical Mathematics*, 21, 243-247.

[5] ALTONJI, J.G. and L.M. SEGAL (1994): "Small Sample Bias in GMM Estimation of Covariance Structures", NBER Technical Working Paper no. 156, National Bureau of Economic Research, Cambridge, MA.

[6] ALTONJI, J.G. and L.M. SEGAL (1996): "Small Sample Bias in GMM Estimation of Covariance Structures", *Journal of Business and Economic Statistics*, 14, 353-366.

[7] ALVAREZ, J., M. BROWNING, and M. EJRNAES (2001): "modelling Income Processes with Lots of Heterogeneity," *Mimeo*.

[8] ANDERSON, T. W. (1963): "Asymptotic Theory for Principal Component Analysis,," *Ann. Math. Stat.,* 34, 122-148.

[9] ANDERSON, T. W. (1984): "An Introduction to Multivariate Statistical Analysis, New York: Wiley.

[10] ARCIDIACONO, A., and J. JONES (2003): "Finite Mixture Distributions, Sequential Likelihood and the EM Algorithm," *Econometrica*, 71, 933-946.

[11] ASHENFELTER, O. and J. MOONEY, "Graduate Education, Ability and Earnings," *Review of Economics and Statistics*, 50 (February 1968): 78-86.

[12] ATTANASIO, O., and S. DAVIS (1996): "Relative Wage Movements and the Distribution of Consumption," *Journal of Political Economy*, 104(6) , 1227-1262.

[13] BACK, A. D. and A. S. WEIGEND (1997), "A First Application of Independent Component Analysis to Extracting Structure from Stock Returns," *International Journal of Neural Systems*, Vol. 8, No.4, 473-484.

[14] BAI, J. (2003): "Inferential Theory for Factor Models of Large Dimensions," *Econometrica,* 71, 135-171.

[15] BAI, J., and S. NG (2002): "Determining the Number of Factors in Approximate Factor Models ," *Econometrica,* 70, 191-221.

[16] BARTHOLOMEW, D. J. (1970): *Stochastic Models for Social Processes.* New York: Wiley.

[17] BAYER, P., N. KEOHANE, and C. TIMMINS (2005): "Migration and Hedonic Valuation: The Case of Air Quality," *mimeo.*

[18] BLANCHARD, O. J., and D. QUAH (1989): "The Dynamic Effects of Aggregate Demand and Supply Disturbances," *American Economic Review*, 79, 655-673.

[19] BLUMEN, I., M. KOGAN, and P. J. McCARTHY (1955): *The Industrial Mobility of labour as a Probability Process.* Ithaca: Cornell University Press.

[20] BLUNDELL, R., and I. PRESTON (1998): "Consumption Inequality and Income Uncertainty," *Quarterly Journal of Economics* 113 (2), 603-640.

[21] BOWLUS, A., and J.M. ROBIN (2004), "Twenty years of Rising Inequality in US Lifetime labour Income Values," *Review of Economic Studies*, vol. 71, 3, pp. 709-742..

[22] BROWN, C. (1980): "Equalizing Differences in the Labor Market," *Quarterly Journal of Economics*, 94, 113-134.

[23] BUCHINSKY, M., D. FOUGÈRE, and F. KRAMARZ (1998): "La Mobilité Salariale en France : 1967-1987," *Revue Economique*, 49, 879-90.

[24] BUCHINSKY, M., and J. HUNT (1999): "Wage Mobility in the United States," *Review of Economics and Statistics*, 351-68.

[25] CARD D. (2001): "Estimating the Return to Schooling: Progess on Some Persistent Econometric Problems," *Econometrica,* 69(5), 1127-1160.

[26] CARDOSO J.-F. (1998): "Blind signal separation : statistical principles," *Proc. IEEE,* 9(10), 2009-2025.

[27] CARDOSO J.-F. (1999): "High-order contrasts for independent Component Analysis," *Neural Computation,* 11, 157-192.

[28] CARDOSO J.-F., and A. SOULOUMIAC (1993): "Blind Beamforming for Non-Gausiian Signals," *IEE-Proceedings-F,* 140, 362-370.

[29] CARDOSO J.-F., and A. SOULOUMIAC (1996): "Jacobi Angles for Simultaneous Diagonalization," *SIAM J. Mat. An. Appl.,* 17, 161-164.

