Variable importance measures in semiparametric and high-dimensional models with or without error-in-variables

Cabral Amilcar Chanang Tondji

UNIVERSITÉ PARIS-EST

École Doctorale Mathématiques et Sciences et Technologie de l'Information et de la Communication (MSTIC)

Thèse de doctorat

Discipline : Mathématiques

Présentée par

Cabral Amilcar CHANANG TONDJI

Variable Importance Measures in Semiparametric and High-Dimensional Models with or without Error-in-Variables

---

Soutenue le 11 Décembre 2020 devant le Jury composé de :

Mme.	Cristina	Butucea	ENSAE - Institut Polytechnique de Paris	Directrice
M.	Antoine	Chambaz	Université de Paris	Co-directeur
Mme.	Gabriela	Ciuperca	Université Claude Bernard de Lyon I	Rapporteure
Mme.	Céline	Duval	Université de Paris	Examinatrice
M.	Eric	Gautier	Université Toulouse I	Rapporteur
M.	Viet Chi	Tran	Université Gustave Eiffel	Examinateur

Nèlɑ̀btɛ̌

« Mbà zîd mɛn nu nkə nku’ ɑ̂ ntɑ̂m tà’ la’ »

Mǎ cwɛ̌d mfɑ məbam kə̀ siaŋ nzô ghǎcàgtə̀ sɑm fã’ nədiàgə : Cristina bô Antoine, kuà’ nənən bìn nɑ̂ kə̀ bə nsi mbɑ̂ mə̀ kə̂’ kù’nĭ nə̀ kìtə mân nkămncob mì mǎ cwɛ̌d nkî kuà’ ndɔ̌nni lɑ ; Nkămncob mənɑ mì bo cwɛ̌d nkudtû ntswə’njɑ̂ ngə’ mbà shʉɑlo njɔ̌ŋle’e nŭm sɑ̂m laŋ ŋwà’nì ntɑ̂m bă ngɑ̂mmfămbu lɑ. Njòg mbə ɑ̂ tiàŋtə yɑmə. Kɑ̂ bìn lăb wɑ̂d zin mbɑ mǎ lɛ̀n mbə fà’ ndʉb yŏg nkə nə to’tə. Mətswə bŏ nŭm ncob mbə « ò mă’ lɑg njàm mbad mbɑ̂ ò à’ yən bɑ mbalə » ? sə̂ fã’ nədiàg lì kə̂’ bə ɑ̂ njɑ̀b kə̀ nkʉnə nă’ mbə mǎ cʉɑtə̀ bwə nkwî’tə zənù ɑm mbɑ sɑ̂m lɛ̌n nŭm ngɑ̂mmfămbu mbà nŭm Mɛ̀n nyănŋtu. Cristina, mǎ fɑ məbam wŭ nŭm ntagə, yù nəcʉɑgə mbɑ yù nətəlɑg nŭm mfɛ̃nkʉnə. Yên màd lì mbə miɑgtə mfə̂n dŭ’ntswə nùm ɑm nă’ mbə mfã’ nsi mbwə ɑm yɑm nɑ kɑ mə̀ bə mbɑ̂ wù lɑ. Kuà’ nənənə mǎ fɑ məbam wù nŭm yù màd nə̀ lò’ ntsiaŋtə ɑm yi nùm nkŏ’fà’ mbɑ̆ mbɑ ò bên kɑ̂g nzì mǎ nə fã’ nùm cɛd lɑ. Antoine, məbam kə̀ siaŋ nzə bə̂ wŭ yù nŭ ntag mì ò fɑ̂ mə lɑ kə̀ lɑ̆g jûjʉ zənù ò kem ɑ̂m yi Nanterre mbɑ̆ Panthéon nɑ, Laŋnzə nətɛ̆d bàn nkə̂’ jwìmtə̌ yɑmə nă’ mbə ŏ tswə mbaŋ àm yên ngə̀laŋ nɑ. Nju zĭn tu ɑm nɑ yɑmə.

« Bɑm nə̀ nkə nzi nùm lə̂dtu » cʉ̆ ncob zənù bən bŏ bwə̀ ɑm nɑ nŭ nkə njôb mə̀ kə̀ bod nə̀ gʉ̀ njə mə̀ bâm nŭm miɑ̂glo ncʉb fã’ mbɑ̂ bûn mɛnfi kà kə̀ kàg lɑ. À nyɑ̂ kə̀ bə ɑ̂ bo mbɑ̂ mə̀ kə̂’ miàgtə̀ sə̂ fã’ nədiàg li. Lɑ ă bwɔ̌ mə lɑbtə yub nŭ yên mìbntag li à midtə nùm ɑm nɑ. Tɑ, Mɑ mǎ làbtə̀ zin nŭ njŏŋ cu fa ; məbwɔ məbwɔ mbɑ̆ nŭ nkɔ̀nì kə̀ siaŋ mì bĭn nkə kɑbtə ɑm yi lɑ.

Linda mə̌ nû mbə tomtə mɑ̂m mfă Master ŏ tagtə̀ mə̀ mbə̌ mə̀ co ntʉ̂ nədiɑg mbə Ndɛngɑm. Ò nû kə̀ lɛ̀n yên ngə̀laŋ nɑ mbə̀ mə̀ à’ lŏ’ mân ntag lɑ nkwâ jù yi kɑ nco yɑ̆m nka. Lɑ mǎ làbtə̀ o. Rita bô Winnie, bin bè tə̂ ntom tŭn nkù mɑm ntɑ̂m yên nkadtə̀ nədiàg lì lɑ mǎ fɑ məbam bin.

Hulda, mbà ŏ yən nɑ sən fà’ li kə̂’ lŏ’ ngŭ’ gham, mǎ nkə ndɑbtə o nŭm ntag mì ò fɑ mə̀ lɑ, bà ngə̀laŋ kòmzwì mbà ke’təwud ò nkə ndo’ ɑm ncu yi bwə Paris lɑ. Mə̌ làbtə̀ o nŭm ntom, nkɔ̀nì ngʉ̂ tsə̀mô’ cu ò nɑg ɑm yi kɑ̂ mə̌ nku’ni nə̀ tum ntɑ̂m yên ndɑ̂njà lì lɑ.

Mă cwĕd nkə ndɑ̂btə mɛntɑ̂ Jean Marie nŭm ntag mì à còŋ ɑm yi lɑ. Ŏ nu mbə mbaŋ ɑ̆m Paris kà mbə njə̀ ò nû mbə kə̀ bə nsi mbɑ̂ mə̀ nû kə̀ kù'nĭ nə̀ miàgtə̀ sə̂n fà' lì. Bə̂ mə̀ kə̀ làgtə̆ Ngònmɑ̂ Véronique : mə̆ làbtə̆ nŭm yù nə lò' ndo' ɑ̂m yi. Mə̆ fɑ məbam bə̆nmɑ cɑm Roussel, Daniel mbɑ Hughes. Nyûlɑg mĕn mbɑ̂ mɛntɑ̂, Tɑ̂ zə̀ndʉb Tchakam à bə̂ ntʉ̂m kwàtə̀ mɑm njŏŋ le' fa lɑ.

Nə̀ tum bwə mə̆ fɑ məbam njŏŋ bə̀tʉ̀n fa bò yɑb mbu nùm sə̂n fà' li lɑ, kə bŏ bə̂ ngʉɑ̂ nə̀sà kə mbɑ̆ mbaŋ ə̀m nù : Olivia, Nelsone, Loic, Yannick, Ronald, Alexandre bô Landry. Məbam kə̀ siaŋ nzə bə̂ tŭnndɑ̂ Nana nŭm mbà bo ghâ' mbu mub kɑ ntiaŋtə ɑm yi lɑ. Bin lâgtə mfɑ̂ mə̀ mbə zə̀ mə̀ lâg mènntʉ̀n mbà ncʉ̀' lɑ. À bə ɑ̂ zwiàg nə̀ bĕn dʉ̀' i nkə ndam ɑm mbum yə̂n ngə̀laŋ nə̀ kud tûnù lì. Mə̆ kwatə mbə bìn nkə laŋ ɑ̂m mba' nùm bə.

Nə Məmtə

Tunù : Nətə ndom nŭm jûncob ncàŋ Mbàtənjòŋ nshʉɑtə nŭm nə tɑmtə nkă-nkɑ̀n mbɑ̀ kə̀-kɑ̆n ntû mfì’ ndʉb.

Màd zə ngɑ̂mnkŭn cwɛ̆d nkûndə cən ngŭ’gham nì lɑ kwì’tə̆ zəbən nəta nə diɑ̆g cûnkɑ̀n yi ntû nzə̆nzə̆ ngɑ̂mfà’ mbɑ̂ nəcwìntə̆ ngòkɛd, Nə dʉ’ca’a, Nə nɑ̆b yaŋmbà’, ngû tsəmo’o. Bə̆ cɑɑtə̀ bwə nco ntûm laŋ njʉ « Nkɑ̀n ndûbndʉb ». Bŏ nɑ̆b cûnkɑ̀n nzwə nə na nkɑ̀n yi nsô cu tsə̀ bô bŏ fâ’ yi lɑ bwə. Mfì’ bə ɑ̂ mbə ngɑ̂msiaŋtə̀ à à’ kə mbə fà’ nkə ntəlɑg nŭ Ntotô Mbàtənjòŋ mì ă fà’ yi lɑ, ncʉɑtə bwə ntswî sô lĕn mfì’ nə ma’a.

Caŋ yên lĕn mfì’ nə mà’ lɑ bə ɑ̂ nə totə nkò’ à bô nətɛ̆d tà’ Nkŭnù (zə bŏ nkə ntsiâŋ mŏngəlaŋ Mbɑ̀ntô Y) bo ɑ bɑ̀ Mbàtənjòŋ mì bò gʉ i lɑ (kə nə fəl i lo, bŏ nkə ntsiâŋ mŏ ngəlaŋ X lɑ) zə bo nŭm mbĕn ntsiâŋ Ntotô Mbàtənjòŋ lɑ. Cŭ măd zə bô bŏ kî yên nkò’ lɑ li Y = f(X), dŭ’ zə f bô tà’ Nəfə zə ă cwɛ̆d ju’ ɑ̂ Tûmnù bò kû’ni nə bə kə̀ lèn nɑ (yên ngəlaŋ nɑ bŏ tsiâŋ lĕn kɑn), kə̀ nə bə kə̀ lèn nû (yên ngəlaŋ nɑ bŏ tsiâŋ lĕn kə̀-kɑ̀n). À nû bə mbwə̀ bŏ nkə nsiaŋtə Ntotô Mbàtənjòŋ ɑ gham ngû ncòb tsə bwə. Ndɔ̆nì, bŏ kŭm nguɑ̂n nca’ kɑ ncûb ju bwə, mfə ɑ̂ bă fà’ bò bô bwə lɑ. Mbɑ̀ bò bô kə̀ kù’nĭ nə lo’ yob fâ’ mfà yi lɑ bŏ nŭm nkâ tsə bô bŏ fâ’ yi lɑ.

Ntûm sô fă nədiɑg lì, bàg sâŋ laŋ yɑm ɑ̂ nŭm màd nzwə nə kă Ntotə Mbàtənjòŋ, məbwɔ məbwɔ ngəlâŋ bŏ cwɛ̆d mfi’tə i ɑ̂ tû laŋfà’ ndʉb lɑ. Băg zi’tə màd nə fa’ bô nkɑ̀n ă nsi lɑ. Cʉɑtə bwə ntswî màd ă nkə nzîn nɑ nə lò’ ndâ’tə njàŋ bĭn nə tɑmtə nŭm bĭn mfă’mbu.

Abstract

Title: Variable Importance Measures in Semiparametric and High-Dimensional Models with or without Errors-in-Variables

During the last few decades, the advancements in technology we witnessed have considerably improved our capacities to collect and store large amount of information. As a consequence, they enhanced our data mining potential. The repercussions, on multiple scientific fields, have been stark. In statistical analysis for example, many results derived under the then common low dimensional framework, where the number of covariates is smaller than the size of the dataset, had to be extended. The literature now abounds with significant contributions in high dimensional settings. Following this path, the current thesis touches on the concept of *variable importance* that is, a methodology used to assess the significance of a variable. It is a focal point in today's era of big data. As an example, it is often used for prediction models in high dimensional settings to select the main predictors. Our contributions can be divided in three parts.

In the first part of the thesis, we rely on semiparametric models for our analysis. We introduce a multivariate *variable importance measure*, defined as a sound statistical parameter, which is complemented by user defined marginal structural models. It allows one to quantify the significance of an exposure on a response while taking into account all other covariates. The parameter is studied through the *Targeted Minimum Loss Estimation (TMLE)* methodology. We perform its full theoretical analysis. We are able to establish consistency and asymptotic results which provide as a consequence p-values for hypothesis testing of the parameter of interest. A numerical analysis is conducted to illustrate theoretical results. It is achieved by extending the implementation of the *TMLE.NPVI* package [20] such that it is able to cope with multivariate parameter.

In the second part, we introduce a *variable importance measure* which is defined through a nonparametric regression model under a high dimensional framework. It is partially derived from the parameter described in the first part of the thesis, without the requirement that the user provides a marginal structural model. The regression model comes with the caveat of having a data structure which, in some cases, is subject to measurement errors. Using a high-dimensional projection on an orthonormal base such as *Fourier series*, *smoothing splines* and the *Lasso* methodology, we establish consistency and the convergence rates of our estimators. We further discuss how these rates are affected when the design of the dataset is polluted. A numerical study, based on simulated and on financial datasets, is provided.

In the third and final part of this thesis, we consider a *variable importance measure* defined through a linear regression model subject to errors-in-variables. This regression model was derived in the previous chapter. The estimation of the parameter of interest is done through a convex optimization problem, obtained by projecting the empirical covariance estimator on the set of symmetric non-negative matrices, and using the *Slope* methodology. We perform its complete theoretical and numerical analysis. We establish sufficient conditions, rather restrictive on the noise variables, under which to attain optimal convergence rates for the parameter of interest and discuss the impact of measurement errors on these rates.

Keywords: errors-in-variables, high dimensional estimation, Lasso, Slope, convex optimization, semiparametric model, inverse problems, nonparametric model, TMLE, variable importance measure, marginal structural model, statistical inference.

Résumé

Titre: Mesures de l'importance de variable au travers de modèles semi-paramétriques et en grande dimension avec ou sans erreurs sur les variables

Les progrès technologiques de ces dernières décennies ont considérablement accru nos capacités à collecter et sauvegarder une quantité importante d'information. Ce faisant, notre marge de manœuvre pour l'exploitation de ces données, a été amplifiée. Les répercussions, sur de nombreux domaines scientifiques, ont été fulgurantes. En analyse statistique par exemple, de nombreux résultats obtenus sous le canevas habituel d'étude en petite dimension, qui consistait à considérer le nombre de variables explicatives inférieur à la taille de l'échantillon, ont dû être étendus. La littérature scientifique abonde maintenant de nombreux résultats qui ont été mis en exergue, en prenant en compte cette nouvelle realité qu'est la présence des données en grande dimension. Nos travaux s'inscrivent dans cette droite lignée. En effet, cette thèse aborde le concept d'*importance de variables*, c'est-à-dire un canevas permettant de déterminer la portée d'une variable. Il s'agit là d'un point crucial dans cette nouvelle ère de données de grande taille. À titre d'exemple, ce concept est largement utilisé dans des modèles de prédiction afin d'améliorer le choix des variables explicatives. Nos contributions peuvent être divisées en trois parties.

Dans la première partie, nous introduisons une mesure multivariée dénommée *mesure de l'importance de variable*, définit en tant que paramètre statistique, assujettie à des modèles de structures marginaux. Nous nous sommes appuyés sur des modèles semi-paramétriques pour son analyse. Cette mesure permet notamment de quantifier la pertinence d'une variable explicative sur une réponse, en prenant en compte le reste des variables du problème. Le paramètre d'intérêt est étudié grâce à la méthode du *TMLE (Tartgeted Minimum Loss Estimation)*. Nous effectuons son analyse théorique complète et sommes ainsi en mesure d'établir la consistance de notre estimateur, ainsi que sa convergence asymptotique. Ce dernier résultat nous permet donc de déduire les intervalles de confiance liés à l'estimateur. Nous effectuons également une analyse numérique afin d'illustrer nos résultats théoriques.

A cet effet, nous avons étendu l'implémentation du package *TMLE.NPVI* [20], de telle sorte qu'il puisse traiter des cas où le paramètre d'intérêt est multivarié.

Dans la seconde partie de cette thèse, nous introduisons une *mesure de l'importance de variable* définie au travers d'un modèle de régression non-paramétrique en grande dimension. Cette mesure provient en partie de celle introduite dans la première partie, sans la contrainte supplémentaire que l'utilisateur doive fournir un modèle de structure marginal. Au-delà, nous considérons également le cas où les données de notre échantillon sont polluées. En s'appuyant sur une décomposition finie sur une base orthonormée du type *base de Fourier* ou *Splines* par exemple, et en utilisant la méthode dite du *Lasso*, nous établissons les vitesses de convergence de notre estimateur. Nous mettons aussi en exergue l'impact des erreurs de mesure, dans notre design, sur ces vitesses de convergence. Au-delà, nous proposons également une étude numérique basée sur des données synthétiques et une application, s'appuyant sur des données financières réelles.

Dans la troisième et dernière partie, nous considérons une *mesure d'importance de variable* définie grâce à un modèle de régression linéaire soumis à des erreurs de mesure sur son échantillon. Ce modèle de régression trouve son origine dans la partie précédente. L'estimation de notre paramètre d'intérêt s'effectue au travers d'un problème d'optimisation convexe, obtenu en projetant la covariance empirique du design sur l'ensemble de matrices définies positives, et en utilisant la pénalisation *Slope*. Nous effectuons ainsi une analyse théorique et numérique complète. Au-delà, nous établissons les conditions suffisantes, assez restrictives concernant les erreurs, à respecter afin d'atteindre des vitesses optimales de convergence de notre paramètre d'intérêt, tout en mettant l'accent sur l'impact de la pollution de notre échantillon sur ces vitesses.

Mots Clés: erreurs de mesure, estimation en grande dimension, inférence statistique, Lasso, Slope, optimisation convexe, modèle semi-paramétrique, modèle non-paramétrique, mesure de l'importance de variable, modèle de structure marginal, problèmes inverses, TMLE

Remerciements.

If I have seen further, it is by standing on the shoulders of giants.
Isaac Newton

Je souhaite exprimer ma profonde gratitude à mes deux directeurs de thèse : Cristina et Antoine. Il est certain que sans vous et votre support inconditionnel, je ne me trouverais point en train d'écrire ces quelques mots, qui marquent la fin de ce long et tortueux parcours scientifique, mais ô combien formateur. Ne vous inquiétez pas car je sais, qu'en réalité, il ne fait que commencer. Ne dit-on pas au pays de *Toussaint Louverture* : "*Beyond the mountains, more mountains*". Cette thèse n'a pas été conventionnelle mais grâce à votre soutien, elle m'a permis de grandir et d'affermir non seulement mes connaissances scientifiques, mais aussi humaines. Cristina, je te dis un grand merci pour tous tes conseils, ta rigueur et cette attention méticuleuse aux détails. Tout ceci a été contagieux mais je sais que j'ai encore du travail à fournir pour atteindre un niveau similaire au tien. Et bien sûr, merci pour tous les accueils chaleureux sur le "*platal*" et aussi pour m'avoir redonné plaisir à travailler sur un tableau ! Antoine, un grand merci à toi également pour les conseils et échanges très instructifs à Nanterre et au *Panthéon*. La distance physique n'a pas aidé mais tu as su rester à mes côtés jusqu'au bout. Je vous dois énormément à tous les deux !

Je suis honoré que Mme Gabriela Ciuperca et M. Eric Gautier aient accepté de rapporter ma thèse. Vos remarques et suggestions ont permis d'améliorer ce manuscrit. Je remercie Mme Céline Duval et M. Viet Chi Tran pour leur participation à mon Jury.

"Ne jamais t'endormir sur tes lauriers" ! C'est avec cette citation, proférée si souvent au détour de nos nombreuses et précieuses discussions, que mes parents n'ont eu de cesse d'essayer de me faire prendre conscience de la profondeur des concepts *d'ardeur au travail* et *de résilience*. Sans ces derniers, je n'aurai pas pu parachever cette thèse. C'est donc tout naturellement que je vous remercie pour ces valeurs qui raisonnent plus que jamais

en moi. Je vous dis merci du fond du cœur pour tout, mais surtout pour votre amour inconditionnel.

Linda, lors de ma cérémonie de graduation de *Master*, tu m'avais susurré l'idée de faire un doctorat. Tu ignorais sans doute à ce moment-là que j'allais prendre ce conseil pour ordre et m'exécuter. À toi, je dis merci. Rita et Winnie, vous qui m'avez encouragé et supporté tout au long de ce voyage quasi *initiatique*, je vous dis également merci.

Hulda, comme tu le constates, il m'aura fallu moins de 20 ans ! Enfin presque ! Merci à toi ma très chère pour ton aide, les fous rires, ta tendresse, le soutien moral incommensurable après mes séjours presque toujours difficiles sur Paris. Un grand merci pour tes encouragements, ton amour et les sacrifices consentis pour me permettre de mener à bien cette thèse.

Je tiens aussi à remercier mon oncle Jean Marie pour son soutien et ses conseils. Merci d'avoir été à mes côtés depuis mon arrivée sur Paris. Je n'aurais sans doute jamais pu effectuer ce parcours académique sans toi. Une pensée toute particulière pour mon oncle et grand père Tchakam Flaubert, qui est toujours dans mes louanges. Un grand merci à mes oncles Roussel, Daniel et Hughes. J'ai aussi une pensée très chaleureuse pour ma tante Véronique : merci pour ton hospitalité légendaire et ton affection.

Pour finir, j'adresse également des remerciements chaleureux à tous ceux qui ont de près, comme de loin, contribué à ce travail : Olivia, Nelsone, Loic et Yannick. Un grand merci tout particulier à la famille Nana, pour m'avoir accueilli à bras ouvert et ainsi facilité, à bien des égards, mes dernières années de thèse. Je ne saurais terminer ce document sans mentionner mes amis de longue date : Landry, Alexandre, Ronald et Vincent. Nos conversations et échanges, surtout durant ces dernières années, m'ont permis d'affronter l'adversité avec force et conviction. J'oublie certainement de nombreuses personnes et je m'en excuse auprès de vous par avance. Sachez que c'est l'esprit *Diopien*, qui m'anime de plus en plus, qui me pousse à cette concision. Je suis convaincu que vous ne m'en tiendrez pas trop rigueur.

À mon père, à ma mère et mes ancêtres.
Modeste expression de mon infini reconnaissance et de ma piété filiale.

Contents

Résumé substantiel

Cette thèse aborde principalement le concept de *mesure de l'importance de variable*. Il s'agit d'un canevas qui permet de déterminer quantitativement, et parfois qualitativement, la portée d'une variable. C'est une notion primordiale dans de nombreux domaines scientifiques, notamment les problèmes statistiques comportant des jeux de données de grande taille.

Notre étude de *mesure d'importance de variable* s'effectue sous deux principaux registres. Premièrement, on définit et étudie une *mesure d'importance de variable* de dimension finie et définit sur un espace non paramétrique. La mesure est ainsi qualifiée de non paramétrique en ceci qu'elle est définie comme une valeur $\psi_f(P_0)$ d'une loi inconnue P_0, provenant d'une fonctionnelle ψ_f. Cette fonctionnelle est définie sur un ensemble non paramétrique de loi $\mathcal{M}$ qui est soumise à des contraintes peu restrictives. Dans ce cadre, f est une fonction fournie par l'utilisateur. Dans le Chapitre 3, nous développons et analysons un estimateur basé sur la méthode dénommée *targeted minimum loss estimation*. Sous des hypothèses bénignes, nous pouvons construire des intervalles de confiance asymptotique de notre estimateur, associés à un niveau donné.

Deuxièmement, en s'inspirant de la définition de ψ_f, nous introduisons une seconde fonctionnelle ψ_{pen}, de dimension finie. Elle est définie sur l'ensemble $\mathcal{M}' \subset \mathcal{M}$, caractérisé sous $\mathcal{M}$ par la présence en son sein d'un modèle de régression. Contrairement à ψ_f, ψ_{pen} ne requiert point une fonction définie par l'utilisateur. En supposant que P_0 soit un élément de $\mathcal{M}'$, $\psi_{pen}(P_0)$ est bien défini et ce vecteur peut être interprété comme une autre mesure d'importance de variable similaire à $\psi_f(P_0)$ et identique à ce dernier pour un choix précis de f. Dans les chapitres 3 et 4, nous développons et analysons deux estimateurs de $\psi_{pen}(P_0)$, basés respectivement sur les méthodes de *Lasso* et *Slope*. Sous des hypothèses propres à ce registre, nous établissons des vitesses de convergence. Des analyses numériques sont effectuées afin d'illustrer nos résultats théoriques.