[30] CARNEIRO, P., K. T. HANSEN, and J. J. HECKMAN (2002): "Estimating Distributions of Treatment Effects with an Application to the Returns to Schooling and Measurement of the Effects of Uncertainty on College Choice," *International Economic Review*, 44(2), 361-422.

[31] CARROLL, R. J. , and P. HALL (1988): "Optimal rates of Convergence for Deconvoluting a Density," *Journal of the American Statistical Association*, 83, 1184-1186.

[32] CHAMBERLAIN, G. , and Z. GRILICHES (1975): "Unobservables with a Variance-Components Structure: Ability, Schooling and the Economic Success of Brothers," *International Economic Review*, 16(2), 422-444.

[33] CHAMBERLAIN, G. , and Z. GRILICHES (1977): "More on Brothers," in *Kinometrics: Determinants of Socioeconomic Success Within and Between Families*, edited by Paul Taubman, New York: North Holland.

[34] CHRISTENSEN, B. J., R. LENTZ, D. T. MORTENSEN, G. R. NEUMANN and A. WERWATZ (2005) "On the Job Search and the Wage Distribution," Forthcoming *Journal of Labor Economics.*

[35] CLARK, A. E., and A. J. OSWALD (1994): "Unhappiness and Unemployment," *Economic Journal*, 104, 648-659.

[36] CLARK, A. E., and A. J. OSWALD (1996): "Satisfaction and Comparison Income," *Journal of Public Economics*, 61, 359-381.

[37] COHEN, D. (1998): "Welfare Differentials across French and US labour Markets," *mimeo.*

[38] COMON, P. (1994): "Independent Component Analysis, a New Concept?," *Signal Processing*, 36(3), 287-314.

[39] COX, D. R. (1975): "Partial Likelihood," *Biometrika*, 62(2), 269-276.

[40] CRAGG, J. G. (1997): "Using Higher Moments to Estimate the Simple Errors-in-Variables Model," *RAND Journal of Economics*, 28, S71-S91.

[41] CUNHA, F., J.J. HECKMAN and S. NAVARRO (2005), "Seperating uncertainty from heterogeneity in life cycle earnings," *Oxford Economic Papers,* 57, 191-261.

[42] DAGENAIS, M. G., and D. L. DAGENAIS (1997): "Higher Moment Estimators for Linear Regression Models with Errors in Variables," *Journal of Econometrics*, 76, 193-221.

[43] DALE-OLSEN, H. (2005): "Estimating Workers' Marginal Willingness to Pay for Safety Using Linked Employer-Employee Data," *Economica*, forthcoming.

[44] DANIEL, C. and C. SOFER (1998): "Bargaining, Compensating Wage Differentials and Dualism of the Labor Market. Theory and Evidence for France," *Journal of Labor Economics*, 16(3), 546-75

[45] DARDANONI, V. (1993): "Measuring Social Mobility," *Journal of Economic Theory*, 61, 372-394.

[46] DARTON, R. A. (1980): "Rotation in Factor Analysis," *The Statistician*, 19, 167-194.

[47] DE LATHAUWER, L. (2003): "Simultaneous Matrix Diagonalization: the Overcomplete Case," *Proc. of the 4th International Symposium on ICA and Blind Signal Separation, Nara, Japan*, 812-825.

[48] DEMPSTER, A., N. LAIRD, and D. RUBIN (1977): "Maximum Likelihood from Incomplete Data via the EM Algorithm," *J. Royal Statist. Soc. Ser. B.*, 39, 1-38.

[49] DEY, M., and C.J. FLINN (2005): "An Equilibrium Model of Health Insurance Provision and Wage Determination," *Econometrica*, 73(2), 571-628.

[50] DICKENS, R. (2000): "Caught in a Trap? Wage Mobility in Great Britain : 1975-1994," *Economica*, 67, 477-497.

[51] DIGGLE, P. J., and P. HALL (1993): "A Fourier Approach to Nonparametric Deconvolution of a Density Estimate," *Journal of the Royal Statistical Society Series B*, 55, 523-531.

[52] DOLTON, P.J., G. MAKEPEACE, and W. VAN DER KLAAUW (1989): "Occupational Choice and Earnings Determination: The Role of Sample Selection and Non-pecuniary Factors," *Oxford Economic Papers*, 41(3), 573-594.

[53] DOZ, C. and E. RENAULT (2005): "Factor Stochastic Volatility in Mean Models: a GMM approach," mimeo.

[54] DROUET-MARI, D., and S. KOTZ (2001): *Correlation and Dependence.* London: Imperial College.

[55] DUFFIE, D. and J. PAN (1997): "An Overview of Value at Risk," *Journal of Derivatives*, 7-49, reprinted in *Options Markets*, edited by G. Constantinides and A. G. Malliaris, London: Edward Elgar, 2001.

[56] DUGUĒ, D. (1951), "Analycité et convexité des fonctions caractéristiques," *Annales de l'Institut Henri Poincaré*, Vol. XII, 45-46.

[57] DUNCAN, O. D. (1966): "Methodological Issues in the Study of Social Mobility," in *Social Structure and Mobility in Economic Developpement*, ed. by N.J. Smelser and S.M. Lipset. Chicago: Aldine.

[58] DUNCAN, G., and B. HOLMLUND (1983): "Was Adam Smith Right After All? Another Test of the Theory of Compensating Differentials," *Journal of Labor Economics*, 1, 366-379.

[59] DUSTMANN C., and C. MEGHIR (2005): "Wages, Experience and Seniority," *Review of Economic Studies*, 72(1), 77-108.

[60] EKELAND I., J. HECKMAN, and L. NESHEIM (2002): "Identifying hedonic models," *American Economic Review*, 92(2), 304-309.

[61] EKELAND I., J. HECKMAN, and L. NESHEIM (2004): "Identification and estimation of hedonic models," *Journal of Political Economy*, 112, S60-S109.

[62] ERICKSON, T., and T. WHITED (2002): "Two-Step GMM Estimation of the Error-in-Variables Model Using High-Order Moments," *Econometric Theory*, 18, 776-799.

[63] ERIKSSON J. and V. KOIVUNEN (2003): "Identifiability and separability of linear ICA models revisited," *4th International Symposium on ICA and Blind Signal Separation*, 23-27.

[64] FAN, J. Q. (1991): "On the Optimal Rates of Convergence for Nonparametric Deconvolution Problems," *Annals of statistics*, 19, 1257-1272.

[65] FAN, J. and J.-Y. KOO (2002): "Wavelet Deconvolution," *IEEE transactions on Information Theory*, Vol. 48, 3, pp. 734-747.

[66] FLURY, B. (1984): "Common Principal Components in K Groups," *J. Am. Stat. Ass.*, 79, 892-898.

[67] FIELDS, G. and E. OK (1999): "The Measurement of Income Mobility: an Introduction to the Literature," in *Handbook on Income Inequality Measurement*, ed. by J. SILBER. Amsterdam: Kluver.

[68] FLINN, C. (2002): "labour Market Structure and Inequality: a Comparison of Italy and the US," *The Review of Economic Studies*, 69(3), 611-645.

[69] FLINN C. J., and J. J. HECKMAN (1982):"New Methods for Analizing Structural Models of Labor Force Dynamics," *Journal of Econometrics*, 18, 115-168.

[70] FLURY, B. (1986): "Asymptotic Theory for Common Principal Component Analysis," *Annals of Statistics*, 14, 418-430.

[71] FOUGÈRE D., and T. KAMIONKA (1992):"Un Modèle Markovien du Marché du Travail," *Annales d'Économie et de Statistiques*, 27, 501-508.

[72] FRÉCHET, M. (1935): "Généralisations du Théorème des Probabilités Totales," *Fund. Math.*, 25, 379-387.

[73] FREEMAN R. B. (1978):"Job Satisfaction as an Economic Variable," *American Economic Review*, Papers and Proceedings, 68, 115-168.

[74] GAGLIARDINI, P., and C. GOURIÉROUX (2002): "Constrained Nonparametric Copulas," *CREST Working Paper.*

[75] GEARY, R. C. (1942): "Inherent Relations Between Random Variables," *Proc. Royal Irish Academy*, 47, 63-76.

[76] GENEST, C., and R. J. MacKAY (1986): "Copules Archimédiennes et Familles de Lois Bidimensionnelles dont les Marges sont Données," *Canadian Journal of Statistics*, 14, 145-159.

[77] GEWEKE, J., and M. KEANE (2000): "An Empirical Analysis of Income Dynamics among Men in the PSID: 1968-1989," *Journal of Econometrics*, 96, 293-356.