La mesure d'importance de variable ψ_f sous $\mathcal{M}$

D'importants progrès ont été réalisés dans l'article [75] dans le domaine de *mesure d'importance de variable* pour les problèmes statistiques. L'auteur y a proposé de définir cette mesure comme un paramètre statistique. Il pouvait ainsi être étudié au travers d'estimateurs adéquats. On pouvait donc en déduire des propriétés asymptotiques pour les estimateurs, facilitant ainsi la construction d'intervalles de confiance pour un niveau donné. Cette approche a inspiré de nombreux articles ultérieurs, dont le plus important pour nous fût [22], sur lequel notre travail se repose.

Considérons le problème statistique s'appuyant sur la structure de données $\mathcal{O} = (X, W, Y)$, provenant d'une unité expérimentale d'intérêt $W \in \mathcal{W} \subset \mathbb{R}^d$, représentant des vecteurs de variables explicatives, $X \in \mathbb{R}$ (une exposition continue) une variable réelle de cause affectant $Y \in \mathbb{R}$ (une réponse continue), une valeur réelle de variables d'effets. On nomme par P_0 la vraie loi générant la structure de données $\mathcal{O}$. On suppose que l'exposition contient un niveau de référence x_0. En d'autres termes, il existe $0 < c < 1/2$ tel que $P_0(X \neq x_0|W) \in [c, 1-c]$ P_0-presque surement.

Notre objectif est de quantifier la relation qui existe entre X et Y, tout en prenant en compte W. Il est primordial de considérer W pour établir cette relation car nous ne pouvons pas tout simplement ignorer son impact. Par analogie avec [22], nous introduisons une *mesure d'importance de variable* $\psi_f(P_0)$ caractérisée par

$$\psi_f(P) = \arg\min_{\beta \in \mathbb{R}^d} E_P\left[\left(Y - E_P(Y|X = x_0, W) - (X - x_0) f_\beta(W)\right)^2\right], \tag{1}$$

où $f_\beta(W) = \beta^\top \cdot f_{CT}(W)$ avec $f_{CT} : \mathcal{W} \to \mathbb{R}^d$ une fonction fournie par l'utilisateur. Nous remarquons ici que $\psi_f(P_0)$ est multivarié et sa taille correspond au nombre d'éléments de W. Cette dernière characteristique de $\psi_f(P_0)$ est cruciale car elle permet d'obtenir une *mesure d'importance de variable* qui capture le rôle que joue chaque élément de W dans la relation qui existe entre X et Y. Nous ne faisons pas l'hypothèse qu'il existe un β tel que

$$Y = E_{P_0}(Y|X = x_0, W) - (X - x_0) f_\beta(W) + \epsilon,$$

où ϵ est un bruit inconnu. En d'autres termes, $\psi_f(P_0)$ est universellement bien défini.

Le paramètre d'intérêt

Afin d'étudier le paramètre d'intérêt, nous utilisons une méthode d'estimation semiparamétrique dénommée *Targeted Minimum Likelihood Estimation* (*TMLE*) par analogie avec [22].

Considérons (1) et supposons sans perte de généralité que $x_0 = 0$. Nous dénommons

$\dot{f}_\beta = \frac{\partial f_\beta}{\partial \beta} = f_{CT}$, le gradient de f_β et $L_0^2(P)$ l'ensemble des fonctions mesurables g tel que $E_P[g^2] < \infty$ et $E_P[g] = 0$. Nous dénommons aussi

$$\theta(P)(X,W) = E_P(Y|X,W), \quad g(P)(W) = P(X \neq 0|W),$$

$$\mu(P)(W) = E_P(X\dot{f}_\beta(W)|W) \quad \text{et} \quad \Sigma(P) = E_P[X^2\dot{f}_\beta(W)^\top \dot{f}_\beta(W)],$$

quelques propriétés importantes de la loi P. Si nous supposons que

- toutes les propriétés de P sont bien définies,
- Σ est inversible,
- il existe un $c \in]0, 1/2[$ tel que $g(P)(W) \in]c, 1-c[$ P-presque surement,

alors, la solution de (1) est unique et est donnée par

$$\psi_f(P) = \Sigma(P)^{-1}\left[E_P\left(X\dot{f}_\beta(W)\left(\theta(P)(X,W) - \theta(P)(0,W)\right)\right)\right].$$

La procédure d'inférence statistique

La procédure d'inférence statistique est au cœur de notre analyse numérique. Elle s'appuie sur n indépendantes copies $\mathcal{O}^{(i)}$, avec $i \in \{1,\ldots,n\}$ de la structure de données observées $\mathcal{O}$. Comme mentionné plus haut, l'estimation de $\psi_f(P_0)$ est basée sur la méthode dite du *TMLE*. Elle peut être divisée en deux principales étapes. Dans la première étape, l'estimateur initial $\psi_f(P_n^0)$ est évalué, avec P_n^0 construit comme un élément de $\mathcal{M}$ s'appuyant sur les données $\mathcal{O}^{(1)},\ldots,\mathcal{O}^{(n)}$.

Nous remarquons que cet estimateur initial peut être biaisé. Ceci ne constitue pas une difficulté car ce biais, si il existe, sera corrigé au travers des mises à jour successives. On remarque également qu'il n'est pas nécessaire d'estimer toute la loi P_0, mais seulement certaines de ses propriétés, à savoir : $\mu(P_0)$, $g(P_0)$, $\Sigma(P_0)$ et $\theta(P_0)$.

Dans la seconde étape, on construit k mises à jour $P_n^1,\ldots,P_n^k$, qui nous permettent d'établir la séquence d'estimateurs $\{\psi_n^j = \psi_f(P_n^j)\}_{j=1,\ldots,k}$. Les mises à jour P_n^k sont construites à partir de la *fonction d'influence éfficace* liée à ψ_f. La procédure s'arrête lorsque la suite d'estimateurs converge. Dans le chapitre 2, nous sommes donc en mesure d'établir un théorème de convergence de notre estimateur. Il prouve ainsi qu'il est convergent et asymptotiquement normal.

Au-delà de ψ_f au travers un sous ensemble de $\mathcal{M}$

Considérons la structure de données $\mathcal{O} = (X, W, Y)$ d'une unité expérimentale d'intérêt où $X \in \mathbb{R}$ est une exposition, $Y \in \mathbb{R}$ est une réponse et $W \in \mathcal{W} \subset \mathbb{R}^d$ représente des variables explicatives. Cette structure de données est extraite de la vraie loi $P \in \mathcal{M}'$, où $\mathcal{M}' \subset \mathcal{M}$ et $\mathcal{M}'$ est caractérisé par $(\eta_P^\star, \theta_P^\star) \in \mathbb{R}^m \times \mathbb{R}^k$ tel que

$$Y = \bar{\Phi}_m^\star(W)\eta_P^\star + X \cdot \bar{\Phi}_k^\star(W)\theta_P^\star + \varepsilon,$$

où $\bar{\Phi}_i^\star = \{\Phi_1^\star, \ldots, \Phi_i^\star\}$ est un vecteur contenant les i premières composantes d'un *espace d'Hilbert*.

La mesure d'importance de variable $\psi_f(P)$ (1) s'appuie sur une fonction f_β qui est fournie par l'utilisateur, pour des raisons de simplicité. Cette fonction est par nature une propriété inconnue de notre problème statistique et ce faisant une source de complexité. Dans le chapitre 3, nous décidons d'aborder le problème sous un autre angle. Nous utilisons la même mesure d'importance de variable sans toutefois demander à ce que la fonction f_β soit spécifiée par l'utilisateur.

Notre objectif demeure de quantifier la relation qui existe entre la réponse Y et l'exposition X, tout en prenant en compte W. On dénomme par $\mathcal{F}$, l'ensemble des fonctions non paramétriques et continues sur $\mathbb{R}^d$. On introduit par analogie avec (1), un paramètre statistique $\psi_{pen}(P)$ caractérisé par

$$\psi_{pen}(P) = \arg\min_{f \in \mathcal{F}} E_P\left[(Y - E_P(Y|X = x_0, W) - (X - x_0) \cdot f(W))^2\right]. \tag{2}$$

Nous remarquons que le paramètre d'intérêt est maintenant une fonction. Ce faisant, nous introduisons une méthode plus flexible permettant de caractériser les interactions entre les éléments de W. Ceci nous permettra ainsi d'avoir une lecture plus complète de notre mesure d'importance de variable.

La fonction d'intérêt

Considérons $\Phi = (\Phi_1, \ldots, \Phi_p, \ldots) = \{\Phi_j\}_{j=1}^\infty$ une base orthonormée $(b.o.n.)$ de $\mathcal{H}$, un *espace d'Hilbert*. Sachant que $P \in \mathcal{M}'$, on sait qu'il existe $(\eta_P, \theta_P) \in \mathbb{R}^m \times \mathbb{R}^k$ tel que, sous P,

$$Y = \bar{\Phi}_m(W)\eta_P + X \cdot \bar{\Phi}_k(W)\theta_P + \varepsilon, \tag{3}$$

où ε est une variable aléatoire centrée gaussienne, ayant pour variance σ^2, indépendante de (X, W), et $\bar{\Phi}_j = \{\Phi_1, \ldots, \Phi_j\}$ représente un vecteur dont les éléments sont les j premières composantes de la *b.o.n.* $\mathcal{H}$. En plus, $g(P)(W) = E_P[Y|X = 0, W] = \bar{\Phi}_m(W)\eta_P$ et, si on choisit $f(W) = \bar{\Phi}_k(W)\theta_P$, alors $\psi_f(P) = \theta_P$.

Par analogie avec (1), $\bar{\Phi}_k(W)\cdot\theta_P$ est notre nouvelle *mesure d'importance de variable*. Nous n'exploitons par la forme de $\psi_{pen}(P)$ définit sous (2), mais plutôt nous nous appuyons sur le modèle de régression (3) pour son analyse.

On réécrit (3) au travers de

$$Y = \mathbb{X}\cdot\beta + \epsilon, \tag{4}$$

où $\beta = \begin{pmatrix} \eta \\ \theta \end{pmatrix} \in \mathbb{R}^p$ avec $p = m + k$ et $\mathbb{X}$ correspondant à

$$\mathbb{X} = \begin{bmatrix} \Phi_1(W_1) & \dots & \Phi_m(W_1) & X_1\cdot\Phi_1(W_1) & \dots & X_1\cdot\Phi_k(W_1) \\ \vdots & & \vdots & \vdots & & \vdots \\ \Phi_1(W_n) & \dots & \Phi_m(W_n) & X_n\cdot\Phi_1(W_n) & \dots & X_n\cdot\Phi_k(W_n) \end{bmatrix}. \tag{5}$$

On assume $n \ll p$. Nous faisons ainsi face à un problème de grande taille. Les vecteurs θ et η sont supposés *creux* avec des coefficients respectifs donnés par s_θ et s_η. Par conséquent, β est aussi *creux* (c'est-à-dire, $\sum_{i=1}^p I(\beta_j \neq 0) = s \ll p$) avec pour coefficient $s = s_\eta + s_\theta$.

Estimateurs pénalisés

Dans le chapitre 3, nous optons pour la méthode du *Lasso* afin de trouver un estimateur à notre paramètre d'intérêt. Le problème peut donc s'écrire

$$\hat{\beta} = \begin{pmatrix} \hat{\eta} \\ \hat{\theta} \end{pmatrix} = \arg\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2n}\sum_{i=1}^n\left(Y_i - (\mathbb{X}\beta)_i\right)^2 + \lambda\|\beta\|_1\right\}, \tag{6}$$

où $\|\beta\|_1 = \sum_{i=1}^p |\beta_i|$ et $\lambda > 0$ est un paramètre de réglage.

Au travers d'une *condition de compatibilité* (voire [67, 74]), on arrive à établir les vitesses de convergence ainsi que la consistance de notre estimateur $\hat{\beta}$, et par conséquent $\hat{\eta}$ et $\hat{\theta}$.

Il est important de se rappeler que notre nouvelle mesure d'importance de variable est donnée par $\bar{\Phi}_k(W)\cdot\theta_P$. Ce faisant, il peut être avantageux à bien des égards de considérer un problème d'optimisation dont la pénalité s'appuie uniquement sur θ. Ceci nous permettra ainsi d'avoir une vision plus appropriée de l'estimateur de la fonction d'intérêt. Avec ce cadre, m est supposé largement inférieur à n, mais l'on conserve $n \ll k$.

On dénomme par Ψ_m, la matrice donnée par les m premières fonctions de la base $\{\Phi\}_{j=1}^\infty$ au points $W_1, \dots, W_n$. Elle correspond à

$$\Psi_m = \begin{bmatrix} \Phi_1(W_1) & \dots & \Phi_m(W_1) \\ \vdots & & \vdots \\ \Phi_1(W_n) & \dots & \Phi_m(W_n) \end{bmatrix}$$

Notre nouveau problème d'optimisation peut donc s'écrire

$$\begin{pmatrix} \hat{\eta} \\ \hat{\theta} \end{pmatrix} = \arg \min_{\eta \in \mathbb{R}^m, \theta \in \mathbb{R}^k} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left[Y_i - (\Psi_m^i \cdot \eta + X_i \cdot (\Psi_k^i \cdot \theta)) \right]^2 + \lambda \|\theta\|_1 \right\}. \tag{7}$$

Grâce à une *condition de compatibilité*, nous arrivons à établir les vitesses de convergence de nos estimateurs $\hat{\eta}$ et $\hat{\theta}$, de même que leur consistance. Ceci améliore les vitesses d'estimation de θ, lié au modèle (6).

Estimateurs pénalisés avec erreurs sur les variables

Nous savons que dans de nombreux cas de la vie réelle, les données observées sont en général polluées. Ceci a été pris en considération dans la seconde phase de notre analyse. Précisément, nous assumons que le modèle (4) est soumis à des erreurs additives sur les variables correspondantes à $Z = X + \nu$, où les éléments de ν sont extraits d'une distribution centrée gaussienne de variance μ^2. Nous introduisons ainsi le nouveau modèle caractérisé par

$$\begin{cases} Y &= \mathbb{X} \cdot \beta + \epsilon \\ \mathbb{Z} &= \mathbb{X} + \mathbb{K} \end{cases}, \tag{8}$$

où $\beta = \begin{pmatrix} \eta \\ \theta \end{pmatrix}$, et $\mathbb{K}$ est donné par

$$\mathbb{K} = \begin{bmatrix} 0 & \dots & 0 & \nu_1 \cdot \Phi_1(W_1) & \dots & \nu_1 \cdot \Phi_k(W_1) \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \dots & 0 & \nu_n \cdot \Phi_1(W_n) & \dots & \nu_n \cdot \Phi_k(W_n) \end{bmatrix}.$$

La présence d'erreurs sur les variables a naturellement un impact sur la formulation du problème d'optimisation. Il s'écrit dorénavant sous la forme

$$\hat{\beta} \in \arg \min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2} \beta^\top \Gamma \beta - \gamma^T \beta + \lambda \|\beta\|_1 \right\} \quad \text{tel que } \|\beta\|_1 \leq c_0 \sqrt{s},$$

où $\Gamma = \frac{1}{n} \widetilde{\mathbb{X}}^\top \widetilde{\mathbb{X}} - \mu^2 \zeta^\top \zeta$, $\gamma = \frac{1}{n} \widetilde{\mathbb{X}}^\top Y$, $\widetilde{\mathbb{X}} = [\Psi_m, D_Z \Psi_k]$, $\xi = (0, \dots, 0, 1, \dots, 1)$, un vecteur de taille p et $c_0 > 0$ une constante relativement grande.

En s'appuyant sur une condition dénommée *Restricted Eigenvalues* (voire [8]), nous pouvons établir de nouvelles vitesses de convergence, ainsi que la consistance de nos estimateurs, sous certaines conditions bénignes.

Tout comme dans le cas précédent, il est intéressant d'étudier le problème d'optimisation en pénalisant uniquement la variable θ. À cet effet, le nouveau problème d'optimisation s'écrit donc

$$\hat{\theta} \in \arg \min_{\|\theta\|_1 \leq c_0 \sqrt{s_\theta}} \left\{ \frac{1}{2} \theta^\top \Gamma \theta - \gamma^\top \theta + \lambda \|\theta\|_1 \right\}.$$

En s'appuyant de nouveau sur la condition *Restricted Eigenvalues*, mentionnée ci-dessus, nous établissons également la consistance et les vitesses de convergence de nos estimateurs sous certaines conditions.

Estimateurs Slope avec erreurs sur les variables

Dans le chapitre 4, nous généralisons le modèle de régression (8) et introduisons le modèle

$$\begin{cases} Y = \bar{X} \cdot \beta^\star + \epsilon \\ W = \bar{X} + U \end{cases} \tag{9}$$

où $Y = (Y_1, \dots, Y_n)^\top$ est un vecteur de réponses, $\bar{X} = \left[\bar{X}_{ij}\right]_{1 \le i \le n, 1 \le j \le p}$ est une $n \times p$ matrice de design et $\epsilon = (\epsilon_1, \dots, \epsilon_n)^\top$ est un vecteur gaussien dont les éléments sont indépendants et identiquement distribués, de variance σ^2. Au-delà, $W = [W_{ij}]_{1 \le i \le n, 1 \le j \le p}$ est une $n \times p$ matrice de design polluée par $U \in \mathbb{R}^{n \times p}$. On suppose que les lignes de la matrice U sont extraites indépendamment au travers d'une gaussienne centrée et multivariée ayant une matrice de variance covariance de taille $p \times p$ dénommée C_U.

Notre objectif est d'estimer le paramètre d'intérêt $\beta^\star$. Il est supposé *s-parcimonieux*, tel que $0 < s \ll p$. On rappelle que les données sont en grande dimension et que l'échantillon de données de taille n est bien plus petit que p. Les variables ϵ et U sont supposées indépendantes. L'estimateur *Slope* $\hat{\beta}$ de $\beta^\star$ est donné par

$$\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \left\{ \beta^\top \tilde{\Sigma} \beta - \frac{2}{n} Y^\top W \beta + \|\beta\|_\star \right\}, \tag{10}$$

où

$$\tilde{\Sigma} \in \arg\min_{M \in \mathcal{S}_{\ge 0}} \|\hat{\Sigma} - M\|_2 \quad avec \quad \hat{\Sigma} = \frac{1}{n} W^\top W - C_U,$$

et l'ensemble $\mathcal{S}_{\ge 0}$ est un ensemble de matrices symétriques définies positives découlant d'une projection au travers de la norme de *Frobenius*. La norme *Slope* est donnée par $\|\beta\|_\star = \sum_{i=1}^p \lambda_i |\beta|_{(i)}$, où $\lambda_1 \ge \dots \ge \lambda_p > 0$ sont des paramètres de réglage.

En utilisant la projection $\tilde{\Sigma}$ au lieu de $\hat{\Sigma}$, nous assurons la convexité du critère (10). Au-delà, nous évitons de restreindre l'ensemble où β varie. Nous obtenons une vitesse de convergence de $\hat{\beta}$ sous une condition de compatibilité et en utilisant les propriétés de la covariance empirique. Nous énonçons des conditions suffisantes sur les erreurs (relativement restrictives) qui permettent d'atteindre les vitesses optimales, contrairement au *Lasso*.

Chapter 1

Introduction

During the last couple of decades, we have witnessed an unprecedented development in technology. It has ultimately allowed us to collect a tremendous amount of data. Commonly called *the era of big data*, its impact looms large, touching a wide variety of scientific fields such as economics, biology and medicine. Its repercussions on our daily life are becoming more and more visible since it is affecting a large portion of the *"real" economy*. As examples, we can mention the health-care and automotive industries. The potential of this *era* has not yet been fully harnessed. One of the reasons, for this state of affairs, is that it comes with a lot of challenges. As far as statisticians are concerned, one of these challenges is the *curse of dimensionality*. It has prompted an extensive amount of research. The outcomes of these studies, in many cases, led to the extension of results that were proven true in low dimension settings. Furthermore, the high dimensionality of available data has also brought to light the need to carefully choose *adequate covariates*, when defining a model. As such, several tools have been introduced among which the concept of variable importance measure. It is at the core of this thesis.

Overview of our contributions. Our study of variable importance measure is achieved under two main frameworks. On the one hand, we define and study a new nonparametric, finite-dimensional, covariate-adjusted variable importance measure. It is nonparametric in the sense that it is defined as the value $\psi_f(P_0)$ at the law P_0 of the data of a functional ψ_f, defined on a nonparametric set $\mathcal{M}$ of laws subject to mild constraints. Here, f is a user-supplied function. In Chapter 2, we develop and analyze an estimator based on the targeted minimum loss estimation methodology. Under mild assumptions, the estimator lends itself to the construction of confidence regions of given asymptotic confidence level and to the derivation of asymptotic p-values for related hypothesis testing.

On the other hand, drawing inspiration from the definition of ψ_f, we introduce a second

high-dimensional functional ψ_{pen}, defined on a set $\mathcal{M}' \subset \mathcal{M}$ characterized within $\mathcal{M}$ by a regression model. As opposed to ψ_f, ψ_{pen} does not rely on a user-supplied function. Assuming that P_0 falls in $\mathcal{M}'$, $\psi_{pen}(P_0)$ is well defined and a subvector thereof can be interpreted as another variable importance measure akin to $\psi_f(P_0)$, and equal to it for a particular choice of f. In Chapters 3 and 4, we develop and study two estimators of $\psi_{pen}(P_0)$, based respectively on the Lasso and the Slope methodologies. Under assumptions that are typical of this kind of framework, we derive some convergence rates. Simulation studies illustrate the theoretical results.

1.1 Variable importance measure

The analysis of *variable importance* has been simultaneously developed in several fields such as *engineering, economics, biology, statistics* and *machine learning*, to only mention those.

The *era of big data* has made these techniques of paramount importance. The authors in [81] provide a rather exhaustive list of *variable importance measures (VIM)*, covering a broad spectrum of disciplines such as astronautic engineering, chemistry and environmental science. They argue that these *measures* could be regrouped in seven main categories: *Difference-based VIMs* (i.e.: Derivatives-based methods [65]), *Parametric regression techniques* (i.e.: correlation coefficient), *Nonparametric regression techniques* (i.e.: Generalized Additive Model), *Hypothesis test techniques* (i.e.: entropy based measure [71]), *Variance based VIMs* [42], *Moment-independent VIMs* [12] and *Graphs VIMs*[11].

There is not an unified definition for the notion of *variable importance measure*. However, within the predictive model world, it can be viewed as a scale which helps to determine the dependence of an outcome of a regression model upon a single or a set of input variables. We can cite as an example *Random Forest* [15], which is a well-known machine learning algorithm for regression and classification. It relies on two main measures to evaluate a *variable importance*. The first one, called *Mean Decrease Impurity importance (MDI)* or *Mean Decrease Gini*, is based on the aggregation of the *Gini Impurity* across all nodes of the decision tree, which helps to determine how to split the data at each node. The second one, called *Permutation Importance*, is also based on the aggregation of a measure of accuracy of the predictor, computed at each node of the tree. *Random Forest* and similar decision tree techniques have a lot of benefits. They are well design for high dimensional dataset. The construction of decision trees requires, at each node, only a subset of the entire dataset. Furthermore, *Random Forest* is well-known to still be robust when the dataset contains outliers.

Similar positive features can be presented regarding other *VIMs* of the same family. However, it is also important to emphasize that most of them do have a few drawbacks.

We know as an example that decision tree methods sometimes produce optimal predictors which have very few variables, given the initial dataset (see [10]). *Random Forest* is known for over-fitting learning datasets that are particularly noisy. Furthermore, most *VIMs* techniques are known not to allow users to infer p-values, for hypothesis testing, from the quantified *variable importance*. Some of these methods have been questioned by researchers because they do not faithfully represent the *true importance* of a variable (see [54]).

1.2 The variable importance measure ψ_f over model $\mathcal{M}$

Significant breakthroughs were achieved in the article [75] within the field of *variable importance measure* for statistical problems. The author proposed to define the latter measure as a sound statistical parameter. It could then be studied through adequate estimators. One could therefore infer asymptotic properties for these estimators, hence easing the derivation of p-values for hypothesis testing. This work inspired several subsequent articles, with the most important for us being [22], on which our work is based.