[78] GIBBONS R. and L. KATZ (1991): "Layoffs and Lemons," *Journal of Labor Economics*, 9(4), 351-380.

[79] GIBBONS R. and L. KATZ (1992): "Does Unmeasured Ability Explain Inter-Industry Wage Differentials," *Review of Economic Studies*, 59(3), 515-535.

[80] GODDEERIS J. (1988): "Compensating Differentials and Self-Selection: An Application to Lawyers," *Journal of Political Economy*, 96(2), 411-28.

[81] GOTTSHALK, P. (1997): "Inequality, Income Growth and Mobility: the Basic Facts," *Journal of Economic Perspectives*, 11(2), 21-40.

[82] GOTTSHALK, P., and R. MOFFIT (1994): "The Growth of Earnings Instability in the US labour Market," *Brookings Papers on Economic Activity*, 2, 217-72.

[83] GOTTSHALK, P., and E. SPOLAORE (2002): "On the Evaluation of Economic Mobility," *Review of Economic Studies*, 69(1), 191-208.

[84] GRILICHES, Z. (1977): "Estimating the Returns to Schooling: Some Econometric Problems," *Econometrica*, 45, 1-22.

[85] GRONBERG, T., and R. REED (1994): "Estimating Workers' Marginal Willingness to Pay for Job Attributes Using Duration Data," *Journal of Human Resources*, 29, 911-931.

[86] GROOT, W., and M. VERBENE (1997): "Aging, Job Mobility, and Compensation," *Oxford Economic Papers*, New Series, 49, 380-403.

[87] GUILLOTIN, Y., and A. Bigard (1992): "La Mobilité Hiérarchique des Salaires en France de 1967 à 1982," *Economie et Prévision*, 102-103, 189-204.

[88] HALL, P., and Q. YAO (2003): "Inference in Components of Variance Models with Low Replications," *Annals of Statistics*, 31, 414-441.

[89] HAMERMESH, D. S. (1999): "The Changing Inequality for Workplace Amenities," *Quarterly Journal of Economics*, 114, 1085-1123.

[90] HAMERMESH, D. S. and S. G. DONALD, "The Effects of College Curriculum on Earnings: Accounting for Non-ignorable Non-response Bias," Texas University, *mimeo*.

[91] HANSEN, L. P. (1982): "Large Sample Properties of Generalized Method of Moments Estimators," *Econometrica*, 50(4), 1029-54.

[92] HAUSE, J. C. (1980): "The Fine Structure of Earnings and the On-the-job Training Hypothesis," *Econometrica*, 48(4), 1013-29.

[93] HECKMAN, J. J. (1981), "The Incidental Parameters Problem and the Problem of Initial Conditions in Estimating a Discrete Time - Discrete Data Stochastic Process," in *Structural Analysis of Discrete Data with Econometric Applications*, ed. by C.F. Manski and D. McFadden. Cambridge: MIT Press.

[94] HECKMAN, J. J. (2001), "Micro Data, Heterogeneity, and the Evaluation of Public Policy: Nobel Lecture," *Journal of Political Economy*, 109(4), 673-748.

[95] HECKMAN, J.J. and S. NAVARRO (2005), "Dynamic Discrete Choice and Dynamic Treatment Effects," University of Chicago, *mimeo*.

[96] HECKMAN, J.J., and B. SINGER (1984): "A Method For Minimizing the Impact of Distributional Assumptions in Econometric Models For Duration Data," *Econometrica*, 52, 271-320.

[97] HECKMAN, J., R. MATZKIN, and L. NESHEIM (2005): "Simulation and Estimation of Hedonic Models," in T.J. Keohe, T.N. Srinivasan, and J. Whalley, ed., *Frontiers in Applied General Equilibrium Modeling, In Honor of Herbert Scarf.* Cambridge.

[98] HERZOG, H.W., and A.M. SCHLOTTMAN (1990): "Valuing Risk in the Workplace: Market Price, Willingness to Pay, and the Optimal Provision of Safety," *Review of Economics and Statistics*, 72(3), 463-470.

[99] HIRANO, K. (2002): "Semiparametric Bayesian Inference in Autoregressive Panel Data Models," *Econometrica*, 70, 781-799.

[100] HOROWITZ, J. L. (1998): *Semiparametric Methods in Econometrics.* New-York: Springer-Verlag.