Let us consider a statistical problem with a data structure $\mathcal{O} = (X, W, Y)$ of an experimental unit of interest $W \in \mathcal{W} \subset \mathbb{R}^d$, representing the vectors of covariates, $X \in \mathbb{R}$ (a continuous exposure) a real valued variable of cause affecting $Y \in \mathbb{R}$ (a continuous response), a real valued variable of effect. We denote by P_0 the true data-generating distribution of the data structure $\mathcal{O}$. The *exposure* is assumed to feature a reference level x_0. In words, there exists $0 < c < 1/2$ such that $P_0(X \neq x_0|W) \in [c, 1-c]$ P_0-almost surely.

Our objective is to quantify the relationship, if any, that exists between X and Y, while taking into account the covariates W. Taking the covariates in consideration is critical for us since we cannot rule out their impact on the relationship between the exposure and the response. By analogy with [22], we introduce a *variable importance measure* $\psi_f(P_0)$ characterized by

$$\psi_f(P) = \arg\min_{\beta \in \mathbb{R}^d} E_P\left[\left(Y - E_P(Y|X = x_0, W) - (X - x_0) f_\beta(W)\right)^2\right], \tag{1.1}$$

where $f_\beta(W) = \beta^\top \cdot f_{CT}(W)$ with $f_{CT} : \mathcal{W} \to \mathbb{R}^d$ is a user-supplied function. The assumption here is that one could predefine the type of relationship that exists between the covariates. As an example, with $d = 2$, we could have $f_{CT}(W) = (W_1^2, W_2^2)$, where $W = (W_1, W_2)$. As such, we can infer that $(X - x_0) f_{\psi_f(P_0)}(W)$ is the best approximation of the form $(X - x_0) f_\beta(W)$ of $Y - E_P(Y|X = x_0, W)$. It is important to note here that $\psi_f(P_0)$ is multivariate since its size corresponds to the number of covariates. This feature is crucial since it allows us to have a *variable importance measure* which captures the role played by each covariate in the relationship between the exposure and the response.

We emphasize here that we do not assume that there exists β such that

$$Y = E_{P_0}(Y|X = x_0, W) + (X - x_0) f_\beta(W) + \epsilon,$$

where ϵ is an unknown noise. In words, $\psi_f(P_0)$ is universally defined.

1.2.1 The parameter of interest

In order to study the parameter of interest, we use a semiparametric estimation methodology called targeted minimum loss estimation (*TMLE*), by analogy with [22]. We need to identify and study its key properties in order to deploy the *semiparametric model theory*.

Influence Curve. a *semiparametric model* is a statistical model in which parameters of interest are both an Euclidean vector and an infinite-dimensional parameter (also called *nonparametric component*). When dealing with these models, one is usually interested in estimating the Euclidean vector.

To solve this problem, the theory of asymptotic efficiency, as developed for parametric models, was extended. As such, the notion of *semiparametric efficiency bounds* was introduced in [68] and further developed in [3, 9, 44, 55].

Keeping the above in mind, we can infer that the *influence function* plays a similar role as the *normalized score function* in parametric models (see [79]). However, the *influence function* is not unique. A corresponding function can be found through a projection on the closure of a linear span of the *tangent space* (see [79]). This function, result of a projection, is unique and is called the *efficient influence curve*. Through [9], we can establish that by knowing the *influence function* (or even better the *efficient influence function*) of an estimator, we can deduce its asymptotic distribution.

Now, considering (1.1), we assume without loss of generality that $x_0 = 0$. Let us denote by $\dot{f}_\beta = \frac{\partial f_\beta}{\partial \beta} = f_{CT}$, the gradient of f_β and $L_0^2(P)$ the set of measurable functions g such that $E_P[g^2] < \infty$ and $E_P[g] = 0$. We also denote by

$$\theta(P)(X,W) = E_P(Y|X,W), \quad g(P)(W) = P(X \neq 0|W),$$

$$\mu(P)(W) = E_P(X\dot{f}_\beta(W)|W) \quad \text{and} \quad \Sigma(P) = E_P[X^2\dot{f}_\beta(W)\dot{f}_\beta(W)^\top],$$

some relevant features of the distribution P. If we assume that

- all features of P are well-defined,
- Σ is invertible,
- there exists a $c \in]0, 1/2[$ such that $g(P)(W) \in]c, 1-c[$ P-almost surely,

then the solution to (1.1) is unique and is given by

$$\psi_f(P) = \Sigma(P)^{-1}\left[E_P\left(X\dot{f}_\beta(W)\left(\theta(P)(X,W) - \theta(P)(0,W)\right)\right)\right].$$

Furthermore, we can also infer that the functional ψ_f is path-wise differentiable at P. Let us consider a bounded function $s \in L_0^2(P)$ and $\epsilon \in \mathbb{R}^d$ such that $\|s\|_\infty < \infty$ and $\|\epsilon\|_\infty < \|s\|_\infty^{-1}$. We can characterize a distribution $P_\epsilon \in \mathcal{M}$ such that

$$dP_\epsilon(\mathcal{O}) = \left(1 + \epsilon^\top s(\mathcal{O})\right) dP(\mathcal{O}), \tag{1.2}$$

If P_ϵ verifies the conditions defined above, then $\psi_f(P_\epsilon)$ is differentiable at $\epsilon = 0$ and its derivative is given by

$$\lim_{\epsilon \to 0} \frac{\psi_f(P_\epsilon) - \psi_f(P)}{\epsilon} = E_P\left[s(\mathcal{O})^\top D^\star(P)(\mathcal{O})\right],$$

where $D^\star(P)$ is the *efficient influence curve*, given by

$$\begin{aligned} D^\star(P) = {} & \Sigma(P)^{-1}\left[\theta(P)(X,W) - \theta(P)(0,W) - X f_\beta(W)\right] X\dot{f}(W) \\ & + \Sigma(P)^{-1}\left[(Y - \theta(P)(X,W))\left(X\dot{f}(W) - \frac{1_{X=0}}{g(P)(0|W)}\mu(P)(W)\right)\right]. \end{aligned}$$

We note here that this influence curve enjoys a key feature. It is **double-robust** (see [22]) in the sense that for any $(P, P') \in \mathcal{M}^2$, if either $(\mu(P') = \mu(P)$ and $g(P') = g(P))$ or $\theta(P')(0,\cdot) = \theta(P)(0,\cdot)$ holds then $PD^\star(P') = 0$ implies that $\psi_f(P') = \psi_f(P)$. This property is of importance because it plays a crucial role in proving the consistency and asymptotic property of the estimator of $\psi_f(P_0)$.

1.2.2 Statistical inference procedure

Having laid the groundwork of the underlying theory of the estimation methodology, we turn our attention to the statistical inference procedure. It is at the heart of the numerical analysis. The procedure is based on n independent copies $\mathcal{O}^{(i)}$, with $i \in \{1, \ldots, n\}$ of the observed data structure $\mathcal{O}$. As mentioned above, the estimation of $\psi_f(P_0)$ is based on the *TMLE* methodology which can be divided in two main steps. In the first step, an initial estimate $\psi_f(P_n^0)$ is evaluated, where $P_n^0 \in \mathcal{M}$ is built as an element of $\mathcal{M}$ based on $\mathcal{O}^{(1)}, \ldots, \mathcal{O}^{(n)}$.

It is important to note that this initial estimate can be biased. It does not constitute a drawback for the rest of the procedure, since this bias, if it exists, is corrected with subsequent updates. We also emphasize that it is not required to estimate the whole distribution P_0 for our objective. In fact, only a few features of P_0 must be estimated, namely : $\mu(P_0)$, $g(P_0)$, $\Sigma(P_0)$ and $\theta(P_0)$.

In the second step, using (1.2), we then construct k successive updates $P_n^1, \ldots, P_n^k$ which allow us to build a sequence of estimates $\{\psi_n^j = \psi_f(P_n^j)\}_{j=1,\ldots,k}$. For k large enough, the procedure converges. In Chapter 2 (see 2.1), we then derive the following result:

Theorem 1.1. *Let us denote $\psi_0 = \psi_f(P_0)$. Suppose that performing k_n iterations of the updating procedure guarantees that $\|P_n D^\star(P_n^{k_n})\|_\infty = o_P(\frac{1}{\sqrt{n}})$. Suppose moreover that there exists a function f_1 with $P_0 f_1 = 0$ such that $\|P_0(D^\star(P_n^{k_n}) - f_1)(D^\star(P_n^{k_n}) - f_1)^\top\|_\infty = o_P(1)$, and that $\|\psi_f(P_n^{k_n})\psi_f(P_0) - P_0 D^\star(P_n^{k_n})\|_\infty = o_P(\frac{1}{\sqrt{n}})$. In addition, suppose that S_n estimates consistently $E_{P_0}[f_1(\mathcal{O})f_1(\mathcal{O})^\top]$. Then, $\psi_n^\star = \psi_f(P_n^{k_n})$ satisfies $\sqrt{n}(\psi^\star - \psi_0) = (P_n - P_0)f_1 + o_P(1)$, hence $\sqrt{n}S_n^{-\frac{1}{2}}(\psi_n^\star - \psi_0)$ converges in law to the d-multivariate Gaussian law with zero mean and identity covariance matrix.*

This result proves the consistency of the *TMLE* estimator and its asymptotic normality. It therefore allows one to construct confidence regions given a confidence level. Furthermore, one can derive the parameter of interest p-values for hypothesis testing.

1.3 Going beyond ψ_f within a submodel of $\mathcal{M}$

Let us consider the data structure $\mathcal{O} = (X, W, Y)$ of an experimental unit of interest, where $X \in \mathbb{R}$ is an exposure, $Y \in \mathbb{R}$ is a response and $W \in \mathcal{W} \subset \mathbb{R}^d$ represents our covariates, assumed drawn from the true data generating distribution $P_0 \in \mathcal{M}'$, where $\mathcal{M}' \subset \mathcal{M}$ and $\mathcal{M}'$ is characterized by $(\eta_P^\star, \theta_P^\star) \in \mathbb{R}^m \times \mathbb{R}^k$ with potentially high dimensions m and k, such that

$$Y = \bar{\Phi}_m^\star(W)\eta_P^\star + X \cdot \bar{\Phi}_k^\star(W)\theta_P^\star + \varepsilon,$$

where $\bar{\Phi}_j^\star = \{\Phi_1^\star, \ldots, \Phi_j^\star\}$ is a vector containing the first j elements of an orthonormal basis (*o.n.b*) of a *Hilbert space*. We recall that (1.1) is given by

$$\psi_f(P) = \arg\min_{\beta \in \mathbb{R}^p} E_P\left[\left(Y - E_P(Y|X = x_0, W) - (X - x_0) \cdot f_\beta(W)\right)^2\right],$$

where $f_\beta(W) = \beta \cdot f_{CT}(W)$, with f_{CT} a function provided by the user. We chose to have f_β linear in β for simplicity. It allowed a clearer path towards strong convergence rates results. Furthermore, we emphasize here that the function f_{CT} plays the role of representing the interactions that exist between the covariates. Strictly speaking, it is an unknown feature of our problem which is, in most cases, a source of considerable complexity. In Chapter 3, we decide to analyze the problem through different lenses. We consider the same variable importance measure without requesting that f_β be specified by the user.

Our objective remains to quantify the relationship that exists between the response Y and the exposure X (having a reference level x_0), while taking into account the covariates

W. Let us denote by $\mathcal{F}$, the nonparametric set of all continuous functions on $\mathbb{R}^d$. We introduce, by analogy with (1.1), a statistical parameter $\psi_{pen}(P)$ characterized by

$$\psi_{pen}(P) = \arg\min_{f\in\mathcal{F}} E_P\left[(Y - E_P(Y|X = x_0, W) - (X - x_0)\cdot f(W))^2\right]. \tag{1.3}$$

It is important to note here that the parameter of interest is now a function. As such, we are aiming at introducing a more flexible way of characterizing the interactions between the covariates. It allows us to obtain a complete picture of the *variable importance measure* with respect to the covariates in the dataset. From now on, we assume $x_0 = 0$, without loss of generality.

1.3.1 The function of interest

Nonparametric regression models have been extensively studied in the literature. Several groups of techniques used to solve these problems have emerged through the years. We can mention as examples the *Kernel smoothing* method [73], *k-nearest-neighbors* (see [1, 37] - with few variations such as the *weighted k-nearest-neighbors regression* [63]), *Wavelets* (see [51]), *Smoothing spline* (see [64, 73]) and *nonparametric least squares*. We rely upon this last technique for what follows.

Let us consider $\Phi = (\Phi_1, \ldots, \Phi_p, \ldots) = \{\Phi_j\}_{j=1}^{\infty}$ a complete orthonormal base of $\mathcal{H}$, a *Hilbert space*. Given that $P \in \mathcal{M}'$, we know that there exists $(\eta_P, \theta_P) \in \mathbb{R}^m \times \mathbb{R}^k$ such that, under P,

$$Y = \bar{\Phi}_m(W)\eta_P^\top + X\cdot\bar{\Phi}_k(W)\theta_P + \varepsilon, \tag{1.4}$$

where ε is a centered Gaussian random variable with variance σ^2 that is independent of (X, W), and $\bar{\Phi}_j = \{\Phi_1, \ldots, \Phi_j\}$ represents a vector whose elements are the first j components of an *o.n.b.* of $\mathcal{H}$. Furthermore, $g(P)(W) = E_P[Y|X = 0, W] = \bar{\Phi}_m(W)\eta_P$ and, if one chooses $f(W) = \bar{\Phi}_k(W)\theta_P$, then $\psi_f(P) = \theta_P$. We emphasize here that by analogy to (1.1), $\bar{\Phi}_k(W)\cdot\theta_P$ is our new *variable importance measure*. Furthermore, we note that the whole base $\{\Phi_j\}_{j=1}^{\infty}$ is not used in our decomposition. As such, there are potentially residual terms which one would want to take into consideration. We assume here that they are null under some smoothness conditions.

It is important to note that in what follows, we do not exploit the form of $\psi_{pen}(P)$ given by (1.3), but rather rely on the regression model (1.4) for its analysis. The indices $\{m, k\}$ are assumed known and increase with n, but are function of P. The latter dependence is not problematic in this study since we only consider a single data generating distribution P.

We can now rewrite (1.4) as

$$Y = \mathbb{X}\cdot\beta + \epsilon, \tag{1.5}$$

where $\beta = \begin{pmatrix} \eta \\ \theta \end{pmatrix} \in \mathbb{R}^p$ with $p = m + k$ and $\mathbb{X}$ is given by

$$\mathbb{X} = \begin{bmatrix} \Phi_1(W_1) & \dots & \Phi_m(W_1) & X_1 \cdot \Phi_1(W_1) & \dots & X_1 \cdot \Phi_k(W_1) \\ \vdots & & \vdots & \vdots & & \vdots \\ \Phi_1(W_n) & \dots & \Phi_m(W_n) & X_n \cdot \Phi_1(W_n) & \dots & X_n \cdot \Phi_k(W_n) \end{bmatrix}. \tag{1.6}$$

We assume $n \ll p$. As such, our problem is in high dimension. Giving this setting, the vectors θ and η are assumed sparse with respective coefficients given by s_θ and s_η. As a result, β is also sparse (in essence: $\sum_{i=1}^{p} I(\beta_j \neq 0) = s \ll p$) with a sparsity coefficient denoted by $s = s_\eta + s_\theta$.

Nonparametric Least square. Let us consider the statistical problem (1.5). A *nonparametric least square* estimate of β is given by

$$\arg\min_{\beta \in \mathbb{R}^p} \|Y - \mathbb{X}\beta\|_2. \tag{1.7}$$

In high dimensional settings, we know that the matrix $\mathbb{X}^\intercal\mathbb{X}$ cannot be positive definite. As such, (1.7) does not admit a unique solution.

One way of solving this issue in the literature has been to target a set of potential solutions $\beta^\star$ which are *sparse*. This sparsity assumption implies that we do not want to have values of $\beta^\star$ that are large, even if they lead to better fit. It is then common to adjust the problem (1.7) by introducing a penalty term such that it is now given by

$$\hat{\beta}^{LASSO} = \arg\min_{\beta \in \mathbb{R}^p} \|Y - \mathbb{X}\beta\|_2 + pen(\beta), \tag{1.8}$$

where $pen(\cdot)$ is a real-valued function of β. Several alternatives have been considered for the choice of $pen(\cdot)$. The most natural one appeared to be $pen(\beta) = \lambda \cdot \|\beta\|_0$, where $\lambda > 0$ is a tuning parameter and $\|\cdot\|_0$ represents the total number of non-zero elements. However, with this choice of penalization problem, the resulting problem is impossible to compute in polynomial time (it is NP-hard). The Lasso estimator, introduced in [72], has opted for a convex relaxation of the penalty term with $pen(\beta) = \lambda \cdot \|\beta\|_1$, where $\|\cdot\|_1$ is the l_1-norm. The authors in [4] have shown that there exists a $\lambda^\star$ such that the *Lasso estimator* reaches the optimal minimax rates of prediction and estimation, under some conditions on the design matrix $\mathbb{X}$.

We note that $\lambda^\star$ remains a function of the sparsity coefficient s of β, which is unknown in practice. A solution that has been explored to alleviate this difficulty is to consider the *Slope estimator* (see [50]). It relies on a penalty term given by $pen(\beta) = \sum_{j=1}^{p} \lambda_j |\beta|_{(j)}$, where the tuning parameters verifies $\lambda_1 \geq \dots \geq \lambda_p > 0$ and $|\beta|_{(j)}$ is the j^{th} largest component of $|\beta|$. With this form, the objective is to penalize more the larger components of β, compare

to the smaller ones. The authors in [4] have shown that there exist tuning parameters $\{\lambda_i^\star\}_{i=1,\dots,p}$ such that the *Slope estimator* attains the optimal minimax rates of estimation and prediction.

1.3.2 The optimization problems and corresponding convergence rates using Lasso

In Chapter 3, we opt for the *Lasso* methodology to find the estimator of the parameter of interest. As mentioned above, it appears natural to use the penalty function $pen(\beta) = \lambda\|\beta\|_1$ for our optimization problem. However, we recall that $\bar{\Phi}_k(W)\cdot\theta_P$ is now our new *variable importance measure*. One can then prefer to use the penalty function $pen(\theta) = \lambda\cdot\|\theta\|_1$. By only penalizing θ_P, our variable of interest, it might be possible to infer a better estimator. Both cases are explored below.

Penalty function based on β.

The optimization problem is given by

$$\hat{\beta} = \begin{pmatrix} \hat{\eta} \\ \hat{\theta} \end{pmatrix} = \arg\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2n}\sum_{i=1}^{n}(Y_i - (\mathbb{X}\beta)_i)^2 + \lambda\|\beta\|_1\right\} \tag{1.9}$$

where $\|\beta\|_1 = \sum_{i=1}^p |\beta_i|$ and $\lambda > 0$ is a tuning parameter. We rely on the well-known *compatibility condition* (see [67, 74]). It stipulates that, on an index set $\mathcal{S}_0$ having s_0 elements, if for some $\phi_0 > 0$ and all $\gamma \in \mathbb{R}^p$ such that $\|\gamma_{\mathcal{S}_0^c}\|_1 \le 3\|\gamma_{\mathcal{S}_0}\|_1$, it holds that

$$\|\gamma_{\mathcal{S}_0}\|_1^2 \le \frac{s_0\|\mathbb{X}\gamma\|_2^2}{n\phi_0^2}.$$

It plays an important role in establishing the convergence rate of the estimator $\hat{\beta}$. In Chapter 3 (see 3.1), we then derive the following result:

Proposition 1.2. *Assume that the variable X takes values in a compact space. Furthermore, assume that the compatibility condition holds. Then, with probability larger than $1 - 2\exp(-\frac{t^2}{2})$ and $\lambda = 2\sigma\|X\|_\infty\sqrt{\frac{t^2+2\log(p)}{n}}$ with arbitrary $t > 0$, the estimator $\hat{\beta}$ of (1.9) verifies*

$$\frac{1}{2n}\|\mathbb{X}\left(\hat{\beta}-\beta\right)\|_2^2 + \lambda\|\hat{\beta}-\beta\|_1 \le 4\sigma^2\|X\|_\infty^2\frac{8s}{\phi_0}.\frac{t^2+2\log(p)}{n}.$$

As a consequence, with the same probability,

$$\|\hat{\beta}-\beta\|_1 \le 2\sigma\|X\|_\infty\cdot\frac{8s}{\phi_0^2}\cdot\sqrt{\frac{t^2+2\log(p)}{n}}.$$

We can then infer similar results for the estimator $\hat{\eta}$, and ultimately, $\hat{\theta}$ the estimator of our parameter of interest.

Penalty function based on θ.

This framework relates to a partially linear model with high-dimensional linear component, i.e. m is assumed smaller than n, but we keep $n \ll k$. We put more emphasis on the estimation of θ as it is a key component of our parameter of interest. Let us denote by Ψ_m, the matrix given by the first m functions of the base $\{\Phi\}_{j=1}^{\infty}$ at points $W_1, \ldots, W_n$, that is,

$$\Psi_m = \begin{bmatrix} \Phi_1(W_1) & \ldots & \Phi_m(W_1) \\ \vdots & & \vdots \\ \Phi_1(W_n) & \ldots & \Phi_m(W_n) \end{bmatrix}$$

and by D_X the diagonal matrix with elements $\{X_1, \ldots, X_n\}$. Thus, $\mathbb{X} = [\Psi_m, D_X\Psi_k]$. The new optimization problem can now be written as

$$\begin{pmatrix} \hat{\eta} \\ \hat{\theta} \end{pmatrix} = \underset{\eta\in\mathbb{R}^m, \theta\in\mathbb{R}^k}{\arg\min} \left\{ \frac{1}{2n} \sum_{i=1}^{n} \left[Y_i - (\Psi_m^i \cdot \eta + X_i \cdot (\Psi_k^i \cdot \theta)) \right]^2 + \lambda \|\theta\|_1 \right\}. \tag{1.10}$$

We solve this problem in two steps. In the first step, we solve for $\hat{\eta}$ which minimizes (1.10) for any θ.

It corresponds to a *classic least square estimator* whose solution is given by

$$\hat{\eta}(\theta) = \frac{1}{n}\Psi_m^\top \cdot (Y - (D_X\Psi_k)), \text{ where } Y = \{Y_i\}_{i=1,\ldots,n}.$$

In the second step, we replace the solution of the previous solution in (1.10), and then obtain a resulting classic l_1 optimization problem solely based on θ and given by

$$\hat{\theta} = \arg\min_{\theta\in\mathbb{R}^k} \left\{ \frac{1}{n} \|V \cdot \left(Y - \tilde{\Psi}_k \cdot \theta\right)\|_2^2 + \lambda \|\theta\|_1 \right\}, \tag{1.11}$$

where $\lambda > 0$ is a tuning parameter, $V = \left(I_{n\times n} - \frac{1}{n}\Psi_m \cdot \Psi_m^\top\right)$ and $\tilde{\Psi}_k = D_X\Psi_k$. We are then able to deduce convergence rates of our estimator. In Chapter 3 (see 3.2), we then derive the following result:

Proposition 1.3. *Let us assume that the compatibility condition holds. Thus with probability larger than* $1 - 2\exp(-\frac{t^2}{2})$ *and for* $\lambda = 2\sigma\|X\|_\infty \sqrt{\frac{t^2+2\log(k)}{n}}$*, the estimator* $\hat{\theta}$ *is such that*

$$\frac{1}{2n}\|V \cdot \tilde{\Psi}_k \cdot \left(\hat{\theta} - \theta\right)\|_2^2 + \lambda\|\hat{\theta} - \theta\|_1 \leq 4\sigma^2\|X\|_\infty^2 \cdot \frac{8s_\theta}{\phi_0^2} \cdot \frac{t^2 + 2\log(k)}{n}$$

for some arbitrary $t > 0$*. Hence,*

$$\|\hat{\theta} - \theta\|_1 \leq 2\sigma\|X\|_\infty \frac{8s_\theta}{\phi_0^2} \cdot \sqrt{\frac{t^2 + 2\log(k)}{n}}.$$

This result shows an improvement in the estimation of θ. However, we can notice that the estimation rate of η deteriorates with respect to the case where θ and η are estimated at the same time.