[101] HOROWITZ, J. L., and M. MARKATOU (1996): "Semiparametric Estimation of Regression Models for Panel Data," *Review of Economic Studies*, 63, 145-168.

[102] HU, Y. and G. RIDDER (2005), "On Deconvolution as a First Stage Nonparametric Estimator". IEPR Working Paper No. 05.29 http://ssrn.com/abstract=786604.

[103] HWANG, H.S., D. MORTENSEN, and W.C. REED. (1998): "Hedonic Wages and Labor Market Search," *Journal of Labor Economics*, 16, 815-847.

[104] HWANG, H. S., W. C. REED, and C. HUBBARD (1992): "Compensating Differentials and Unobserved Productivity," *Journal of Political Economy*, 100, 835-858.

[105] HYVARINEN A. (1999), "Fast and Robust Fixed-Point Algorithms for Independent Component Analysis," *IEEE Transactions on Neural Networks*, 10(3):626–634.

[106] HYVARINEN A., J. KARHUNEN and E. OJA (2001), *Independent Component Analysis*, John Wiley & Sons, New York.

[107] IKEDA, S. and K. TOYAMA (2000), "Independent component analysis for noisy data–MEG data analysis," *Neural Networks*, Vol.13, No.10, 1063-1074.

[108] JOE, H. (1997): *Multivariate Models and Dependence Concepts.* London: Chapman & Hall.

[109] JUHN, C., K.M. MURPHY, and B. PIERCE (1993), "Wage Inequality and the Rise in Returns to Skill," *Journal of Political Economy*, 101, 410-442.

[110] KATZ, L., and D. AUTOR (1999): "Changes in the Wage Structure and Earnings Inequality," in *Handbook of labour Economics*, ed. by Aschenfelter and Card. Amsterdam: Elsevier.

[111] KOSTIUK P. (1990): "Compensating Differentials for Shift Work," *Journal of Political Economy*, 98(5), 1054-75.

[112] KOTLARSKI, I. (1967): "On Characterizing the Gamma and Normal Distribution," *Pacific Journal of Mathematics*, 20, 69-76.

[113] LANG, K., and S. MAJUMDAR (2004): "The Pricing of Job Characteristics When Markets do not Clear: Theory and Policy Implications," *International Economic Review*, 45, 1111-1128.

[114] LAWLEY, D.N, and A.E. MAXWELL (1971): *Factor Analysis as a Statistical Method.* London: Butterworth.

[115] LEMIEUX, T. (2006): "Composition Effects, Wage Measurement, and the Growth in Within-Group Wage Inequality," *forthcoming American Economic Review.*

[116] LEVY, F., and R. MURNANE (1992): "US Earnings Levels and Earnings Inequality : a Review of Recent Trends and Proposed Explanations," *Journal of Economic Literature*, 30, 1333-81.

[117] LEWBEL, A. (1997): "Constructing Instruments for Regressions with Measurement Error When No Additional Data are Available, with an Application to Patents and R&D," *Econometrica*, 65, 1201-1213.

[118] LEWBEL, A. (2004): "Identification of Heteroskedastic Endogenous or Mismeasured Regressor Models," Boston College, *mimeo.*

[119] LI, T. (2002): "Robust and Consistent Estimation of Nonlinear Errors-in-Variables Models, " *Journal of Econometrics*, 110, 1-26.

[120] LI, T., and Q. VUONG (1998): "Nonparametric Estimation of the Measurement Error Model Using Multiple Indicators," *Journal of Multivariate Analysis*, 65, 139-165.

[121] LILLARD, L., and Y. WEISS (1979): "Components of Variation in Panel Earnings Data: American Scientists 1960-70," *Econometrica*, 47, 437-454.

[122] LILLARD, L., and R. WILLIS (1978): "Dynamic Aspects of Earnings Mobility," *Econometrica*, 46, 985-1012.

[123] LIN, Y.-N. and K. HUNG (2005): "The Volatility Risk Premium Embedded in S&P 500 Index Returns," mimeo

[124] LINTON, O., and Y. J. WHANG (2002): "Nonparametric Estimation with Aggregated Data," *Econometric Theory*, 18, 420-468.

[125] MACURDY, T. (1982): "The Use of Time Series Processes to Model Structure of Earnings in a Longitudinal Data Analysis," *Journal of Econometrics*, 18, 83-114.

[126] MADANSKY, A. (1959): "The Fitting of Straight Lines When Both Variables are Subject to Error," *Journal of the American Statistical Assiociation*, 54, 173-205.

[127] MANSKI, C. F. (1988): "Identification of Binary Response Models," *Journal of the American Statistical Association*, 83, 729-738.

[128] MANSKI, C. (1989): "Anatomy of the Selection Problem," *Journal of Human Ressources*, 24, 343-360.

[129] MARKANDYA, A. (1984): "The Welfare Measurement of Changes in Economic Mobility," *Economica*, 51(204), 457-471.

[130] MARTIN, J. P. (1996): "Measures of Replacement Rates for the Purpose of International Comparisons: a Note," *OECD Economic Survey,*, 26, 99-115.

[131] McCLENDON, McKEE J. (1977): "Structural and Exchange Components of Vertical Mobility," *American Sociological Review*, 42(1), 56-74.

[132] MEGHIR, C., and L. PISTAFERRI (2004): "Income Variance Dynamics and Heterogeneity," *Econometrica*, 72, 1-32.

[133] MENDELSON, S. (2003), "A few notes on Statistical Learning Theory," in *Advanced Lectures in Machine Learning*, (S. Mendelson, A.J. Smola Eds), LNCS 2600, 1-40, Springer.

[134] MEYER, B. (1995): "Natural and Quasi-experiments in Economics," *journal of Buisiness and Economic Statistics*, 13(2), 151-161.

[135] MOFFIT, R., and P. GOTTSCHALK (2002): "Trends in the Transitory Variance of Earnings in the US," *The Economic Journal*, 112, C68-C73.

[136] MOULINES, E. J.-F. CARDOSO and E. GASSIAT (1997), "Maximum likelihood for blind separation and deconvolution of noisy signals using mixture models," Proc. ICASSP'97 Munich, vol. 5, 3617-20.

[137] NELSEN, R. B. (1998): *An Introduction to Copulas.* New-York: Springer Verlag.

[138] NEWEY, W.K., and D.L. McFADDEN (1994): "Large Sample Estimation and Hypothesis Testing," *in* R.F. Engle and D.L. McFadden, eds., *Handbook of Econometrics* vol 4: 2111-2245. Amsterdan: Elsevier Science.

[139] PAL, M. (1980): "Consistent Moment Estimators of Regression Coefficients in the Presence of Errors-in-Variables," *Journal of Econometrics*, 14, 349-364.

[140] PIERCE, B. (2001): "Compensation Inequality," *Quarterly Journal of Economics*, 116(4), 1493-1525.

[141] PIAZZESI, M. (2003): "Affine Term Structure Models," prepared for the *Handbook of Financial Econometrics.*

[142] PIKETTY, T. (2001): *Les hauts revenus en France au vingtième siècle. Inégalités et redistributions, 1901-1998.* Paris: Grasset.

[143] PLACKETT, R. L., (1965): "A Class of Bivariate Distributions," *J. Am. Statis. Assoc.*, 60, 516-522.

[144] POLLARD, D. (1984): "Convergence in Distributions of Stochastic Processes." Springer, New-York.

[145] POSTEL-VINAY, F. and J.-M. ROBIN (2002): "Equilibrium Wage Dispersion with Worker and Employer Heterogeneity", *Econometrica* 70(6), 2295-350.

[146] PRAIS, S. J., (1955): "Measuring Social Mobility," *J. Royal Statis. Soc. Ser. A*, 118, 56-66.

[147] RAO, P. (1992): "Identifiability in Stochastic Models." New-York, Academic Press.

[148] REDNER, R., and H. WALKER (1984): "Mixture Densities, Maximum Likelihood and the EM Algorithm," *SIAM Review*, 26, 195-239.

[149] REIERSOL, O. (1950): "Identifiability of a Linear Relation Between Variables which are Subject to Error," *Econometrica*, 9, 1-24.

[150] ROBIN, J.M., R.J. SMITH (2000): "Tests of rank," *Econometric Theory*, vol. 16, 151-175

[151] ROSEN, S. (1974): "Hedonic Prices and Implicit Markets: Product Differentiation in Pure Competition," *Journal of Political Economy*, 82, 34-55.