1.3.3 Adding errors-in-variables and their impact on convergence rates using Lasso

We know that in most real life examples, the dataset used for statistical analysis is polluted. We have taken the latter fact in consideration in this second phase of our analysis. Specifically, we assume that the model (1.5) is subject to additive measurement errors corresponding to $Z = X + \nu$, where elements of ν are drawn from a centered Gaussian distribution with variance μ^2. The variables ϵ and ν are assumed independent. Thus, we introduce the model given by

$$\begin{cases} Y &= \mathbb{X} \cdot \beta + \epsilon \\ \mathbb{Z} &= \mathbb{X} + \mathbb{K} \end{cases}, \tag{1.12}$$

where $\beta = \begin{pmatrix} \eta \\ \theta \end{pmatrix}$, and $\mathbb{K}$ is given by

$$\mathbb{K} = \begin{bmatrix} 0 & \dots & 0 & \nu_1 \cdot \Phi_1(W_1) & \dots & \nu_1 \cdot \Phi_k(W_1) \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \dots & 0 & \nu_n \cdot \Phi_1(W_n) & \dots & \nu_n \cdot \Phi_k(W_n) \end{bmatrix}.$$

We further consider that there exists $L > 0$ such that

$$\|\Psi_k(W)\theta\|_2^2 < L, \quad P\text{-almost surely.}$$

This assumption plays a key role in deriving the convergence rates of the estimators. Similarly to Section 1.3.2, two types of penalty functions are considered.

Penalty function based on β.

We note that the presence of measurement errors surely has an impact on the optimization problem as developed in (1.9). In fact,

$$\begin{aligned} E[\|Y - (\Psi_m\eta + (D_Z\Psi_k)\theta)\|_2^2] &= E[\|Y - (\Psi_m\eta + (D_X\Psi_k)\theta)\|_2^2] + E[\|(D_\nu\Psi_k)\theta\|_2^2] \\ &\quad - 2E[\epsilon^\top (D_\nu\Psi_k)\theta] \\ &= E[\|Y - (\Psi_m\eta + (D_X\Psi_k)\theta)\|_2^2] + n\mu^2\|\theta\|_2^2. \end{aligned}$$

From the above, it appears that the regularization requires a correction by a factor of $n\mu^2\|\theta\|_2^2$. As such, the estimators $\hat{\eta}$ and $\hat{\theta}$ are then defined by

$$\begin{pmatrix} \hat{\eta} \\ \hat{\theta} \end{pmatrix} = \underset{\eta\in\mathbb{R}^m, \theta\in\mathbb{R}^k}{\arg\min}\, \mathcal{L}(\eta, \theta)$$

with

$$\mathcal{L}(\eta,\theta) = \frac{1}{2n}\left\| Y - (\Psi_m \cdot \eta + (D_Z\Psi_k)\cdot\theta)\right\|_2^2 - \frac{1}{2}\mu^2\|\theta\|_2^2 + \lambda_1\|\eta\|_1 + \lambda_2\|\theta\|_1,$$

where $\lambda_1 \geq 0$ and $\lambda_2 \geq 0$.

We denote by $\widetilde{\mathbb{X}}$, the equivalent of $\mathbb{X}$ in (1.5) with X replaced by its counterpart Z. The matrix $\widetilde{\mathbb{X}}$ corresponds to $\widetilde{\mathbb{X}} = [\Psi_m, D_Z\Psi_k]$, where D_Z is a diagonal matrix with diagonal elements $Z_1,\ldots,Z_n$. We further consider the vector $\xi = (0,\ldots,0,1,\ldots,1)$, of size p, where the first m elements are equal to 0 and the last k are all equal to 1. Decomposing the loss function, we obtain:

$$\begin{aligned}\mathcal{L}(\eta,\theta) &= \frac{1}{2}\left(\frac{1}{n}\|\Psi_m\eta + (D_Z\Psi_k)\theta\|_2^2 - \mu^2\|\theta\|_2^2\right) - \frac{1}{n}\langle Y, \Psi_m\eta + (D_Z\Psi_k)\theta\rangle \\ &\qquad + \frac{1}{2n}\|Y\|_2^2 + \lambda_1\|\eta\|_1 + \lambda_2\|\theta\|_1 \\ &= \frac{1}{2}\left(\frac{1}{n}\|\widetilde{\mathbb{X}}\beta\|_2^2 - \mu^2\|\zeta\beta\|_2^2\right) - \frac{1}{n}\langle Y, \widetilde{\mathbb{X}}\beta\rangle + \lambda\|\beta\|_1 + \frac{1}{2n}\|Y\|_2^2 \quad \text{with } \lambda = \begin{pmatrix}\lambda_1\\ \lambda_2\end{pmatrix},\end{aligned}$$

where $\beta = \begin{pmatrix}\eta\\ \theta\end{pmatrix}$ and $\zeta = diag(\xi)$. Thus

$$\mathcal{L}(\eta,\theta) = \frac{1}{2}\beta^\top\Gamma\beta - \gamma^T\beta + \lambda\|\beta\|_1 + \frac{1}{2n}\|Y\|_2^2,$$

with $\Gamma = \frac{1}{n}\widetilde{\mathbb{X}}^\top\widetilde{\mathbb{X}} - \mu^2\zeta^\top\zeta$ and $\gamma = \frac{1}{n}\widetilde{\mathbb{X}}^\top Y$. The optimization problem becomes then

$$\hat{\beta} = \arg\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2}\beta^\top\Gamma\beta - \gamma^T\beta + \lambda\|\beta\|_1\right\}.$$

For $\mu \neq 0$, the matrix Γ is not positive definite. As such, we are dealing with a non convex quadratic problem. Furthermore, if Γ has negative eigenvalues, the problem is unbounded. Thus, we need to add constraints to the optimization problem (see 3.3.1). Therefore, the problem is now given by

$$\hat{\beta} \in \arg\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{2}\beta^\top\Gamma\beta - \gamma^T\beta + \lambda\|\beta\|_1\right\}, \quad \text{such that } \|\beta\|_1 \leq c_0\sqrt{s}, \tag{1.13}$$

where $c_0 > 0$ is large enough.

Furthermore, we use a well-known *Restricted Eigenvalues* condition (see [8]) on the matrix Γ. It stipulates that for some integer s, such that $1 \leq s \leq p$, $t_0 > 0$, a positive number c_0, the condition holds for the index set $J_0 \in \{1,\ldots,p\}$ and for all $\delta \in \mathbb{R}^1_\star$ such that $|J_0| \leq s$ and $\|\delta_{J_0^c}\|_1 \leq c_0\|\delta_{J_0}\|_1$, if we have

$$\frac{\delta^\top\Gamma\delta}{\|\delta_{J_0}\|_2^2} \geq t_0. \tag{1.14}$$

The above condition is crucial since it allows us to control the rank deficiency problem related to the definition of Γ. In Chapter 3 (see 3.3), we then derive the following result:

Theorem 1.4. *We assume that the variable X takes values in a compact space. Furthermore, we also assume that the Restricted Eigenvalues condition holds. Then, there exist $t > 0$, $t_0 > 0$ and $\lambda_0 = \left(\|X\|_\infty(\sigma + \mu\sqrt{L}) + \mu\sqrt{L}\right)\sqrt{\frac{2(1+t)\log(p)}{n}}\,\tau_0$, with*

$$\tau_0 = \max\left\{\frac{2\mu^2\sqrt{L}}{\sqrt{n}}\sqrt{t}, \frac{4\mu^2 L}{n}t\right\} + (\sigma^2+\mu^2)\cdot\max\left\{\sqrt{\frac{t}{n}}, 2\delta\frac{t}{n}\right\}$$

such that for $2\lambda_0 \le \lambda$, the estimators $\hat{\eta}$ and $\hat{\theta}$ verify

$$\|\hat{\beta}-\beta\|_1 = \|\hat{\theta}-\theta\|_1 + \|\hat{\eta}-\eta\|_1 \le 12\lambda\cdot\frac{s}{t_0}$$

with probability greater than $1 - 5\exp(-\frac{t}{2})$.

Penalty function based on θ.

As in Section 1.3.2, the resolution of the optimization problem is done in a two-step process. The index m is assumed smaller than n, but we keep $n \ll k$. We put more emphasis on the estimation of θ as it is a key component of our parameter of interest. In the first step, we solve for $\hat{\eta}$ since it corresponds to a *classic least square estimator*. In the second step, the optimization problem corresponding to the estimator $\hat{\theta}$ is adjusted due to the presence of measurement errors. It is tentatively given by

$$\hat{\theta} = \arg\min_{\theta\in\mathbb{R}^k}\left\{\frac{1}{2}\theta^\top\Gamma\theta - \gamma^\top\theta + \lambda\|\theta\|_1\right\}, \tag{1.15}$$

where $\Gamma = \frac{1}{n}(V\widetilde{\Psi}_k)^\top(V\widetilde{\Psi}_k) - \frac{1}{n}\mu^2(V\Psi_k)^\top(V\Psi_k)$ and $\gamma = \frac{1}{n}(V\widetilde{\Psi}_k)^\top Y$.

Again, Γ is singaular, we have to further constrain the problem. Hence, following [48], the optimization problem (1.15) is replaced by

$$\hat{\theta} \in \arg\min_{\|\theta\|_1\le c_0\sqrt{s_\theta}}\left\{\frac{1}{2}\theta^\top\Gamma\theta - \gamma^\top\theta + \lambda\|\theta\|_1\right\}$$

where $c_0 > 0$ is a large enough constant. We can then derive the convergence rate for our estimator (see 3.4).

Theorem 1.5. *We assume that the variable X takes values in a compact space. Furthermore, we assume that the Restricted Eigenvalues condition holds.*

Then, there exists $t > 0$, $t_0 > 0$ and $\lambda_0 = \|X\|_\infty\left(\sigma + \mu\sqrt{L}\right)\sqrt{\frac{2(1+t)\log(k)}{n}}\,\tau_0$, with

$$\tau_0 = \max\left\{\frac{2\mu^2\sqrt{L}}{\sqrt{n}}\sqrt{t}, \frac{4\mu^2 L}{n}t\right\} + (\sigma^2+\mu^2)\cdot\max\left\{\sqrt{\frac{t}{n}}, 2\delta\frac{t}{n}\right\}$$

such that for $\lambda_0 \le 2\lambda$, the estimator $\hat{\theta}$ verifies

$$\|\hat{\theta}-\theta\|_1 \le 12\lambda\frac{s_\theta}{t_0}$$

with probability greater than $1 - 5\exp(-\frac{t}{2})$.

1.3.4 Convergence rates using Slope

We now generalize the regression model (1.12) and consider the model

$$\begin{cases} Y = \bar{X} \cdot \beta^\star + \epsilon \\ W = \bar{X} + U \end{cases} \tag{1.16}$$

where $Y = (Y_1, \dots, Y_n)^\top$ is the vector of responses, $\bar{X} = \left[\bar{X}_{ij}\right]_{1 \le i \le n, 1 \le j \le p}$ is the $n \times p$ design matrix of covariates and $\epsilon = (\epsilon_1, \dots, \epsilon_n)^\top$ is a Gaussian vector having centered independent and identically distributed elements, with variance σ^2. Furthermore, $W = [W_{ij}]_{1 \le i \le n, 1 \le j \le p}$ is a $n \times p$ noisy design matrix, polluted by $U \in \mathbb{R}^{n \times p}$. We assume here that the rows of the matrix U are drawn independently from a centered Gaussian multivariate distribution with a $p \times p$ covariance matrix denoted by C_U.

Our objective is to estimate the parameter of interest $\beta^\star$. It is assumed s-sparse, with $0 < s \ll p$. We recall that the dataset is in high dimension and the sample size n is much smaller than p. The variables ϵ and U are assumed independent. The *Slope* estimator $\hat{\beta}$ of $\beta^\star$ is tentatively defined through the following optimization problem

$$\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{\sqrt{n}} \|Y - \bar{X} \cdot \beta\|_2^2 + \|\beta\|_\star \right\}, \tag{1.17}$$

with $\|\beta\|_\star = \sum_{i=1}^p \lambda_i |\beta|_{(i)}$, known as the *Slope* norm, where $\lambda_1 \ge \dots \ge \lambda_p > 0$ are tuning parameters. The presence of errors-in-variables requires a correction of the optimization problem (1.17).

As such, instead of (1.17) we tentatively consider

$$\begin{aligned} \hat{\beta} &\in \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{n} \|Y - W\beta\|_2^2 - \beta^\top C_U \beta + \|\beta\|_\star \right\} \\ &\in \arg\min_{\beta \in \mathbb{R}^p} \left\{ \beta^\top \left(\frac{1}{n} W^\top W - C_U\right)\beta - \frac{2}{n} Y^\top W \beta + \frac{1}{n} Y^\top Y + \|\beta\|_\star \right\} \\ &\in \arg\min_{\beta \in \mathbb{R}^p} \left\{ \beta^\top \hat{\Sigma} \beta - \frac{2}{n} Y^\top W \beta + \|\beta\|_\star \right\} \end{aligned}$$

The problem above is not convex because the matrix $\hat{\Sigma}$ could have negative eigenvalues. Thus, we propose to project $\hat{\Sigma}$ on the set $\mathcal{S}_{\ge 0}$ of symmetric positive semi-definite matrices using the *Frobenius* norm. Therefore, we introduce $\tilde{\Sigma}$ defined by

$$\tilde{\Sigma} \in \arg\min_{M \in \mathcal{S}_{\ge 0}} \|\hat{\Sigma} - M\|_F,$$

which is equivalent to

$$\tilde{\Sigma} \in \arg\min_{M \in \mathcal{S}_{\ge 0}} \|\hat{\Sigma} - M\|_2,$$

and finally consider the convex optimization problem given by

$$\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \left\{ \beta^\top \tilde{\Sigma} \beta - \frac{2}{n} Y^\top W \beta + \|\beta\|_\star \right\}. \tag{1.18}$$

In order to derive the bounds on the convergence rates of this estimator, we need to rely on the well-known *Weighted Restricted Eigenvalues* condition, introduced in [4] and applied to the design $\bar{X}$. It was a key factor in deriving our convergence rates.

The condition states that the design matrix $\bar{X}$ satisfied the $WRE(s, c_0)$ condition if

$$\max_{j=1,\dots,p} \|\bar{X}e_j\|_n \leq 1,$$

and

$$\kappa = \min_{\delta \in C_{WRE(s,c_0)}} \frac{\|\bar{X}\delta\|_n^2}{\|\delta\|_2} > 0 \tag{1.19}$$

where $C_{WRE(s,c_0)} = \{\delta \in \mathbb{R}^p : \|\delta\|_\star \leq c_0\|\delta\|_\star \sqrt{\sum_{j=1}^p \lambda_j^2}\}$ is a cone in $\mathbb{R}^p$. We show that under mild conditions and if $\bar{X}$ verifies the $WRE(s, c_0)$ condition, then there exists a $c_0 \in \mathbb{R}$ such that

$$\frac{\kappa}{2} = \min_{\delta \in C_{WRE(s,c_0)}} \frac{\delta^\top \tilde{\Sigma}\delta}{\|\delta\|_2^2} > 0 \tag{1.20}$$

with probability at least $1 - \frac{4}{n} - 2\exp(-\frac{t^2}{2})$, where $t > 0$. The convergence rate for our estimator $\hat{\beta}$ (see 4.1) is then given by

Theorem 1.6. *Let $s \in \{1, \dots, p\}$ and assume that the WRE condition holds. We choose the following tuning parameters*

$$\lambda_j = \begin{cases} \gamma' \sqrt{\frac{\log(2p/j)}{n}} & \text{if} \quad \sum_{j=1}^s \log(\frac{2p}{j}) \geq R_{n,p}^2 \\ \frac{R_{n,p}}{\sqrt{s}} & \text{if} \quad \sum_{j=1}^s \log(\frac{2p}{j}) < R_{n,p}^2 \end{cases},$$

and assume that $\delta_0 < e2^{-1/2}$ and

$$\gamma' \geq 2\sigma(4 + \sqrt{2}).$$

Then, with probability at least $1 - \frac{8}{n} - 4p^{1-2\log(2/\delta_0)} - \delta_0$, we have

$$\|\hat{\beta} - \beta^\star\|_2 \leq \max\left\{C_1\sqrt{\frac{s}{n}\log(\frac{2ep}{s})}, C_2\sqrt{\frac{log^2(1/\delta_0)}{sn\log(2ep/s)}}, C_3 R_{n,p}\right\}$$

where $t = \sqrt{2\log(2/\delta_0)}$, $C_1 = \frac{(3+4\sigma)\gamma'}{\kappa}$, $C_2 = 4(4+\sqrt{2})^2$ and $C_3 = \frac{28}{\kappa}$. Furthermore,

$$R_{n,p}(C_U, \|\beta^\star\|_2, t) = R_{n,p} = t^2\sqrt{\frac{\|C_U\|_2}{n}} + \|\beta^\star\|_2\left(2A_{n,p}(C_U) + 3tp\sqrt{\frac{2\log(p)}{n}\max_{1\leq j\leq p} C_U^{jj}}\right)$$

where

$$A_{n,p}(C_U) = c \cdot \max\left\{Tr(C_U)\frac{\log(pn)}{n}, \sqrt{Tr(C_U)\|C_U\|_2\frac{\log(pn)}{n}}\right\}.$$

.

We note that the convexity of the problem (1.18) allows to minimize overall possible values of β without restrictions. However, the rates are driven by the estimation rates of the large covariance matrix of the design in the convolution model. Under very restrictive conditions on the noise, we recover the optimal rates for estimating $\beta^\star$.

1.4 Computational contributions and simulation studies

1.4.1 Computational contributions

Our contributions are two-fold. On the one hand, we extended the *TMLE* package [20] so that it could cope with the estimation of the multivariate variable importance measure, that we developed in Section 1.2. The extension allows the user to derive confidence regions.

On the other hand, we have relied on a bespoke version of *coordinate descent* to evaluate the estimators of ψ_{pen}. In order to conduct our numerical studies, we could have chosen a well-known algorithm such as *Lars*([32]). However, a majority of the problems we were faced with assumed that our datasets were subject to measurement errors. This feature is not always dealt with adequately in most of the existing solving packages. So we decided to develop our own version of the *coordinate descent* algorithm, known to be very efficient. By fully implementing our procedures, we could then control our setup and easily include the intricacies of our problems. Furthermore, this choice allowed us to extend our procedures initially written in *R*, to *C++*, in order to improve the overall speed.

Coordinate descent. The intuition behind *Coordinate descent algorithm* is quite straightforward: apply *Newton Raphson* algorithm consecutively and sequentially to each element of a parameter of interest of the problem until convergence. It was first used to resolve *Lasso like* problems in [35], where it was called *Modified Newton Raphson*. The author analyzed the structures of bridge estimators and developed an overall approach to solve bridge regression (penalized regression with penalty function given by $\sum |\beta_j|^\gamma$ and $\gamma \geq 1$). We can also mention [28] where the authors studied a linear inverse problem assumed to have a sparse expansion. The corresponding quadratic regression problem was adjusted through weighted l^α penalties, where $1 \leq \alpha \leq 2$. However, it is truly the articles [34, 82] which have been a catalyst for a greater interest by the statistician community in the methodology. Since then, it has been utilized in different forms and for a large variety of domain: *Block coordinate descent*([46]), *Cyclic coordinate descent* ([46]) and *Stochastic coordinate descent* ([26]) are some of the known variations.

We recall here the regression model (1.16) is given by

$$\begin{cases} Y = \bar{X} \cdot \beta^\star + \epsilon \\ W = \bar{X} + U \end{cases}$$

where $\bar{X} \in \mathbb{R}^{n \times p}$ is a design matrix, ϵ is a centered Gaussian vector, $Y \in \mathbb{R}^n$ and $\beta \in \mathbb{R}^p$ is our parameter of interest. Furthermore, U is a $n \times p$ matrix whose elements are drawn from a p-multivariate centered Gaussian distribution of covariance $C_U \in \mathbb{R}^p$. The model is set in

a high dimensional framework with $n \ll p$. The *Lasso* optimization problem corresponding to the regression model above is defined by

$$\hat{\beta} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{n} \|Y - \bar{X} \cdot \beta\|_2^2 + \lambda \|\beta\|_1 \right\}, \tag{1.21}$$

where $\lambda > 0$. In order to account for the pollution of the dataset, (1.21) has to be adjusted. Instead of constructing $\hat{\beta}$ given by (1.21), we aim for

$$\begin{aligned} \hat{\beta} &= \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{n} \|Y - W \cdot \beta\|_2^2 - \beta^\top \cdot C_U \cdot \beta + \lambda \|\beta\|_1 \right\} \\ &= \arg\min_{\beta \in \mathbb{R}^p} \left\{ \bar{f}(\beta) + \bar{g}(\beta) \right\}, \end{aligned} \tag{1.22}$$

where $\bar{g}(\beta) = \lambda \|\beta\|_1$ and $\bar{f}(\beta) = \frac{1}{n} \|Y - W \cdot \beta\|_2^2 - \beta^\top \cdot C_U \cdot \beta$. In order to solve numerically for $\hat{\beta}$, using *coordinate descent algorithm*, one has to find the gradient of the functions $\bar{f}$ and $\bar{g}$. The gradient associated with the function $\bar{f}$ can be derived easily. However, the function $\bar{g}$ is not differentiable for all β. To solve this problem, the methodology relies on a sub-gradient theory to compute $\nabla \bar{g}$. We have

$$\partial_i \bar{g}(\beta) = \lambda \cdot \begin{cases} 1 & \text{if} \quad \beta_i > 0 \\ -1 & \text{if} \quad \beta_i < 0 \\ [-1, 1] & \textit{if} \quad \beta_i = 0 \end{cases}$$

The *coordinate descent algorithm* can then be described as follows: Let us denote by $\nabla_i h(\beta) = \nabla_i \bar{f}(\beta) + \nabla_i \bar{g}(\beta)$ the i^{th} component of the gradient $\nabla h(\beta) = \nabla \bar{f}(\beta) + \nabla \bar{g}(\beta)$. At each point of the iterative process, we update the i_k^{th} element of our estimate $\beta^k; \beta_{i_k}^k$ while fixing all the remaining elements. As such, (1.22) becomes a unidimensional minimization problem. We can then rely on the gradient $\nabla h(\beta^k)$ to produce an update of $\beta_{i_k}^k$. The procedure can be briefly resumed in the following pseudo code.

Algorithm 1.1 Coordinate descent for (1.22)

Set $k \leftarrow 0$ and choose $\beta^0 \in \mathbb{R}^p$.

While Termination test is not satisfied **do**

Choose index $i_k \in \{1, 2, \ldots, p\}$.

$\beta^{k+1} \leftarrow \beta^k - \alpha_k [\nabla h(\beta^k)]_{i_k} e_{i_k}$ for some $\alpha_k > 0$.

$k \leftarrow k + 1$.

End While

The resolution of a *Slope* based optimization problem follows a similar pattern.

1.4.2 Simulation studies

We provide in this thesis two main approaches to estimate the variable importance measures ψ_f and ψ_{pen} that we introduce. Naturally, one would like to compare them. This is easy when f, the user-supplied function, satisfies $f = \bar{\Phi}_k$ in which case $\bar{\Phi}_k(W) \cdot \psi_f(P_0) = \psi_{pen}(P_0)$.

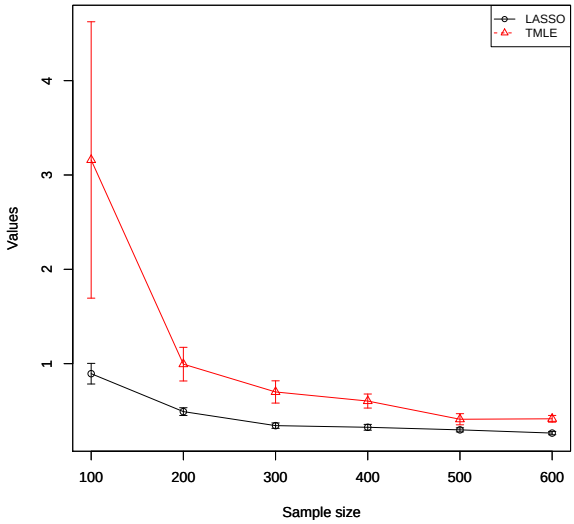
Let us denote by $\hat{\psi}_f$ and $\hat{\psi}_{pen}$ the estimators of ψ_f and ψ_{pen} in this special estimation. One expects by design that the prediction risk linked to $\hat{\psi}_{pen}$ will be smaller than the one linked to $\hat{\psi}_f$. On the contrary, the confidence regions built from $\hat{\psi}_{pen}$ will surely be less precise than the ones built from $\hat{\psi}_f$.

Comparison of Prediction and Estimation Risks - No measurement errors Given the above, we now introduce a framework, relying on the model (1.4), which will be the base case for our analysis. We assume that the function g is given by $g(W) = W$ and the function f, is given by $f(W) = 20 \cdot cos(2\pi W) + 5 \cdot sin(2\pi 2W)$. We also assume that the variables W and X are drawn from a $[0,1]$-uniform distribution. We then force randomly 15% of values of X to be equal to 0, so as to be in line with the constraints introduced in [23]. The variable Y is generated through (1.4). We use a *Fourier basis* for our decomposition. We generate data samples $O = (X, W, Y)$ of size $n = \{100, 200, 300, 400, 500, 600\}$.

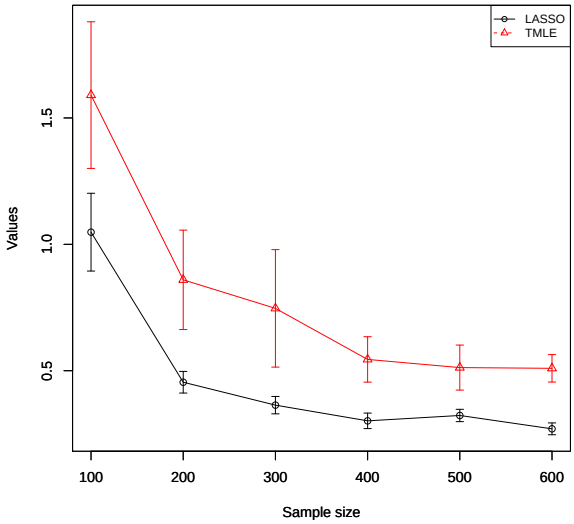
The *Prediction Risk* calculation is done through a two-step cross-validation process. We partition the sample O in l distinct subsets O^{subset}, of equal size. In the firs step, we set aside one of the $O^{subset}_{\{i\}}$, which is used as the test set, and use the $l-1$ remaining subsets as training sets. For each test set, we compute the estimator $\hat{\theta}_{\{-i\}}$. In the second step, we use the $O^{subset}_{\{i\}}$ to determine how close the estimator $\hat{\theta}_{\{-i\}}$ is to the real value. These two steps are repeated successively l times. For each iteration calculated the difference between the estimator and the true value function using the test set. The empirical average of these l distances give us the desired *Prediction and Estimation Risks.*

We see from figures (1.1b) and (1.1a) that the *TMLE* estimators perform relatively poorly compared to *Lasso* based ones for very small sample size. However, as data increases, the performance, of both methods, tends to converge.

Comparison of Prediction and Estimation Risks - With measurement errors We rely here on the model (1.12) in order to build the framework of our analysis. We assume that W is drawn from a $[0,1]$-uniform distribution. The elements of Z are given by $Z = X + \nu$ where X is drawn from a $[0,1]$-uniform distribution, and ν is a vector coming from a centered Gaussian distribution with standard deviation $\mu = 20\%$. As before, we force 15% of the elements of Z to be equal to 0. he variable Y is generated through (1.4). We use a *Fourier basis* for our decomposition. We generate data samples $O = (Z, W, Y)$ of



(a) Evolution of the Prediction Risk of the function f using both TMLE and Lasso methodology

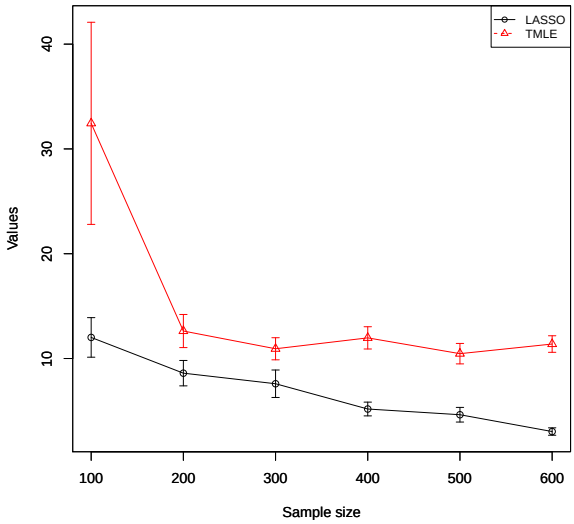


(b) Evolution of the Estimation Risk of the function f using both TMLE and Lasso methodology

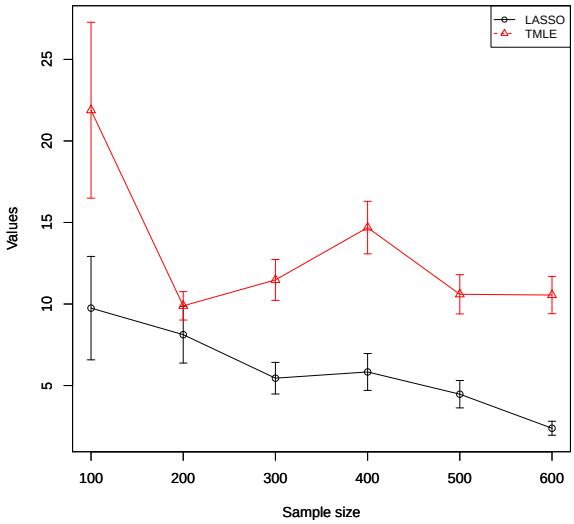
Figure 1.1 – Model without measurement errors

size $n = \{100, 200, 300, 400, 500, 600\}$. We rely on the procedure described in the previous paragraph to infer our *Prediction and Estimation Risks* for this case as well.

We see from figures (1.2b) and (1.2a) that the *TMLE* performs poorly for small sample size and further struggles to improve even when we increase the sample size. As expected, it is biased for cases when data sample is polluted.



(a) Evolution of the Prediction Risk of the function f using both TMLE and Lasso methodologies with data subject to error - μ = 20%



(b) Evolution of the Estimation Risk of the function f using both TMLE and Lasso methodologies with data subject to error - μ = 20%

Figure 1.2 – Model with measurement errors

Chapter 2

Inference of a non-parametric covariate-adjusted variable importance measure of a continuous exposure

This chapter is based on the manuscrit [23].

Abstract

We consider a setting where a real-valued variable of cause X affects a real-valued variable of effect Y in the presence of a context variable W. We aim at quantifying the impact of (X, W) on Y, while making only realistic assumptions on the true data generating distribution of (W, X, Y). To do so, we introduce a non-parametric, context-adjusted variable importance measure, whose definition relies on a user-supplied marginal structural model. It generalizes the variable importance measure introduced by [22]. We show how to infer it by targeted minimum loss estimation (TMLE), conduct a simulation study and present an illustration of its use.

2.1 Introduction

The statistical literature offers different ways to assess the importance of a variable. One way essentially consists in comparing the prediction performances of two models. The bigger model uses all the prediction variables whereas the smaller one, a subset of the bigger model, also relies on all prediction variables except the variable of interest. More algorithmic in nature, another way consists in assessing the importance of a variable by counting "the number of times" the variable of interest is used in the process of making predictions. This is the case for instance in random forests [15]. To the best of our knowledge, neither approach lends itself well to the construction of confidence regions. Following the seminal article [75], a third way consists in defining the variable importance measure as a sound statistical parameter and developing tailored estimators which, by the central limit theorem, can be used to construct confidence regions. An example of such a variable importance measure is introduced and studied by [22], upon which this manuscript builds.

The setting. Consider the situation where a real-valued variable of cause, $X \in \mathbb{R}$ (a continuous exposure), affects a $[0,1]$-valued variable of effect, Y (a continuous response), in the presence of a variable of context $W \in \mathcal{W}$ (covariates). The objective is to assess to what extent (X,W) influences Y while making only realistic assumptions on the unknown distribution P_0 of $O = (W,X,Y)$. This requires both the definition of a tailored statistical parameter and the elaboration of a semi-parametric inferential procedure to construct confidence regions of a given asymptotic level based on independent copies of O drawn from P_0.

Marginal structural models are very useful tools in this regard. The notion of marginal structural model was introduced in [60]. Widely used in the epidemiology literature, they are parametric classes of regression functions. Let $\{\mathrm{msm}_\beta : \beta \in B\}$ be such a class of functions mapping $\mathbb{R} \times \mathcal{W}$ to $\mathbb{R}$. It yields a parameter, defined as a minimizer in β of

$$\beta \mapsto E_{P_0}\left([Y - E_{P_0}(Y|X = x_0, W) - \mathrm{msm}_\beta(X,W)]^2\right),$$

where x_0 is a reference value for X for which there exists $0 < c < 1/2$ such that $P_0(X \neq x_0|W) \in [c, 1-c]$ P_0-almost surely. For instance, choosing a marginal structural model with $B = \mathbb{R}$ and msm_β given by $\mathrm{msm}_\beta(X,W) = \beta(X - x_0)$ yields the non-parametric variable importance measure studied in [21, 22].

For technical reasons, we focus on marginal structural models of the form

$$\{(X,W) \mapsto (X - x_0) f_\beta(W) : \beta \in B\}, \tag{2.1}$$

where f_β is linear in β. The statistical parameter of interest is formally defined as

$$\psi_0 = \arg\min_{\beta \in B} E_{P_0}\left([Y - E_{P_0}(Y|X = x_0, W) - (X - x_0) f_\beta(W)]^2\right), \tag{2.2}$$

assuming that the minimizer exists and is unique. We interpret $(X - x_0)f_{\psi_0}(W)$ as the best approximation of the form $(X - x_0)f_\beta(W)$ to $(E_{P_0}(Y|X,W) - E_{P_0}(Y|X = x_0, W))$. It quantifies the influence of X on Y, using x_0 as a reference value, while accounting for the covariates W on a linear scale.

Relevant literature. Our main source of inspiration is [21, 22]. These articles were motivated by an application to the analysis of the effect of DNA copy number variations (the exposure X) on gene expression (a response Y) accounting for DNA methylation (the covariates W). Their parameter of interest corresponds to the marginal structural model

$$\{(X,W) \mapsto \beta(X - x_0) : \beta \in \mathbb{R}\}. \tag{2.3}$$

Contrary to [2, 49, 56, 70, 80], [21, 22] do not assume that the real-valued variable of cause X is discrete (or do not discretize it), but rather exploit the fact that X has a reference value x_0 ($x_0 = 2$ for DNA copy variations) and features a continuum of other values. Moreover, contrary to [58, 59], [21, 22] do not assume a semi-parametric model but rather exploit one. We do too exploit the marginal structural model (2.1) and do not assume that $E_{P_0}(Y|X,W) - E_{P_0}(Y|X = x_0, W)$ belongs to it.

Our parameter of interest ψ_0 (2.2) points to an element of (2.1) such that $E_{P_0}(Y|X,W) - E_{P_0}(Y|X = x_0, W)$ is best approximated by $(X - x_0)f_{\psi_0}(W)$. In words, W is not averaged out completely like in [21, 22]. Instead, the effect of X on Y is quantified as $(X - x_0)$ times a function of the linear expression $f_{\psi_0}(W)$ of W. Hence, contrary to [21, 22], we are able to capture the role played by each component of W in the relationship between X and Y.

We will view ψ_0 as the value of a functional Ψ at P_0, where Ψ maps the statistical model describing the experiment to B. Since Ψ is path-wise differentiable, we can carry out the inference of ψ_0 by targeted minimum loss estimation (TMLE) [21, 22, 76, 78]. The TMLE template is a stepwise procedure. Its instantiation when the negative log-likelihood is chosen as loss function can be described as follows. Set $k = 0$. The first step consists in defining the parameter of interest as the value of a smooth functional Ψ at P_0. The second step consists in using machine learning to estimate some relevant features of P_0 which are needed to derive an initial substitution estimator $\psi_n^k = \Psi(P_n^k)$ of ψ_0 and the derivative $D^*(P_n^k)$ of the functional Ψ at P_n^k. This initial estimator is typically biased, or does not satisfy a central limit theorem. The third step consists in defining a parametric model through P_n^k whose score at P_n^k spans $D^*(P_n^k)$. The fourth step is a maximization of the log-likelihood of this parametric model, which yields an update P_n^{k+1} of P_n^k, hence and updates $\psi_n^{k+1} = \Psi(P_n^{k+1})$, $D^*(P_n^{k+1})$ of $\psi_n^k = \Psi(P_n^k)$, $D^*(P_n^k)$. The third and fourth steps are iterated (in k) till convergence (specifically, in our case, till $\int D^*(P_n^k)dP_n = o_P(1/\sqrt{n})$, where P_n is the empirical measure). The TMLE estimator is the last of the updated substitution estimators, denoted $\Psi(P_n^\star)$.

Organization. Section 2.2 defines and studies the functional Ψ associated to ψ_0 (2.2).

Section 2.3 describes the inference procedure tailored to the construction of confidence regions of a given asymptotic level for ψ_0. Section 4.4 presents the results of a simulation study. Section 2.5 gives an illustration based on real data on climate change. Relevant materials and proofs are gathered in the appendix.

2.2 Studying the parameter of interest

Without loss of generality, we assume from now on that $x_0 = 0$.

Differentiability and robustness. We denote $\dot{f} = \frac{\partial}{\partial\beta} f_\beta$ the gradient of f_β which, by choice, does not depend on β. Denote d the dimension of the space where $\dot{f}(W)$ lives. For every possible data-generating distribution P of O, we denote $\theta(P)(X,W) = E_P(Y|X,W)$ and $g(P)(W) = P(X \neq 0|W)$. Assume that P is chosen such that

1. $\mu(P)(W) = E_P(X\dot{f}(W)|W)$ and $\Sigma(P) = E_P[X^2\dot{f}(W)^\top\dot{f}(W)]$ are well-defined features of P;
2. $\Sigma(P)$ is invertible;
3. there exists $c \in]0, 1/2[$ such that $g(P)(W) \in [c, 1-c]$ P-almost surely.

Condition 3 is called a "positivity assumption". It guarantees that, in every stratum of W, the conditional probabilities of $X = 0$ and $X \neq 0$ are larger than $c > 0$. Conditions 1 and 2 on the joint distribution of $(X, \dot{f}(W))$ are technical but standard. They are met if $X\dot{f}(W)$ is a bounded random variable and if there is no deterministic linear combination of the components of $X\dot{f}(W)$ which equals 0 P-almost surely. For such a P, the equation

$$\Psi(P) = \underset{\beta\in B}{\arg\min}\, E_P\left(\left[\theta(P)(X,W) - \theta(P)(0,W) - Xf_\beta(W)\right]^2\right) \tag{2.4}$$

uniquely characterizes a parameter of P such that $\Psi(P_0) = \psi_0$, if P_0 meets the constraints, which we assume from now on to be true. It occurs that $\Psi(P)$ rewrites as

$$\Psi(P) = \Sigma(P)^{-1}[E_P(X\dot{f}(W)(\theta(P)(X,W) - \theta(P)(0,W)))]. \tag{2.5}$$

The functional Ψ is path-wise differentiable at P, with an efficient influence curve given by $D^\star(P) = D_1^\star(P) + D_2^\star(P)$ where $D_1^\star(P)$ and $D_2^\star(P)$ are two $L_0^2(P)$-orthogonal components characterized by

$$\begin{aligned} D_1^\star(P)(O) &= \Sigma(P)^{-1}\left[(\theta(P)(X,W) - \theta(P)(0,W) - Xf_\beta(W))\right]X\dot{f}(W), \text{ and} \\ D_2^\star(P)(O) &= \Sigma(P)^{-1}\left[(Y - \theta(P)(X,W))\left(X\dot{f}(W) - \frac{1_{\{X=0\}}}{g(P)(0|W)}\mu(P)(W)\right)\right]. \end{aligned}$$

This means that, for any bounded $s \in L_0^2(P)$ taking values in $\mathbb{R}^d$ and $\varepsilon \in \mathbb{R}^d$ with $\|\varepsilon\|_\infty < \|s\|_\infty^{-1}$, if we characterize a data-generating distribution P_ε of O by setting

$$\frac{dP_\varepsilon}{dP}(O) = 1 + \varepsilon^\top s(O),$$

then for ε small enough, P_ε meets Conditions 1, 2, 3, hence $\Psi(P_\varepsilon)$ is well-defined, and moreover $\varepsilon \mapsto \Psi(P_\varepsilon)$ is differentiable at $\varepsilon = 0$ with a derivative given by

$$\lim_{\varepsilon \to 0} \frac{\Psi(P_\varepsilon) - \Psi(P)}{\varepsilon} = E_P[s(O)^\top D^\star(P)(O)].$$

The efficient influence curve $D^\star$ enjoys a remarkable property: if P, P' are two data-generating distributions of O satisfying Conditions 1, 2, 3 and such that $E_P(D^\star(P')(O)) = 0$, then $\Psi(P) = \Psi(P')$ whenever $\theta(P')(0,\cdot) = \theta(P)(0,\cdot)$ or $(\mu(P) = \mu(P')$ and $g(P') = g(P))$. The validity of all the statements made in this section can be checked by adapting, *mutatis mutandis*, the proofs in [22]. Although not straightforward, the exercise of adapting the prior proofs was direct enough for us to decide to omit them in this document.

It is common sometimes to give a causal interpretation to a variable importance measure. It is possible to give such an interpretation in our case.

Causal interpretation. The causal interpretation partly relies on untestable assumptions. Assume, in this section only, that there exists a collection $(Y_x)_{x\in\mathbb{R}}$ of random variables such that *(i)* $(Y_x)_{x\in\mathbb{R}} \perp X|W$ (randomization assumption), and *(ii)* $Y = Y_X$ (consistency assumption). The above holds for instance in the following structural equation model: *there exists three deterministic functions f_W, f_X, f_Y and three independent random variables U_W, U_X, U_Y such that $W = f_W(U_W), X = f_X(W, U_X)$ and $Y = f_Y(W, X, U_Y)$.* In addition, assume that the conditional laws of X given W are all dominated by a common measure μ. Then, there exists a collection of conditional densities $\phi(\cdot|W)$ of X given W, all with respect to μ.

Let us denote by $\mathbb{P}$ the law of the full data $(W, X, (Y_x)_{x\in\mathbb{R}})$. It holds that $E_P(Y|X = x, W) = E_{\mathbb{P}}(Y_x|X = x, W) = E_{\mathbb{P}}(Y_x|W)$, by independence of Y_x and X. Furthermore, for each $\beta \in B$,

$$E_P\{(E_P(Y|X,W) - E_P(Y|X=0,W) - Xf_\beta(W))^2\} = \int E_P\left[(E_{\mathbb{P}}(Y_x - Y_0 - xf_\beta(W)|W))^2\phi(x|W)\right]\mu(dx). \quad (2.6)$$

proof of (2.6). The following series of equalities proves (2.6), where the third one is a consequence of Fubini's theorem:

$$\begin{aligned}
&E_P\{(E_P(Y|X,W) - E_P(Y|X=0,W) - Xf_\beta(W))^2\} \\
&\quad = E_P\left\{E_P\left((E_P(Y|X,W) - E_P(Y|X=0,W) - Xf_\beta(W))^2|W\right)\right\} \\
&\quad = E_P\left\{\int (E_{\mathbb{P}}(Y_x|W) - E_{\mathbb{P}}(Y_0|W) - xf_\beta(W))^2\phi(x|W)\mu(dx)\right\} \\
&\quad = \int E_P\left[(E_{\mathbb{P}}(Y_x|W) - E_{\mathbb{P}}(Y_0|W) - xf_\beta(W))^2\phi(x|W)\right]\mu(dx) \\
&\quad = \int E_P\left[(E_{\mathbb{P}}(Y_x - Y_0 - xf_\beta(W)|W))^2\phi(x|W)\right]\mu(dx).
\end{aligned}$$

□

We recall here that, as aforementioned, we interpret $(X - x_0)f_{\psi_0}(W)$ as the best approximation of the form $(X - x_0)f_\beta(W)$ to $(E_{P_0}(Y|X,W) - E_{P_0}(Y|X = x_0, W))$. Thus, given the above, $\Psi(P)$ can be interpreted as the coefficient associated with the regression of Y_x on $Y_0 + f_\beta(x, W)$ based on a weighted L^2-loss function.

2.3 Inference

Inference is based of n independent random variables $O^{(i)}$ $(i = 1, \dots, n)$ drawn from P_0. We infer $\psi_0 = \Psi(P_0)$ by TMLE.

Initialization. The initialization consists in estimating the following features of P_0: marginal distribution of W, $\mu(P_0)$, $g(P_0)$, $\theta(P_0)$, $\Sigma(P_0)$ and, for each of them, a companion feature required to update them at the next step [see 22, Lemma 1]. We denote P_n^0 a data-generating distribution chosen such that *(i)* each estimator η_n of a feature $\eta(P_0)$ among the above features of interest can be rewritten $\eta_n = \eta(P_n^0)$, and *(ii)* we can sample (W, X) from P_n^0. As soon as we have built estimators of the marginal distribution of W, $\mu(P_0)$, $g(P_0)$, $\theta(P_0)$ and $\Sigma(P_0)$, we can also estimate ψ_0 and $D^\star(P_0)$. This initial estimator of ψ_0 can be biased. The evaluation of ψ_0 is performed by Monte-Carlo simulation: we simulate B independent random variables $(W^{(0,b)}, X^{(0.b)})$ from the marginal joint distribution of (W, X) under P_n^0, then compute

$$\psi_n^0 = B^{-1} \sum_{b=1}^{B} \Sigma(P_n^0)^{-1} \Big[X^{(0,b)} \dot{f}(W^{(0,b)}) \Big(\theta(P_n^0)(X^{(0,b)}, W^{(0,b)}) - \theta(P_n^0)(0, W^{(0,b)}) \Big) \Big].$$

The construction of the marginal distribution of X given $(W, X) \neq 0$ under P_n^0, has been subject to some modifications relatively to [22]. In fact, the conditional distribution can be any distribution whose conditional mean is deduced from $\mu(P_n^0)$ by

$$E_{P_n^0}(X \dot{f}(W) | X \neq 0, W) = \frac{\mu(P_n^0)(W)}{1 - g(P_n^0)(0|W)}, \tag{2.7}$$

and such that the variable Σ verifies :

$$E_{P_n^0}\big[(1 - g(P_n^0)(0|W)) E_{P_n^0}(X^2 \dot{f}(W)^\top \dot{f}(W))\big] = \Sigma(P_n^0). \tag{2.8}$$

Iterative updating. Say we have built $(k-1)$ updates $P_n^1, \dots, P_n^{k-1}$ of P_n^0. The k^{th} update goes as follows. Set $0 < \rho < 1$ a constant close to 1, for instance $\rho = 0.99$ and, for each $\varepsilon \in \mathbb{R}^d$, $\|\varepsilon\|_\infty \leq \rho \|D^\star(P_n^{k-1})\|_\infty$, introduce $P_n^{k-1}(\varepsilon)$ given by

$$\frac{dP_n^k(\varepsilon)}{dP_n^{k-1}}(O) = 1 + \varepsilon^\top D^\star(P_n^{k-1})(O)$$

where $D^\star(P_n^{k-1})(O)$ is the current estimator of the efficient influence curve. This defines a d-dimensional parametric model through P_n^{k-1} fluctuating it in the direction of $D^\star(P_n^{k-1})$. We let ε_n^{k-1} be the maximum likelihood estimator of ε in this model and characterize the k^{th} update as $P_n^k = P_n^{k-1}(\varepsilon_n^{k-1})$. This yields updated estimators of the features of interest in the spirit of [22, Lemma 1], since it holds that

$$\begin{aligned}
\theta(P_\epsilon) &= \frac{\theta(P)(X,W) + \epsilon E_P(Y s(O)|X,W)}{1 + \epsilon E_P(s(O)|X,W)}, \\
\mu(P_\epsilon)(W) &= \frac{\mu(P)(W) + \epsilon E_P(X s(O)|W)}{1 + \epsilon E_P(s(O)|X,W)}, \\
g(P_\epsilon)(0|W) &= \frac{g(P)(0|W) + \epsilon E_P(1\{X=0\} s(O)|W)}{1 + \epsilon E_P(s(O)|X,W)}, \\
\Sigma(P_\epsilon) &= \Sigma(P) + \epsilon E_P[D^\star(O) X^2 \dot{f}(W) \dot{f}^T(W)].
\end{aligned}$$

The k^{th} update of ψ_n^0 is obtained by simulating B independent random variables $(W^{(k,b)}, X^{(k,b)})$ from the marginal joint distribution of (W,X) under P_n^k then computing

$$\psi_n^k = B^{-1} \sum_{b=1}^{B} \Sigma(P_n^k)^{-1} \left[X^{(k,b)} \dot{f}(W^{(k,b)}) \left(\theta(P_n^k)(X^{(k,b)}, W^{(k,b)}) - \theta(P_n^k)(0, W^{(k,b)}) \right) \right]. \quad (2.9)$$

Theorem 2.1 (**Central limit theorem**). *Suppose that performing k_n iterations of the updating procedure guarantees that $P_n D^\star(P_n^{k_n}) = o_P(1/\sqrt{n})$. Suppose moreover that there exists a function f_1 with $P_0 f_1 = 0$ such that $P_0(D^\star(P_n^{k_n}) - f_1)(D^\star(P_n^{k_n}) - f_1)^\top = o_P(1)$, and that $\Psi(P_n^{k_n}) - \psi_0 - P_0 D^\star(P_n^{k_n}) = o_P(1/\sqrt{n})$. In addition, suppose that S_n estimates consistently $E_{P_0}[f_1(O) f_1(O)^\top]$. Then $\psi_n^\star = \Psi(P_n^{k_n})$ satisfies $\sqrt{n}(\psi_n^\star - \psi_0) = (P_n - P_0) f_1 + o_P(1)$, hence $\sqrt{n} S_n^{-1/2}(\psi_n^\star - \psi_0)$ converges in law to the d-multivariate Gaussian law with zero mean and identity covariance matrix. We refer the reader to [22, appendix] for the proof of a similar result.*

2.4 Simulation study

Simulation scheme. We essentially reproduce the simulation framework that has been substantially developed in [22]. Let $O_1 = (O_1^W, O_1^X, O_1^Y)$, $O_2 = (O_2^W, O_2^X, O_2^Y)$ and $O_3 = (O_3^W, O_3^X, O_3^Y)$ be the same real data structures as in [22, Section 6.4]. Let $p = (p_1, p_2, p_3)$ be such that $p_1, p_2, p_3 \geq 0$ and $p_1 + p_2 + p_3 = 1$ and $w = (w_1, w_2, w_3)$ be a vector of positive numbers. Let $\lambda_0 : [0,1] \to [0,1]$ be a non-increasing mapping, σ_2 be a positive number and Σ_1, Σ_3 be two 2×2 covariance matrices.