[152] ROSEN, S. (1986): "The Theory of Equalizing Differences," *in* O. Ashenfelter and D. Card, eds., *Handbook of Labor Economics* vol 1: 2 641-692. Amsterdan: Elsevier Science.

[153] ROSS, S. (1976): "The Arbitrage Theory of Capital Asset Pricing," *Journal of Economic Theory*, 13, 341-360.

[154] SCHENNACH, S. (2004): "Estimation of Nonlinear Models with Measurement Error," *Econometrica*, 72, 33-75.

[155] SENICK, C. (2005): "What Can We Learn from Subjective Data: the Case of Income and Well-Being," *Journal of Economic Surveys*, 19(1), 43-63.

[156] SHORROCKS, A. (1976): "Income Mobility and the Markov Asumption," *The Economic Journal*, 86, 566-578.

[157] SHORROCKS, A. (1978): "The Measurement of Mobility," *Econometrica*, 46, 1013-1024.

[158] SINGER, B., and S. SPILERMAN (1976): "Some Methodological Issues in the Analysis of Longitudinal Surveys," *Annals of Economic and Social Measurement*, 5, 447-474.

[159] SKLAR, A. (1959): "Fonctions de Répartition à n Dimensions et leurs Marges," *Publ. Inst. Statist. Univ. Paris*, 8, 229-231.

[160] SMITH, A.: *An Inquiry into the Nature and Causes of the Wealth of Nations*, 1776.

[161] SOBEL, M. (1983): "Structural Mobility, Circulation Mobility and the Analysis of Occupational Mobility: a Conceptual Mismatch," *American Sociological Review*, 48(5), 721-727.

[162] SPEARMAN, C. (1904): "General intelligence, objectively determined and measured," *American Journal of Psychology*, 15, 201-293.

[163] SPIEGELMAN, C. (1979): "On Estimating the Slope of a Straight Line when Both Variables are Subject to Error," *Annals of Statistics*, 7, 201-206.

[164] STEFANSKI, L., and R. J. CARROLL (1990): "Deconvoluting Kernel Density Estimators," *Statistics*, 2, 169-184.

[165] STOCK, J., and M. W. WATSON (2002): "Forecasting Using Principal Components from a Large Number of Predictors," *Journal of the American Statistical Economic Association*, 97, 1167-1179.

[166] THALER, R., and S. ROSEN (1975): "The Value of Saving a Life: Evidence from the Labor Market", in N. Terleckyj, ed., *Household Production and Consumption.* New York: NBER.

[167] VAN den BERG, G. (1992): "A Structural Dynamic Analysis of Job Turnover and the Costs Associated with Moving to Another Job," *Economic Journal*, 102, 1116-1133.

[168] VAN OMMEREN, J., VAN den BERG, G., and C. GORTER (2000): "Estimating the Marginal Willigness to Pay for Commuting," *Journal of Regional Science*, 40, 541-563.

[169] VAN MONTFORT, K., A. MOOIJAART, and J. DE LEEUW (1989): "Estimation of Regression Coefficients with the Help of Characteristic Functions," *Journal of Econometrics*, 41, 267-278.

[170] VILLANUEVA E. (2005): "Compensating Wage Differentials and Voluntary Job Changes: Evidence from West Germany," *mimeo* UPF.

[171] WU, C. F. J. (1983): "On the Convergence Properties of the EM Algorithm," *The Annals of Statistics*, 11, 95-103.

[172] XENOGIANI, T. (2003): "Job Satisfaction, Mobility Decisions and Wage Gains by Gender," LSE Working Paper.

[173] XU, L. (2000): "Temporal BYY Learning for State Space Approach, Hidden Markov Model and Blind Source Separation," *IEEE Trans on Signal Processing*, Vol. 48, No. 7, 2132-2144.

[174] XU, L. (2001): "BYY Harmony Learning, Independent State Space and Generalized APT Financial Analyses," *IEEE Trans. on Neural Networks*, Vol. 12, No.4, 822-849.

[175] XU, L. (2003): "Independent Component Analysis and Extensions with Noise and Time: A Bayesian Ying-Yang Learning Perspective," *Neural Information Processing Letters and Reviews*, Vol.1, No.1, 1-52.

[176] YIP, F. and L. XU (2000): "An Application of Independent Component Analysis in the Arbitrage Pricing Theory," IEEE-INNS-ENNS *International Joint Conference on Neural Networks*, 5, 5279.