The sampling of a generic data structure $O = (W, X, Y)$ from the synthetic data-generating distribution P^s unfolds as follows. We first draw a latent class assignment U from the multinomial distribution with parameter $(1, p)$. Conditionally on U, the first component W_1 of W is

$$W_1 = \text{expit}(\text{logit}(O_U^W) + w_U Z))$$

where Z is a standard normal random variable independent of U. The $(d-1)$ remaining components of W are drawn from the Gaussian distribution with mean zero and identity covariance matrix. Finally, (X, Y) is drawn conditionally on (U, W):

- if $U = 2$, then $(X, Y) = (0, O_2^Y + \lambda_0(W_1) + \sigma_2 Z')$, where Z' is a standard normal random variable independent of (U, W, Z).
- if $U \neq 2$, then (X, Y) is drawn conditionally on (U, W) from the bivariate Gaussian distribution with mean $(O_U^X - O_2^X, O_U^Y)$ and covariance matrix Σ_U.

Implementation. We have substantially adapted the package [20]. The main changes concern:

- the characterization of f_β in (2.1), which takes the form of a `R formula`;
- the adaptation of the fitting procedures of the features g, mu and $sigma$, and the storage of the fitted objects;
- the computation of the confidence regions.

Results of the simulation study. We considered two choices of marginal structural models (2.1): one is based on $f_\beta(W) = \beta_1 W_1 + \beta_2 W_2$, the other based on $f_\beta(W) = \beta_1 W_1 + \beta_2 W_1^2$ ($\beta \in \mathbb{R}^2$). The evaluation of the true value of ψ_0 for each choice of marginal structural model was performed by Monte-Carlo based on (2.5). We report the values in the second row of Table 2.1. For each choice, independently, we repeated independently $B = 1000$ times the simulation of a data set of sample size $n = 1000$ and the simulation of another data set of sample size $n = 2000$. We applied the TMLE procedure described in Section 2.3 to each data set, with the same choice of the fine-tune parameters as in [22] and with the option `flavor="learning"`.

The results are summarized in Table 2.1. The empirical coverage is satisfying.

MSM	$f_\beta(W_1, W_2) = \beta_1 W_1 + \beta_2 W_2$		$f_\beta(W_1, W_2) = \beta_1 W_1 + \beta_2 W_1^2$	
n	$\psi_{0,1} = 0.56$	$\psi_{0,2} = 0$	$\psi_{0,1} = 1.53$	$\psi_{0,2} = -1.42$
1000	94.6%	95.0%	95.9%	94.5%
2000	93.9%	93.9%	95.6%	93.7%

Table 2.1 – Summary of the results of the simulation study. The values of the true parameter are reported in the second row. The third and fourth row give the empirical coverage of the regions of confidence for each coordinate and each sample size n. MSM stands for "marginal structural model".

2.5 Illustration

It is commonly agreed today that human activities have significant impact on climate. Among others institutions, IPCC (Intergovernmental Panel on Climate Change) has been conducting exhaustive studies on the topic for decades. The effect of CO_2 emissions on climate change is now much better understood [57, 66]. However, one of the major remaining challenges is to understand which factors drive climate change. Our parameter (2.4) can prove useful in this regard.

We exploit a publicly available data set of the World Bank[1]. We extract from it our data set. It consists of $n = 126$ observed data-structures $O_1, \dots, O_i = (W_i, X_i, Y_i), \dots, O_n$ where, for the ith country,

- W_i gathers its under-five mortality rate, population growth, urban population growth, CO_2 emissions per unit of Gross Domestic Product (GDP), energy use per unit of GDP, energy use per capita for the year 1998;
- X_i is a thresholded version of total amount of CO_2 emissions per capita for the year 1998; we rely on a thresholded version to enforce the existence a reference value for the exposure;
- Y_i is the 10%-quantile of the projected annual temperature change for the period 2045–2065.

Under-five mortality rate is a reliable indicator of poverty. Population growth and urban population growth are relevant indicators of economical development. CO_2 emissions per unit of GDP is an indicator of industrialization and reliance on fossil fuel. Finally, energy use per unit of GDP and per capita reveal patterns of energy consumption by the industry and by the country's inhabitants.

All the X_i are non-negative. Their empirical distribution is represented in the LHS plot of Figure 2.1. In order to introduce a reference level to the exposure X, we set to

[1] http://data.worldbank.org/data-catalog/climate-change

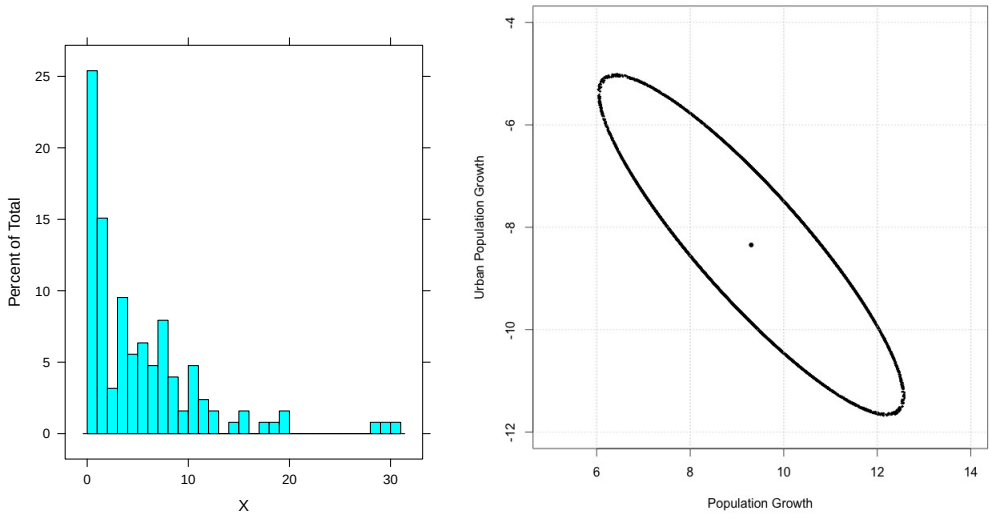


Figure 2.1 – Left: Histogram of the variable X. Right: Confidence region of asymptotic level 95% for parameter $(\psi_{0,2}, \psi_{0,3})$.

$x_0 = 0$ exactly all the X_i smaller than 0.99, which is the 25%-quantile of the empirical distribution of X.

We assume that $O_1, \ldots, O_n$ are independently drawn from a common distribution P_0. We infer $\psi_0 = \Psi(P_0)$ given by (2.5) for the marginal structural model $\{(X, W) \mapsto X f_\beta(W) : \beta \in \mathbb{R}^6\}$ with $f_\beta(W) = \beta^\top W$.

Using the asymptotic normality of the TMLE $\psi_n^\star$, we carry out Student tests of "$\psi_{0,k} = 0$" against "$\psi_{0,k} \neq 0$" for $k = 1, \ldots, 6$. We reject the null for its alternative at level 5% only for $k = 2, 3$, *i.e.*, for population growth and urban population growth, with p-values respectively equal to 3.69×10^{-10} and 2.01×10^{-8}. The corresponding estimates are $\psi_{n,2}^\star = 9.30 \pm 1.33$ and $\psi_{n,3}^\star = -8.34 \pm 1.35$, see also the RHS plot in Figure 2.1. In other words, we estimate $f_{\psi_0}(W)$ with $f_{\psi_n^\star}(W) \approx 9.30 \times W_2 - 8.34 \times W_3$.

The results above teach us that only *population growth* and *urban population growth* seem to be playing key roles in the relationship between *climate change* and CO_2 *emissions per capita*.

Remark. We have carried out the same study with (W, X) corresponding to the years 1990 to 1997. The results of inference and subsequent conclusions were very similar to those presented here (results not shown).

Conclusion. We have generalized the variable importance measure introduced in [22], through the introduction of a more general and flexible marginal structural model, where each covariate W is individually taken into account. Similarly to [22], we proved that its TMLE estimator is consistent and asymptotically convergent, under mild conditions (as mentioned before, although not straightforward, the exercise of adapting the proofs from [22] was direct enough for us to decide to omit them in this document.). Furthermore, we have extended the package [20] so that it can handle the new parameter of interest.

Chapter 3

Linear regression model with functional coefficients and errors-in-variables

This chapter is based on the manuscript [24].

Abstract

We study an univariate linear regression model $Y = g(W) + X \cdot f(W) + \epsilon$, whose coefficients g and f are real-valued functions of (possibly multivariate) covariates W. We assume that g and f admit a finite dimensional expansion on some orthonormal basis in the Hilbert $\mathbb{L}_2(\mathcal{W})$, where $\mathcal{W} \subset \mathbb{R}^d$ is the space of outcomes of the covariates. These dimensions can be large for both g and f, or just for f. We estimate g and f by a *Lasso* like procedure. We also study the same estimation problem in the presence of errors-in-variables, that is, instead of X we observe $Z = X + \nu$ where ν is assumed centered, with known variance and independent of X. We evaluate the behaviour of the procedure on synthetic data and compare our estimator of f with the TMLE in a semi-parametric model for f. Finally, our methodology is applied to a financial dataset in order to search for a (relatively) small portfolio to replicate the predict (daily return) of a given financial index.

3.1 Introduction

Statisticians are often confronted with the problem of selecting a subset within a large sample of variables in order to build a predictor of an outcome of interest. The selection is important because it helps choose relevant explanatory variables, necessary to define a consistent and unbiased predictor.

When dealing with high dimensional data, the selection problem becomes more acute. Consider the case where we use all variables in the dataset to build a regression model. The resulting predictor will likely be poor. In fact, it will yield too many main regressors. One can attempt to change the functional form of the model in order to improve the performance of the predictor, but the selection of a smaller set will always be preferable. As such, many solutions have been developed in the literature to help quantify the importance of a variable in a given regression model or while attempting to define one. They can essentially be divided in two groups. The first group, generally, builds successive predictors with and without a variable (or a set of variables) of interest, respectively, which we wish to measure the importance. The difference in accuracy between both predictors will provide the measure of importance. A general, but detailed, overview of these methods is provided in [41]. The second group consists of *ensemble learning* methods which, for some of them, relies on *decision trees learning* in order to define the best predictor. An early example of such algorithm is called *bagging* and was introduced in [13]. A detailed comparison, of some of these methods (i.e.: *bagging, boosting* and *randomization*), is proposed in [31]. A similar method called *Random Forest*, which is based on a random bootstrapping of the main learning set was also introduced in [14]. For this group of methods, two main measures are commonly used to quantify the variable importance. The first one is known as *Mean Decrease Impurity importance (MDI)* or *Mean Decrease Gini*. It is based on the *Gini impurity* which is computed at each node of the decision tree to help determine how to split the data within the node into smaller datasets. Once aggregated over the entire decision tree, it produces the *MDI* value for each covariate. The second measure is called *Mean Decrease Accuracy (MDA)*, also known as *Mean Permutation Importance*. Similarly to the first measure, at each node of the decision tree, the accuracy of the predictor is evaluated by randomly permutating a variable in the *out-of-bag sample*. An aggregation over the entire decision tree provides the desired value. We note however that these solutions have a few drawbacks. Tree decision methods sometimes produce optimal predictors which have very few variables, given the initial dataset (see [10]). This is far from being a desirable feature. *Random Forest* is known for over-fitting learning datasets that are particularly noisy. Moreover, both of these groups do not allow the calculation of a confidence interval related to the variable importance measure.

A new type of measure was introduced in [77] to solve some of these issues. It relies on

a model which establishes a relation between an outcome Y and a set of covariates W. It is defined as a *real-valued* parameter $\varphi_{vdL}(P)$, predictor of $E_P(Y|A,W)$ and given by

$$\varphi_{vdL}(P)(a) = E_{P^\star}(E_P(Y|A=a,W) - E_P(Y|A=0,W)), \tag{3.1}$$

where (W,Y) are elements of an unknown distribution P, $P^\star$ is a known function of P and A is considered to be either a subset or a function of W. This parameter is interesting because it provides a predictor which measures directly the variable importance of a covariate, and to which we can associate a p-value. A similar variable importance measure was introduced in [22], and extended to a multivariate setup in the manuscript [23].

Let us consider the following statistical problem. We observe the data structure $O = (W,X,Y)$ of an unknown distribution P_0, containing n i.i.d elements, where $W \in \mathbb{R}^d$ represents a vector of covariates, with $d \geq 1$, $X \in \mathbb{R}$ represents an exposure and $Y \in \mathbb{R}$ is a response. The exposure has a reference level x_0 which is considered to be given by $x_0 = 0$, without loss of generality. We want to study the relationship that exists between the response Y and the exposure X, when taking into account the covariates W. Their inclusion makes sense specifically when it is not possible to rule out completely their influence in the relationship. In [22], the parameter φ_{CNvdL} is given by

$$\varphi_{CNvdL}(P) = \arg\min_{b\in\mathbb{R}} E_P\left[(Y - E_P(Y|X=0,W) - X\cdot b)^2\right].$$

Hence, $\varphi_{CNvdL}(P)$ can be seen as the coefficient of the best linear approximation of $Y - E_P(Y|X=0,W)$. It is important to note here that φ_{CNvdL} is a real number. As such, it only provides an aggregated view of the variable importance measure of all covariates in the model through a single real value. One might want to have a granular view of that measure. To be specific, it might be more informative to have a variable importance measure whose elements are uniquely linked to each covariate of the problem. This gap was addressed, in [23], where we considered a function f given by $f(B,W) = B^\top \cdot f_{CT}(W)$, with $f_{CT}(W) : \mathbb{R}^d \to \mathbb{R}^d$ supposed known and $B \in \mathbb{R}^d$ such that the d-dimensional parameter of interest φ_{CT}, was given by

$$\varphi_{CT}(P) = \arg\min_{B} E_P\left[(Y - E_P(Y|X=0,W) - X\cdot f(B,W))^2\right]. \tag{3.2}$$

We note that $\varphi_{CT}(P)$ takes into account each covariate present in the model, when describing the relationship that exists between the exposure X and the response Y. It is important to note that the parameters φ_{CNvdL} and φ_{CT} were both studied through a statistical inference method called *TMLE (Targeted Minimum Likelihood Estimator)*. A brief overview of the analysis of the parameter φ_{CT} is presented in Section 3.4.2.

As aforementioned, we are interested in this article in the estimation of the function f as a mean to derive a more refined expression of the variable importance measure. By

assuming f_{CT} known in [23], the author was constraining his framework to a very specific form in order to define the relation that existed between the covariates. As an example, considering the case where the model had only two covariates, the user may choose to define f_{CT} through $f(W,\beta) = \beta^\top \cdot f_{CT}(W) = \beta_1 W_1^2 + \beta_2 W_2^2$, with $W = (W_1, W_2)$ and $\beta = (\beta_1, \beta_2)$. In fact, for the parameter φ_{CT} to be well defined, due to the definition space of f_{CT}, it had to be of length d. Now, by considering f unknown, we are deriving a flexible relationship between the covariates and as a consequence lifting the constraint on the size of φ_{CT}. It then allows us to obtain a complete picture of the measure of variable importance linked to the covariates of the model, since we are now able, as an example, to potentially take into account a large variety of interaction terms among the covariates.

We recall that $f : \mathbb{R}^d \to \mathbb{R}$ is now the non-parametric function of interest. Keeping in mind the framework developed in [23], the function of interest is an extended, unknown and nonparametric form of the parametric form $f(\varphi_{CT}, W)$, where f was supposed known. Hence, it can be viewed as a real-valued variable importance measure which takes into account the impact of the covariates W. Its estimator is computed through a penalized optimization problem. Furthermore, in order to derive the estimator's convergence rate, we assume the existence of a model which defines the relationship that exists between the exposure X, the response Y and the covariates W.

Consider the model

$$Y = g(X,W) + \epsilon \tag{3.3}$$

where ϵ is a centered Gaussian variable of variance σ^2, independent of X and W, and $g(X,W) = g(0,W) + f_1(X,W)$, with $g(0,W) = E_P(Y|X=0,W)$ and f_1 an unknown function. We further simplify the model (3.3) by analogy to (3.2). We assume that g is linear in X such that

$$g(X,W) = g(0,W) + X \cdot f(W). \tag{3.4}$$

We may see f as the partial derivative of g with respect to X at point 0. For the rest of this article, we denote $g(W) = g(0,W)$, through a slight abuse of notation, with $f, g : \mathbb{R}^d \to \mathbb{R}$. Let us consider $\Phi = (\Phi_1, \cdots, \Phi_p, \cdots) = \{\Phi_j\}_{j=1}^{\infty}$ be an orthonormal base of the Hilbert space $\mathbb{L}_2(\mathbb{R}^d)$. We define the subspace $\mathcal{S}(s,p) \subset \mathcal{H}$, with $\{s,p\} \in \mathbb{N}$. A function $h \in \mathcal{S}(s,p)$, if the function verifies the following properties:

i) can be written as $h(W) = \sum_{j=1}^{p} \Phi_j(W) b_j$ where $b \in \mathbb{R}^p$

ii) the support of b is of size smaller or equal to $s(|\{b_j; b_j \neq 0\}| \leq s)$.

From now on, we assume that there exists $\{s_m, m\} \in \mathbb{N}$ and $\{s_k, k\} \in \mathbb{N}$ such that $g \in \mathcal{S}(s_m, m)$ and $f \in \mathcal{S}(s_k, k)$. Hence,

$$g(X_i, W_i) = \sum_{j=1}^{m} \Phi_j(W_i) \cdot \eta_j + X_i \sum_{j=1}^{k} \Phi_j(W_i) \cdot \theta_j, \text{ with } \eta \in \mathbb{R}^m \text{ and } \theta \in \mathbb{R}^k. \tag{3.5}$$

It is important to note here that, by analogy to (3.2), $\Phi(W)\cdot\theta$ is the new variable importance measure. We note here that we could have assumed that f and g relied, for their definition, on the full list of elements of the basis Φ. However, following this hypothesis, we would have $g(W_i) = \sum_{j=1}^{m} \Phi_j(W_i)\cdot\eta_j + r_{g_m}(W_i)$ and $f(W_i) = \sum_{j=1}^{k} \Phi_j(W_i)\cdot\theta_j + r_{f_k}(W_i)$, where the functions r_{f_k} and r_{g_m} could be considered as residuals, that are small enough under smoothness assumptions. Therefore, it would be necessary to first include the impact of these residuals on consistency and convergence rate when estimating f and g. Secondly, we would also have to derive their appropriate characteristics in order to improve the previous estimation measures. Given (3.5), the linear model (3.3) becomes

$$Y = \mathbb{X}\cdot\beta + \epsilon \tag{3.6}$$

where $\beta = \begin{pmatrix} \eta \\ \theta \end{pmatrix} \in \mathbb{R}^p$ with $p = m + k$. Moreover, $\mathbb{X}$ is given by

$$\mathbb{X} = \begin{bmatrix} \Phi_1(W_1) & \dots & \Phi_m(W_1) & X_1\cdot\Phi_1(W_1) & \dots & X_1\cdot\Phi_k(W_1) \\ \vdots & & \vdots & \vdots & & \vdots \\ \Phi_1(W_n) & \dots & \Phi_m(W_n) & X_n\cdot\Phi_1(W_n) & \dots & X_n\cdot\Phi_k(W_n) \end{bmatrix}. \tag{3.7}$$

Furthermore, we denote by Ψ_m, the matrix given by the first m functions at points $W_1, \cdots, W_n$

$$\Psi_m = \begin{bmatrix} \Phi_1(W_1) & \dots & \Phi_m(W_1) \\ \vdots & & \vdots \\ \Phi_1(W_n) & \dots & \Phi_m(W_n) \end{bmatrix}$$

and by D_X the diagonal matrix with diagonal elements $X_1, \cdots, X_n$. Thus, $\mathbb{X} = [\Psi_m, D_X\Psi_k]$. A *Lasso like* methodology is developed and used to derive a consistent estimator $\hat{\beta}$ of β, an ultimately the estimator $\hat{\theta}$ of θ, our main variable of interest. From now on, we consider that there exists $\delta > 0$, a fixed constant, such that:

$$|\Phi_j(W)| \le \delta, \quad \text{for all } W \text{ in the data structure } O \text{ and } j \in \{1, \dots, p\}. \tag{3.8}$$

Measurement errors occur in most experiments and are integral part of any scientific process. As such, the aforementioned elements (X, W, Y) can rarely be collected without being subject to errors. We consider the case of additive error on the random variable X_i, i.e. we observe $Z_i = X_i + \nu_i$ with $i = 1, \cdots, n$. As such, further generalizing (3.3), we preserve the statistical framework introduced above, except from the sample's elements. Hence, the dataset is given by $\widetilde{O} = (W, Z, Y)$ of a distribution $\tilde{P}$, such that

$$\begin{cases} Y &= g(X, W) + \epsilon \\ Z &= X + \nu \end{cases} \tag{3.9}$$

where ν is a centered Gaussian distribution of variance μ^2. The variables ν and ϵ are considered independent. Our goal is to see how the measurement errors affect the estimation of g and f, as well as the prediction risk of our model. From now on, we assume, without loss of generality that, for all j, $\sum_{i=1}^{n} \Phi_j(W_i)^2 = n$.

Review of Literature

Looking at (3.4), one can realize that f is the partial derivative of $g(X, W)$ in X, at $X = 0$. Hence, estimating the function f, corresponds in this setting at finding the best approximate of the derivative of an unknown function. This problem has been studied in the literature (see [29, 38, 61, 84]). However, we did not pursue this path. Given the model formulation in (3.6), we opted for a *Lasso* based approach.

The *Least Absolute Shrinkage and Selection Operator(Lasso)* was introduced in [72]. It is well known to be attractive for large and sparse set of high dimensional data which is a key characteristic of interest for the models developed in this article. The Lasso has been extensively studied in the literature. The book [40] provides a good overview of its different properties in both low and high dimensional settings. Some of them are worth a reminder. In fact, within the space of linear models, [52] showed that under an appropriate set of conditions, *Lasso* estimator is consistent. We can also cite [83] who proved the consistency of the model selection performed by the *Lasso*. Finally, [43] generalized, to high dimensional settings, the oracle property of estimator based on *adaptative Lasso*, introduced in [85]. The latter is a $\mathbb{L}_1$ weighted optimization problem, relying on a pre-computed starting point. If we suppose an initial estimator $\beta_0 = w_{n_j}$ with $j = 1, \ldots, p_n$, then the least square optimization problem is given by $\|Y - \widetilde{X}\beta\|_2^2 - 2\lambda \sum_{i=1}^{n} w_{n_j}|\beta_j|$ where $\widetilde{X}$ is a matrix. The goal of the weight is to reduce the estimation bias and improve the variable selection accuracy. Under an adaptative irrepresentable condition, [85] proved that the estimator satisfies an oracle property. Some of the properties mentioned above, are covered in this document through the analysis of (3.6), which corresponds to a linear model case. We derived the error bounds of $\hat{\beta}$ and then deduced the ones of the variables $\hat{\eta}$ and $\hat{\theta}$.

Errors-in-variables models have been studied in the literature. Let us cite the semi-parametric model $Y = f_{\theta^0}(X) + \zeta$ and $Z = X + \epsilon$, developed in [18], where the efforts were directed at attaining parametric rates for estimating the finite dimensional θ^0. We further note that high-dimensional linear models have been studied in the presence of measurement errors. We can mention [62], who analyzed a linear regression model given by $Y = R_c \cdot \theta + \epsilon$ and $Z = R_c + U$, where the number of covariates is much larger than the sample size and the matrix R_c is polluted by the matrix U. They proved that the regular *Lasso* estimator, as well as the *Dantzig selector* produced unstable results. They then introduced a *Matrix Uncertainty* selector which, under certain eigenvalues restrictions

applied on R_c, is stable and capable of reproducing a sparse pattern of θ. Furthermore, [67] studied a linear model using the *Lasso* procedure knowing that the covariates are subject to measurement errors and for $p \gg n$. They proved that using a correction method and under certain conditions, the corrected *Lasso* produced sign consistent covariate selection. A similar correction method is used in this document, when dealing with noisy data. [27] completed the study of a similar setting, by introducing a convex corrected Lasso. They established error bounds of the corresponding estimator and showed its asymptotic sign consistent selection property. Finally, we can also mention [48], whose work is based on the same setting, using both multiplicative and additive type errors, through a non convex optimization problem. They established non asymptotic error bounds for the estimator and showed that a gradient descent based algorithm makes the estimator converge in polynomial time, in a neighborhood of the set of all global optimizer. We note here that for most, if not all of the aforementioned articles, measurement errors are applied on the design matrix R_c. In our model (3.9), the noise is affecting the variable multiplicative of the function f. To the best of our knowledge, we have not seen an article treating of measurement errors while considering this type of noise.

Looking closely at (3.5), we realize that it provides us with a very specific model which might find its roots in the family of partial linear models. The latter are characterized by regression of the type $Y = \widetilde{X} \cdot \beta + v(Z) + \epsilon$ where β is the variable of interest, $\widetilde{X}$ a matrix and the function v is unknown. These models have been explored in the literature, usually through the use of polynomial splines, for estimating the non-parametric part v.The authors in [47] used such a method and were able to derive the asymptotic normality of the estimators. Looking also at (3.3) and (3.4), we also realize that it is a regression model which shares some similarities with problems based on the estimation of linear functionals. We can cite for example [17]. In the latter article, the authors considered the regression model $Z_i = X_i + \epsilon$, where $(X_i)_{i\in\mathbb{N}}$ and $(\epsilon_i)_{i\in\mathbb{N}}$ are independent sequences of real valued random variables. They were interested in the estimation of linear functionals linked to the unknown density function of X_i, studying specifically the rate of convergence of its quadratic risk. As previously mentioned, instead of using a spline methodology, we relied on a decomposition in a Hilbert space to provide a desirable form to our unknown functions and hence derived their estimators. We note also that the partial linear models have been studied in the literature when subject to measurement errors such as [45]. They used a non convex corrected penalized least square, as well as a penalized quantile regression to calculate its estimators. They established their convergence rate and asymptotic normality property. A lot of similarities can be drawn between our model and the partial linear model but as mentioned before, they are not identical. The studies that have been achieved so far, when the model is subject to measurement errors, usually assumed that the linear segment of the *partial linear model* has a fixed dimension. In our case, we do not. In fact,

we assume that m can be as large as possible hence adding another layer of complexity and allowing to derive better estimator of the unknown function g. Furthermore, in contrary to both articles mentioned above, we used a *Lasso* type methodology to find our estimators.

The rest of this article is organized as follows. Section 3.2 establishes convergence rates of the estimator of β under two different penalty functions used to define problem (3.6). Section 3.3 establishes convergence rate of the estimator of β based on the errors-in-variables model (3.9). Two different penalty functions are also considered to define the optimization problem from their model. Section 3.4 provides a simulation study which exhibits the aforementioned characteristics of our estimators for both models (3.6) and (3.9). We also compare these results to the ones computed through the *TMLE* estimator. In Section 3.5, we present an application of our methodology in finance by finding an optimal portfolio which replicates the daily return of the $S\&P500$ index. All proofs are developed in the Appendix 3.6.

3.2 High dimensional regression model

Let us consider the model (3.6). The variable ϵ is a vector of size n, whose elements are drawn from a centered Gaussian distribution of variance σ^2.

3.2.1 Convergence rate for globally penalized Lasso

The penalized optimization problem derived from (3.6) is given by

$$\hat{\beta} = \begin{pmatrix} \hat{\eta} \\ \hat{\theta} \end{pmatrix} = \arg\min_{\beta \in \mathbb{R}^p} \left\{ \frac{1}{2n} \sum_{i=1}^{n} (Y_i - (\mathbb{X}\beta)_i)^2 + \lambda \|\beta\|_1 \right\}, \tag{3.10}$$

where $\lambda > 0$. In order to bound the estimation error of the variables of interest, a compatibility condition is needed. It was first coined in [74] for the resolution of a *Lasso* based optimization problem. As we see below, it is an additional constraint, not always verifiable, applied to the matrix $\mathbb{X}$ such as to derive fast convergence rate of the estimator. It is a well known condition, which has been extensively used in the literature (see [67]).

Definition 1. Let us consider the $n \times p$ matrix $\mathbb{X}$. The compatibility condition holds for the index set S_0, having s_0 elements, if for some $\phi_0 > 0$ and all $\gamma \in \mathbb{R}^p$ such that $\|\gamma_{S_0^c}\|_1 \leq 3\|\gamma_{S_0}\|_1$, it holds that

$$\|\gamma_{S_0}\|_1^2 \leq \frac{s_0 \|\mathbb{X}\gamma\|_2^2}{n\phi_0^2}. \tag{3.11}$$

Proposition 3.1. *Assume that the variable X takes values in a compact space. Furthermore, let us assume that the compatibility condition (3.11) holds. Then, with probability larger than $1 - 2\exp(\frac{-t^2}{2})$ and $\lambda = 2\sigma\|X\|_\infty\sqrt{\frac{t^2+2\log(p)}{n}}$ with arbitrary $t > 0$, the estimator $\hat{\beta}$ of (3.10) verifies*

$$\frac{1}{2n}\|\mathbb{X}(\hat{\beta}-\beta)\|_2^2 + \lambda\|\hat{\beta}-\beta\|_1 \leq 4\sigma^2\|X\|_\infty^2\frac{8s}{\phi_0^2}\cdot\frac{t^2+2\log(p)}{n}.$$

As a consequence, with the same probability,

$$\|\hat{\beta}-\beta\|_1 \leq 2\sigma\|X\|_\infty\cdot\frac{8s}{\phi_0^2}\cdot\sqrt{\frac{t^2+2\log(p)}{n}}.$$

Similar results could be inferred directly on the distinct estimators $\hat{\eta}$ and $\hat{\theta}$ through a unique equation. The demonstration of the proposition has been inserted in the Appendix, for reader convenience. The convergence rate of the estimator is then of the order of $s\sqrt{\frac{\log(p)}{n}}$.

3.2.2 Convergence rate for partially penalized Lasso

We study here the model (3.6) by only penalizing the variable θ. As mentioned before, $f(W)$ can be viewed as the partial derivative in X of the function $g(X,W)$. As such, it is natural to assume that we need more information (hence, more coefficients) in order to fully capture its characteristics. The corresponding penalization problem is then given by

$$\begin{pmatrix}\hat{\eta}\\ \hat{\theta}\end{pmatrix} = \arg\min_{\eta,\theta}\left\{\frac{1}{2n}\sum_{i=1}^{n}\left[Y_i - (\Psi_m^i\cdot\eta + X_i(\Psi_k^i\cdot\theta))\right]^2 + \lambda\|\theta\|_1\right\}. \tag{3.12}$$

It can be solved in two steps. We first find the estimator $\hat{\eta}$ which minimizes (3.12), for any θ. It corresponds to a classical *least square estimator*. We assume here that $m \leq n$ and that the matrix $\Psi_m^\top\cdot\Psi_m = n\cdot I_{m\times m}$, with I being the identity matrix. Hence,

$$\hat{\eta}(\theta) = \frac{1}{n}\Psi_m^\top\cdot(Y - (D_X\Psi_k)\cdot\theta). \tag{3.13}$$

The penalization problem (3.12) can then be rewritten using (3.13) as

$$\hat{\theta} = \arg\min_{\theta}\left\{\frac{1}{n}\|V\cdot\left(Y - \widetilde{\Psi}_k\cdot\theta\right)\|_2^2 + \lambda\|\theta\|_1\right\}, \tag{3.14}$$

where $V = \left(I_{n\times n} - \frac{1}{n}\Psi_m\cdot\Psi_m^\top\right)$ and $\widetilde{\Psi}_k = D_X\Psi_k$. Note that V is an orthogonal projection ($V^2 = V$, with V symmetric) and that $rank(V) = n - m$. Thus, the data $V\cdot Y$ lives in the image space of V.

Proposition 3.2. *Let us denote by s_θ the sparsity factor of the variable θ. Let us assume that the compatibility condition (3.11) holds. Thus, with probability larger than $1-2\exp(-\frac{t^2}{2})$ and for $\lambda = 2\sigma\|X\|_\infty\sqrt{\frac{t^2+2\log(k)}{n}}$, the estimator $\hat{\theta}$ is such that*

$$\frac{1}{2n}\|V\widetilde{\Psi}_k\cdot(\hat{\theta}-\theta)\|_2^2+\lambda\|\hat{\theta}-\theta\|_1 \leq 4\sigma^2\|X\|_\infty^2\cdot\frac{8s_\theta}{\phi_0^2}\cdot\frac{t^2+2\log(k)}{n}$$

for some arbitrary $t>0$. Hence,

$$\|\hat{\theta}-\theta\|_1 \leq 2\sigma\|X\|_\infty\frac{8s_\theta}{\phi_0^2}\sqrt{\frac{t^2+2\log(k)}{n}}.$$

Proposition 3.2 gives the expected rate related to the sparsity s_θ. Furthermore, we note that the rate depends on $s_\theta\sqrt{\frac{\log(k)}{n}}$. However, the estimator of η suffers from the estimation of θ which acts here as a nuisance parameter.

$$\begin{aligned}
E\|\hat{\eta}(\hat{\theta})-\eta\|_2^2 &= E\|\frac{1}{n}\Psi_m^\top(Y-\widetilde{\Psi}_k\hat{\theta})-\eta\|_2^2\\
&= E\|\frac{1}{n}\Psi_m^\top(\Psi_m\eta+\widetilde{\Psi}_k\theta+\epsilon-\widetilde{\Psi}_k\hat{\theta})-\eta\|_2^2\\
&= E\|\frac{1}{n}\Psi_m^\top\epsilon+\frac{1}{n}\Psi_m^\top\widetilde{\Psi}_k(\theta-\hat{\theta})\|_2^2\\
&\leq \frac{2}{n^2}\sigma^2 Tr(\Psi_m\Psi_m^\top)+\frac{2}{n^2}E\|\Psi_m^\top\widetilde{\Psi}_k(\theta-\hat{\theta})\|_2^2
\end{aligned}$$

where Tr is the Trace of the matrix. On one hand, we have $\frac{1}{n}Tr(\Psi_m\Psi_m^\top)=m$ and on the other hand,

$$\begin{aligned}
\frac{2}{n^2}E\|\Psi_m^\top\widetilde{\Psi}_k(\theta-\hat{\theta})\|_2^2 &\leq \frac{2}{n^2}E[(\theta-\hat{\theta})^\top\widetilde{\Psi}_k^\top\Psi_m\Psi_m^\top\tilde{\Psi_m}(\theta-\hat{\theta})]\\
&\leq \frac{2}{n}E\|\widetilde{\Psi}_k(\theta-\hat{\theta})\|_2^2\\
&\leq \frac{2}{n}\|X\|_\infty^2E\|\Psi_k(\theta-\hat{\theta})\|_2^2\\
&\leq 2\|X\|_\infty^2E\|\theta-\hat{\theta}\|_2^2\\
&\leq 2\|X\|_\infty^2E\|\theta-\hat{\theta}\|_1^2 \quad \text{since for all } y, \|y\|_2\leq\|y\|_1.
\end{aligned}$$

However,

$$\begin{aligned}
E\|\theta-\hat{\theta}\|_1^2 &= \int_0^\infty P\left(\|\theta-\hat{\theta}\|_1^2\geq u\right)du\\
&\leq 2\int_0^\infty \exp\left(-\frac{1}{2}\left(\frac{nu}{4\sigma^2\|X\|_\infty^2}\left(\frac{\phi_0^2}{8s_\theta}\right)-2\log(k)\right)\right)du\\
&\leq 2\frac{8\sigma^2\|X\|_\infty^2}{n}(\frac{8s_\theta}{\phi_0^2})^2k.
\end{aligned}$$

Finally, we can conclude that

$$E\|\hat{\eta}(\hat{\theta})-\eta\|_2^2\leq\frac{2\sigma^2m}{n}+4k\cdot\frac{8\sigma^2\|X\|_\infty^4}{n}\left(\frac{8s_\theta}{\phi_0^2}\right)^2.$$

Discussion. We emphasize here that in both sections above, β is an aggregate of two signals : η and θ. In Section 3.2.1, the estimators of these variables of interest were computed simultaneously. Hence, the sparsity inherent in the *Lasso like procedure* is present by design in both variables. However, in Section 3.2.2, the objective was to estimate both signals where one is sparse, $\hat{\theta}$. The convergence rate of the estimator $\hat{\theta}$ is similar in both cases. However, the convergence rate of $\hat{\eta}$ is negatively impacted when $\hat{\eta}$ is function of $\hat{\theta}$.

3.3 High dimensional regression model with errors-in-variables

All proofs are developed in Section 3.6. Let us consider the model (3.9) with $g(X,W)$ in (3.4) and such that $g \in \mathcal{S}(s_m, m)$ and $f \in \mathcal{S}(s_k, k)$. It can be written as

$$\begin{cases} Y &= g(W) + Xf(W) + \epsilon \\ Z &= X + \nu \end{cases} \tag{3.15}$$

where the variables $\nu \perp X$, $\epsilon \perp \nu$ and $W \perp \nu$. Moreover, elements of ν are drawn from a centered Gaussian distribution with variance μ^2. However, similar results can be obtained for more general sub-exponential distributions. The above set of equations can be rewritten as:

$$\begin{cases} Y &= \mathbb{X} \cdot \beta + \epsilon \\ \mathbb{Z} &= \mathbb{X} + \mathbb{K} \end{cases} \tag{3.16}$$

where $\beta = \begin{pmatrix} \eta \\ \theta \end{pmatrix}$, $\mathbb{X}$ is the matrix introduced in (3.6) and $\mathbb{K}$ is given by

$$\mathbb{K} = \begin{bmatrix} 0 & \dots & 0 & \nu_1 \cdot \Phi_1(W_1) & \dots & \nu_1 \cdot \Phi_k(W_1) \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \dots & 0 & \nu_n \cdot \Phi_1(W_n) & \dots & \nu_n \cdot \Phi_k(W_n) \end{bmatrix}.$$

Under a certain set of assumptions, [19] established the consistency, as well as the convergence rate of the estimator of our variable of interest. The above methodology seems to be the most natural way to resolve (3.9). However, we use the specificities of our model to provide a Lasso type estimator of $\beta = \begin{pmatrix} \eta \\ \theta \end{pmatrix}$. From now on, we consider that there exists $L > 0$, such that :

$$\|\Psi_k(W)\theta\|_2^2 < L, \quad \text{for all } W \text{ of our data structure } O. \tag{3.17}$$

3.3.1 Convergence rate for globally penalized corrected Lasso

One could assume that even with the presence of measurement errors, the optimization problem could be written exactly as in (3.10), where the variable X is to be replaced by Z. Nonetheless, the regularization problem needs to be adjusted. In fact,

$$\begin{aligned} E[\|Y-(\Psi_m\eta+(D_Z\Psi_k)\theta)\|_2^2] &= E[\|Y-(\Psi_m\eta+(D_X\Psi_k)\theta)\|_2^2]+E[\|(D_\nu\Psi_k)\theta\|_2^2] \\ &\quad -2E[\epsilon^\top(D_\nu\Psi_k)\theta] \\ &= E[\|Y-(\Psi_m\eta+(D_X\Psi_k)\theta)\|_2^2]+n\mu^2E[\|\theta\|_2^2]. \end{aligned}$$

From the above, it appears that the regularization needs to be corrected by the factor $n\mu^2\|\theta\|_2^2$. As such, the estimators $\hat{\eta}$ and $\hat{\beta}$ are then defined by

$$\begin{pmatrix} \hat{\eta} \\ \hat{\theta} \end{pmatrix} = \arg\min_{\eta,\theta} \mathcal{L}(\eta,\theta) \tag{3.18}$$

with

$$\mathcal{L}(\eta,\theta) = \frac{1}{2n}\|Y-(\Psi_m\cdot\eta+(D_Z\Psi_k)\cdot\theta))\|_2^2-\frac{1}{2}\mu^2\|\theta\|_2^2+\lambda_1\|\eta\|_1+\lambda_2\|\theta\|_1,$$

where $\lambda_1\geq 0$ and $\lambda_2\geq 0$. We denote by $\widetilde{\mathbb{X}}$, the equivalent of $\mathbb{X}$ in (3.7) where the variable X has been replaced by its counterpart Z. We also consider the vector $\xi=(0,\cdots,0,1,\cdots,1)$, of size p, where the first m elements are equal to 0 and the last k are all equal to 1. Decomposing the loss function, we obtain:

$$\begin{aligned} \mathcal{L}(\eta,\theta) &= \frac{1}{2}\left(\frac{1}{n}\|\Psi_m\eta+(D_Z\Psi_k)\theta\|_2^2-\mu^2\|\theta\|_2^2\right)-\frac{1}{n}\langle Y,\Psi_m\eta+(D_Z\Psi_k)\theta\rangle \\ &\quad +\frac{1}{2n}\|Y\|_2^2+\lambda_1\|\eta\|_1+\lambda_2\|\theta\|_1 \\ &= \frac{1}{2}\left(\frac{1}{n}\|\widetilde{\mathbb{X}}\beta\|_2^2-\mu^2\|\zeta\beta\|_2^2\right)-\frac{1}{n}\langle Y,\widetilde{\mathbb{X}}\beta\rangle+\lambda\|\beta\|_1+\frac{1}{2n}\|Y\|_2^2 \quad \text{with } \lambda=\begin{pmatrix}\lambda_1\\ \lambda_2\end{pmatrix} \end{aligned}$$

where $\beta=\begin{pmatrix}\eta\\ \theta\end{pmatrix}$ and $\zeta=diag(\xi)$. Thus

$$\mathcal{L}(\eta,\theta)=\frac{1}{2}\beta^\top\Gamma\beta-\gamma^T\beta+\lambda\|\beta\|_1+\frac{1}{2n}\|Y\|_2^2$$

where $\Gamma=\frac{1}{n}\widetilde{\mathbb{X}}^\top\widetilde{\mathbb{X}}-\mu^2\zeta^\top\zeta$ and $\gamma=\frac{1}{n}\widetilde{\mathbb{X}}^\top Y$. The optimization problem becomes then

$$\hat{\beta}=\arg\min_{\beta\in\mathbb{R}^d}\left\{\frac{1}{2}\beta^\top\Gamma\beta-\gamma^T\beta+\lambda\|\beta\|_1\right\}.$$

For $\mu\neq 0$, the matrix Γ is certainly not positive definite. As such, we are dealing with a non convex quadratic problem. Furthermore, if Γ has negative eigenvalues, the problem is unbounded. Thus, we have to add additional constraints to the optimization problem. Hence,

$$\hat{\beta}\in\arg\min_{\beta\in\mathbb{R}^d}\frac{1}{2}\beta^\top\Gamma\beta-\gamma^T\beta+\lambda\|\beta\|_1, \quad \text{such that } \|\beta\|_1\leq c_0\sqrt{s_\beta} \tag{3.19}$$

where $c_0 > 0$ is large enough, and s_β is the sparsity of β. We note here that $s_\beta = s_\theta + s_\eta$, with s_θ and s_θ the sparsity attached to the variables θ and η. Furthermore, we define an assumption that is used in order to bound the estimation error. It is known as the restricted eigenvalues assumption and is discussed in [8].

Definition 2. For some integer s, such that $1 \leq s \leq p$, $t_0 > 0$, positive number c_0, the Restricted Eigenvalues (RE) condition holds for the index set $J_0 \in \{1, \cdots, p\}$ and for all $\delta \in \mathbb{R}^p_\star$ such that $|J_0| \leq s$ and $\|\delta_{J_0^c}\|_1 \leq c_0 \|\delta_{J_0}\|_1$ if we have:

$$\frac{\delta^\top \Gamma \delta}{\|\delta_{J_0}\|_2^2} \geq t_0. \tag{3.20}$$

The integer s here plays the role of an upper bound of the sparsity of the vector δ.

It is important to note that in high dimensional settings, for the least square loss function to be strongly convex, it requires the eigenvalues of the core matrix to be bounded away from 0. In our case, the matrix Γ can be rank deficient. As such, in order to resolve the optimization problem, we need to impose a strong convexity condition. We emphasize here the similarity that exists between the Restricted Eigenvalues (3.20) and the compatibility (3.11) conditions. The RE condition is actually stronger since its implies the compatibility's one. In fact, $\|\delta\|_2^2 \geq \|\delta_J\|_2^2 \geq \frac{1}{s}\|\delta_J\|_1^2$.

Theorem 3.3. *(**Prediction and Estimation Risk**) We assume that the variable X takes values in a compact space. Furthermore, we also assume that the Restricted Eigenvalues condition (3.20) holds. Then, given (3.17), there exists $t > 0$, $t_0 > 0$ and $\lambda_0 = \left(\|X\|_\infty(\sigma + \mu\sqrt{L}) + \mu\sqrt{L}\right)\sqrt{\frac{2(1+t)\log(p)}{n}} + \tau_0$, with*

$$\tau_0 = \max\left\{\frac{2\mu^2\sqrt{L}}{\sqrt{n}}\sqrt{t}, \frac{4\mu^2 L}{n}t\right\} + (\sigma^2 + \mu^2)\max\left\{\sqrt{\frac{t}{n}}, 2\delta\frac{t}{n}\right\},$$

where the constants δ and L are defined respectively in (3.8) and (3.17), and such that with $2\lambda_0 \leq \lambda$, the estimators $\hat{\eta}$ and $\hat{\theta}$ verify

$$\|\hat{\beta} - \beta\|_1 = \|\hat{\theta} - \theta\|_1 + \|\hat{\eta} - \eta\|_1 \leq 12\lambda \cdot \frac{s_\beta}{t_0}$$

with probability greater than $1 - 5\exp(-\frac{t}{2})$.

3.3.2 Convergence rate for partially penalized corrected Lasso

The optimization problem can not be the same as (3.14) with the presence of measurement errors. In fact,

$$\begin{aligned} E[\|V(Y - (D_Z\Psi_k)\theta)\|_2^2] &= E[\|V(Y - (D_X\Psi_k)\theta)\|_2^2] + E[\|V(D_\nu\Psi_k)\theta\|_2^2] \\ &\quad - 2E[\epsilon^\top V^\top (D_\nu\Psi_k)\theta] \\ &= E[\|V(Y - (D_X\Psi_k)\theta)\|_2^2] + \mu^2 E[\|V\Psi_k\theta\|_2^2]. \end{aligned}$$

We recall that $V = \left(I_{n\times n} - \frac{1}{n}\Psi_m \cdot \Psi_m^\top\right)$. As such, the initial optimization problem needs to be adjusted. The estimator is then given by

$$\hat{\theta} = \arg\min_\theta \frac{1}{2n}\|V\cdot(Y-(D_Z\Psi_k)\cdot\theta)\|_2^2 - \frac{1}{2n}\mu^2\|V\cdot\Psi_k\cdot\theta\|_2^2 + \lambda\|\theta\|_1, \tag{3.21}$$

where $\lambda > 0$. We introduce a loss function $\mathcal{L}(\theta)$ defined by $\hat{\theta} = \arg\min_\theta \mathcal{L}(\theta)$. Decomposing the loss function gives us

$$\begin{aligned}\mathcal{L}(\theta) &= \frac{1}{2n}\|V\cdot(Y-D_Z\Psi_k\cdot\theta)\|_2^2 - \frac{1}{2n}\mu^2\|V\cdot\Psi_k\cdot\theta\|_2^2 + \lambda\|\theta\|_1 \\ &= \frac{1}{2}\theta^\top\left(\frac{1}{n}(VD_Z\Psi_k)^\top(VD_Z\Psi_k) - \frac{1}{n}\mu^2(V\Psi_k)^\top(V\Psi_k)\right)\theta \\ &\qquad - \frac{1}{n}(VY)^T(VD_Z\Psi_k)\theta + \lambda\|\theta\|_1 + \|VY\|_2^2.\end{aligned}$$

Since $\|VY\|_2^2$ is not a function of θ, we have

$$\hat{\theta} = \arg\min_\theta \frac{1}{2}\theta^\top\Gamma\theta - \gamma^\top\theta + \lambda\|\theta\|_1,$$

where $\Gamma = \frac{1}{n}(V\widetilde{\Psi}_k)^\top(V\widetilde{\Psi}_k) - \frac{1}{n}\mu^2(V\Psi_k)^\top(V\Psi_k)$ and $\gamma = \frac{1}{n}(V\widetilde{\Psi}_k)^\top Y$. The matrix Γ is degenerate in the sense that it is not positive definite. As such, the eigenvalues of the optimization problem can be negative, thus the problem is unbounded and as a consequence we need to add more constraints to the estimator. Hence following [48], the optimization problem is replaced by:

$$\hat{\theta} \in \underset{\|\theta\|_1\le c_0\sqrt{s_\theta}}{\arg\min}\left\{\frac{1}{2}\theta^\top\Gamma\theta - \gamma^\top\theta + \lambda\|\theta\|_1\right\}, \quad \text{with } c_0 > 0. \tag{3.22}$$

Theorem 3.4. *(**Prediction and Estimation Risk**) We assume that the variable X takes values in a compact space. Furthermore, let us assume that the Restricted Eigenvalues condition (3.20) holds. Then, given (3.17), there exists $t > 0$, $t_0 > 0$ and $\lambda_0 = \|X\|_\infty\left(\sigma + \mu\sqrt{L}\right)\sqrt{\frac{2(1+t)\log(k)}{n}} + \tau_0$, with*

$$\tau_0 = \max\left\{\frac{2\mu^2\sqrt{L}}{\sqrt{n}}\sqrt{t}, \frac{4\mu^2 L}{n}t\right\} + (\sigma^2+\mu^2)\max\left\{\sqrt{\frac{t}{n}}, 2\delta\frac{t}{n}\right\},$$

where the constants δ and L are defined respectively in (3.8) and (3.17), and such that with $\lambda \ge \frac{\lambda_0}{2}$, the estimator $\hat{\theta}$ verifies

$$\|\hat{\theta}-\theta\|_1 \le 12\lambda\frac{s_\theta}{t_0},$$

with probability greater than $1 - 5\exp(-\frac{t}{2})$.

3.4 Simulation study

In this section, we show the accuracy of our estimation methodology, using practical cases. Two main risk metrics are used: *Prediction Risk* and *Estimation Risk*.

3.4.1 Evaluation of estimates

Let us consider a sample data containing the observations $O^i = (Y_i, X_i, W_i)$ of $P^\star$, a known distribution. We assume that $W \in [0,1]^d$, $X \in [0,1]$ and $Y \in \mathbb{R}$, with $d \geq 1$. We know that the observations X_i, Y_i and W_i are linked through (3.3), (3.4) and also (3.6). For what follows, the variables W and X are generated from the $[0,1]-$uniform distribution. We consider a single set of basis functions, for the decomposition of the functions f and g : *the Fourier Basis*. The simulation described here can be adapted with other basis functions such as the *Spline Basis*.

To evaluate the accuracy of the estimator $\hat{f}(W) = \Psi_k(W)\cdot\hat{\theta}$, we computed its *Prediction Risk* value, for several sample sizes through a two-step cross-validation process. Let us denote by n, the data sample size. We partition the sample in l distinct subsets O^{subset}, of equal size. First step: We set aside one of the $O^{subset}_{\{i\}}$, which are used as the test set, and we use the $l-1$ remaining subsets as training sets. Using the latter, we then compute the estimator $\hat{\theta}_{\{-i\}}$. Second Step: we use the $O^{subset}_{\{i\}}$ (test set - left out from the initial estimation), to determine how close the estimator $\hat{\theta}_{\{-i\}}$ is to its real value. These two steps are repeated successively l times. For each iteration, we calculated the difference between the estimator and the true value function using the test set. The empirical average of these l distances give us the desired *Prediction Risk* measure. In brief, the *Prediction Risk* (PR) value of a sample of size n is given by

$$PR = \frac{1}{l}\sum_{i=1}^{l}\frac{1}{|O_{\{i\}}|}\|f(W_{O^{subset}_{\{i\}}}) - \hat{f}_{n-i}(W_{O^{subset}_{\{i\}}})\|_2^2, \tag{3.23}$$

where $|O_{\{i\}}|$ is the number of elements of the subset $O^{subset}_{\{i\}}$, and $\hat{f}_{n-i}(W) = \Psi_k(W)\cdot\hat{\theta}_i$.

For a single sample size, we computed the *Prediction Risk* $u(\geq 20)$ times. It allowed us to evaluate the confidence interval associated with the measure. It is important to note that for a convergent estimation procedure, the PR should converge to 0 as n increases. We show below results for various cases : univariate ($W \in \mathbb{R}$) , bivariate ($W \in \mathbb{R}^2$) and lastly an univariate case where X is measured with errors. Furthermore, for each estimator $\hat{f}$, corresponding to a given sample size, we compared it to its corresponding true function f. To achieve this, we constructed a grid linked to the space on which lies the variable W. We then compared the ordinate difference between the estimator and the true function. The empirical square difference gave us a measure of our *Risk Estimation*. In brief, if we consider a grid κ and an estimator $\hat{f}_j$, the *Estimation Risk* of our function f is given by

$$ER = \frac{1}{Vol(\kappa)}\sum_{\kappa\subset[0,1]}(f(W_j) - \hat{f}_j(W_i))^2, \tag{3.24}$$

where $Vol(\kappa)$ represents the volume of our equidistant grid (i.e.: , if $\kappa = [0,1] \subseteq \mathbb{R}$ and we consider a grid step given by $\delta = 0.01$, then $Vol(\kappa) = 100$. Similarly, if $\kappa = [0,1]^2 \subseteq \mathbb{R}^2$ and we consider a grid step given by $\delta = 0.01$, then $Vol(\kappa) = 100^2$).

Beyond these risk measures, we also introduced a *breaking point* concept. It consists of lowering iteratively as much as possible one of the coefficient associated with the definition of the true function f and identifying the level at which the estimator is not able to recover it, for a given sample size. This is an important exercise as it allows us to test the quality of fit for extreme values. We now present in details three examples, to put in perspective the notions we just described.

Univariate covariate W case ($\mathbf{d = 1}$)

Let us consider the functions $f(W_i) = 20 \cdot \cos(2\pi \cdot 5W_i) + 10 \cdot \sin(2\pi \cdot 4W_i) + \sqrt{5}\sin(2\pi \cdot 8W_i) - 7\sin(2\pi \cdot 11W_i)$ and $g(W_i) = W_i$. The variables W and X are drawn from a $[0,1]-$uniform distribution. We generated data samples of size $n = \{50, 100, 150, 200, 250, 300\}$ and used each sample to compute the estimator of f. As shown in figure 3.1a, we computed the PR corresponding to each sample size, with $u = 30$. Furthermore, we used a *Fourier* basis of size $p = 450 (m = 50; k = 400)$. We note here that we chose a greater value for k compared with m because we wanted to put more emphasis on the estimation of the function f. One can see that for a very small sample size, the methodology provides less accurate estimators. However, it improves rapidly as the sample size reaches $n = 100$.

We also show in figure 3.1b that the *Estimation Risk* changes with increasing sample size. As before, it converges towards 0, as expected. The *Estimation Risk* was computed using a grid with an incremental step $\delta = 0.05$. It is important to note that even for small sample size (i.e: $n = 50$), where the results of both Estimation Risk and Prediction Risk appear relatively off compared to the other samples, the resulting estimator is still of good enough quality. We show in figure 3.2 the different estimators of f per sample size.

Remark: We tried our methodology with different variations of the function g. The results were very convincing, but are not shown here.

To further study the stability of the simulation, we have analyzed its behaviour through the aforementioned *breaking point* concept. In fact, we iteratively reduced the coefficient associated with the component "$\sin(2\pi \cdot 8W_i)$", called B, of the function f. We wanted to see the lowest coefficient for which the methodology was still able capture its impact on the response variable Y. We ran the simulation method for three different sample sizes. For each one, we iteratively increased the value of B, starting at 0. For each value of the coefficient B, we ran the simulation 100 times and counted the number of estimators in which the coefficient associated with the *breaking point* component was null.

What transpires from Table 3.1 is that, the methodology struggles to recover the effect played by the component B when its value is very small. This is magnified when the sample size is small as well. However, the greater the sample size the better the methodology is able to capture the component's effect, even for very small value. On the contrary, for

B	0.2	0.4	0.6	0.8	1	1.2	1.4	1.6	1.8	2
$n = 50$	94.9%	89.9%	86.2%	94.9%	81.9%	99.2%	34.5%	86.2%	45.5%	45.2%
$n = 100$	100%	100%	79.5%	37.5%	83.3%	79.8%	36.1%	1.7%	0.2%	0.034%
$n = 150$	100%	94.9%	91.8%	83.9%	91.1%	12.8%	1.3%	40%	13.4%	0%
$n = 200$	100%	94.7%	78.8%	69.8%	22.2%	18.4%	0%	0.04%	0%	0%
$n = 250$	99.8%	94.9%	74.8%	3.7%	8.5%	4.1%	0%	0.01%	0%	0%

Table 3.1 – *Breaking point analysis* - fraction of simulation runs (out of 100 replications) where the component associated with B is not part of the estimator - $f(W_i) = 20 \cdot \cos(2\pi \cdot 5W_i) + 10 \cdot \sin(2\pi \cdot 4W_i) + B \cdot \sin(2\pi \cdot 8W_i) - 7\sin(2\pi \cdot 11W_i)$

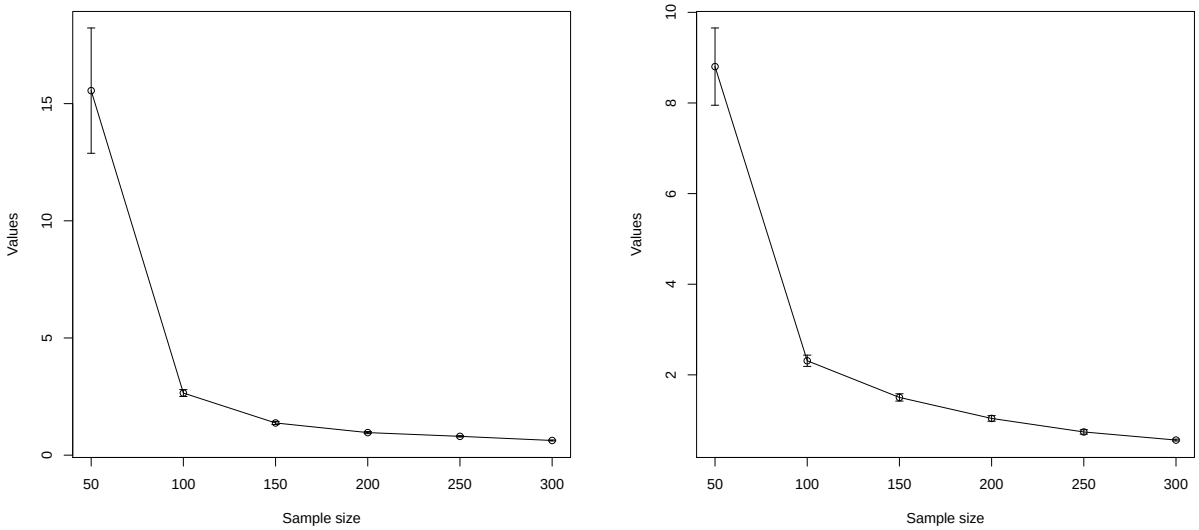
significantly high value of the coefficient, the methodology performs well, even for small sample size.

Bivariate covariates W case (d = 2)

Let us consider a function $f : \mathbb{R}^2 \to \mathbb{R}$ such that $f(W_i) = f(W_i^1, W_i^2) = 10 \cdot \cos(2\pi \cdot 2W_i^1) \cdot \cos(2\pi \cdot W_i^2) + 4 \cdot \sin(2\pi \cdot W_i^1) \cdot \cos(2\pi \cdot W_i^2)$, where $W_i \in [0,1]^2$. We assume the function g to be defined by $g(W_i) = g(W_i^1, W_i^2) = W_i^1 + W_i^2$. The elements of the variables W and X drawn from the $[0,1]$−uniform distribution as before. We computed the risk prediction values for several sample sizes $n = \{50, 100, 150, 200, 250, 300\}$ and with $u = 30$. As one can see in figure 3.3a and 3.3b, both *Estimation Risk* and *Prediction Risk* measures decrease as the sample size increases. As before, the first sample size doesn't perform well compared to the others. The estimation risk measure was computed using 2−dimension grid with an increment given by $\delta = 0.05$.

Univariate covariate W (d=1) with errors-in-variables

We put in perspective in this case the behaviour of the methodology assuming that the data is subject to measurement errors. As such, let us consider the function f given by $f(W_i) = 20 \cdot \cos(2\pi \cdot 5W_i) + 10 \cdot \sin(2\pi \cdot 4W_i) + \sqrt{5}\sin(2\pi \cdot 8W_i) - 7\sin(2\pi \cdot 11W_i)$ and $g(W_i) = W_i$. The sample data is based on (Y, Z, W) where $Z = X + \nu$, with ν is a centered normal distribution of standard deviation μ. The variable W is drawn from the $[0,1]$−uniform distribution, as well as the variable X. The latter, combined with ν allows us to derive the values of Z. We used several sample sizes which were $n = \{50, 100, 150, 200, 250, 300\}$ and $u = 30$. We also used three different values of $\mu = \{0.033, 0.066, 0.099\}$. This allowed us to test the accuracy of the methodology with increasing noise size.



(a) Evolution of the *Prediction Risk*.

(b) Evolution of the *Estimation Risk*.

Figure 3.1 – *Univariate covariate case* with function f from Section 3.4.1 - Confidence Intervals were computed through $u = 30$ replications for each sample size

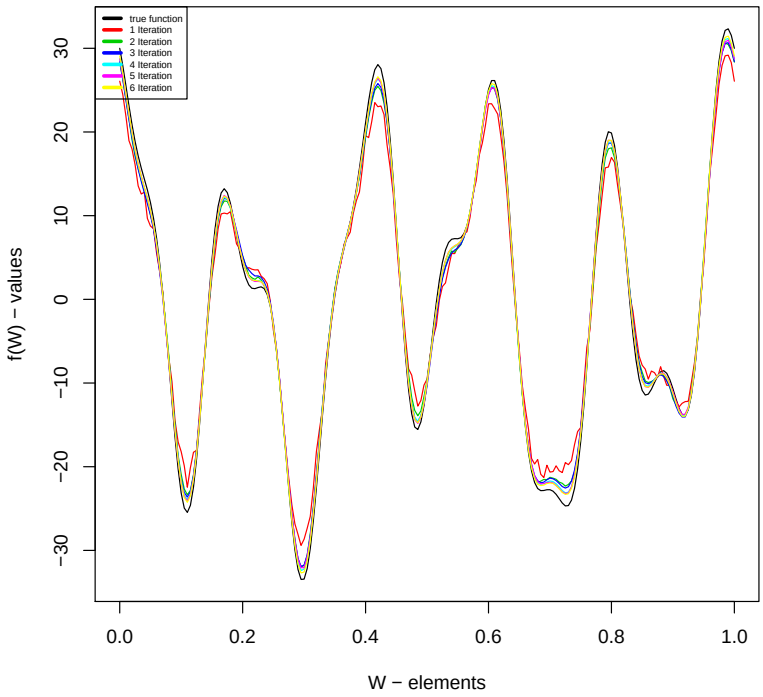
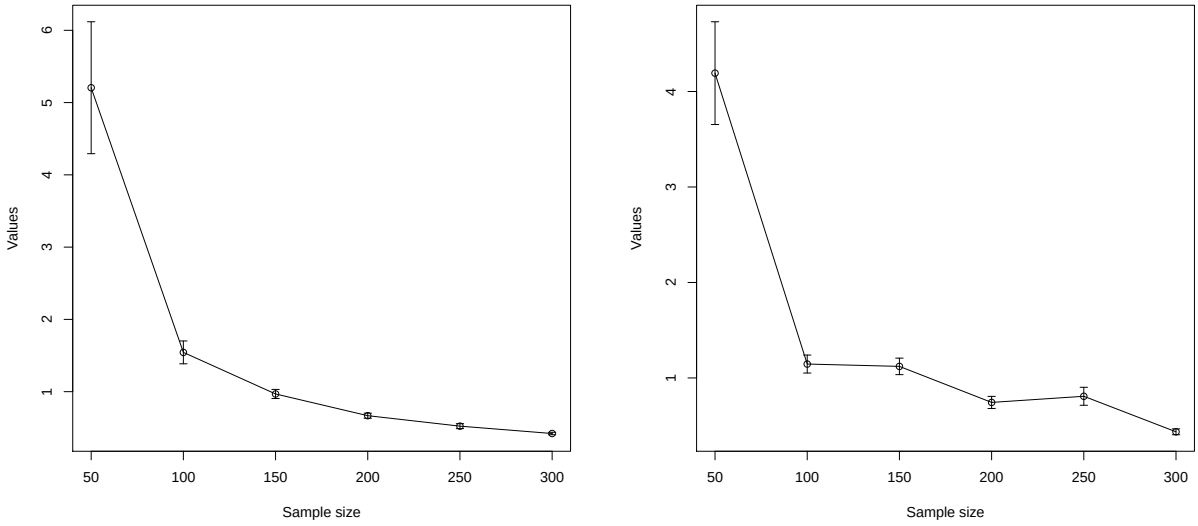


Figure 3.2 – *Univariate covariate case* - Evolution of the true function f from Section 3.4.1 and its corresponding estimators (each iteration linked to an estimator, which is an increase of the previous sample size by 50, with $n = 300$ for the last iteration).



(a) Evolution of the Prediction Risk

(b) Evolution of the Estimation Risk

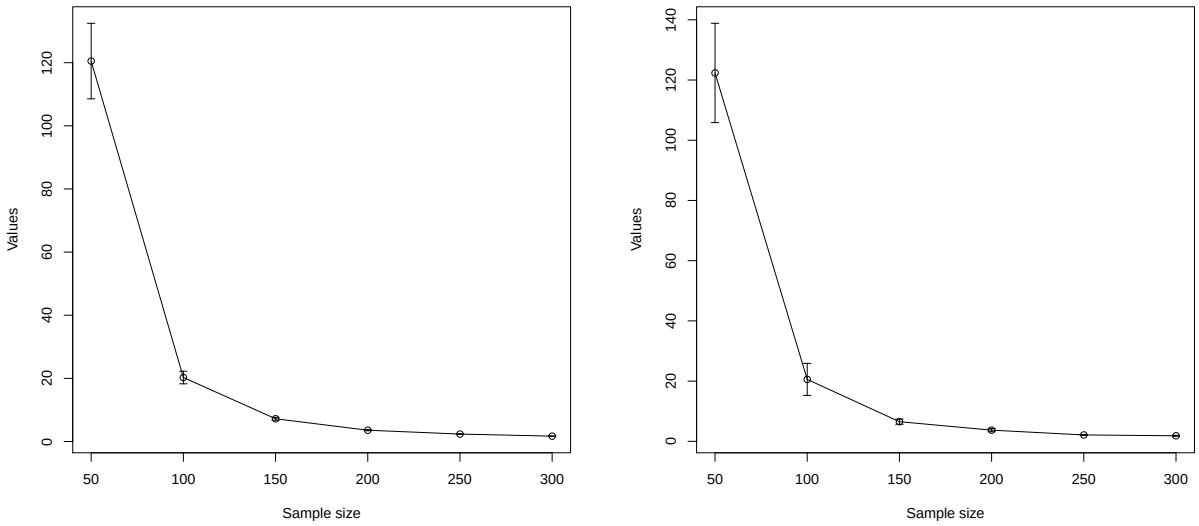
Figure 3.3 – *Bivariate covariate case* with function f from Section 3.4.1 - Confidence Intervals were computed through $u = 30$ replications for each sample size

We implemented a local version of the *Coordinate Descent algorithm* (see [34]), in order to solve the optimization problem and find the estimators. The implementation was completed in *C++*, and linked with the software *R* through packages *Rcpp* and *RcppArmadillo*. We significantly increased the computation time in comparison with a similar algorithm fully set up in *R*.

For all different values of μ, we can see from figures 3.4b, 3.5b, 3.7b, 3.4a, 3.5a and 3.7a that the methodology performs well. However, we note that the magnitude of our risk estimation level, as well as the prediction risk level, are greater than the ones shown in section 3.4.1. Furthermore, we also see that the algorithm needs more information, said differently more data points in the sample, to reach levels seen in section 3.4.1. As before, we can see in figures 3.6 and 3.8 that the estimators of f, for all sample sizes, are of good quality.

3.4.2 Comparison with the TMLE estimator φ_{CT}

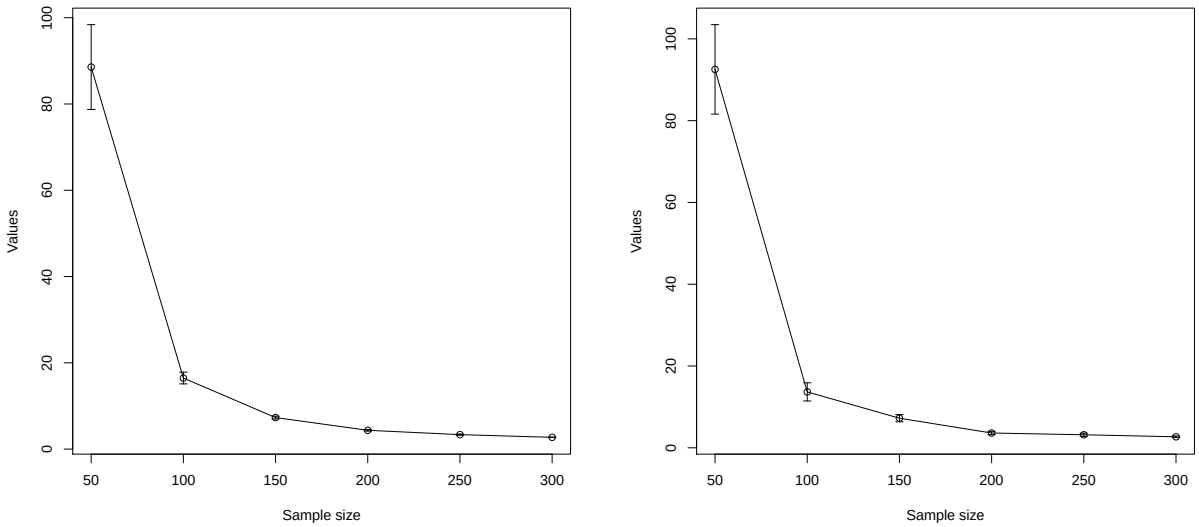
The TMLE (Targeted Minimum Likelihood Estimation) has been used, among others, in the literature to find etimators of the variable importance measure. The latter was introduced in [77] and further generalized in [22] by assuming the covariates to be continuous. It was further extended in the manuscript [23] by adjusting the definition of the parameter of interest such that it becomes multivariate. We recall here the definition of the variable



(a) Evolution of the Prediction Risk

(b) Evolution of the Estimation Risk

Figure 3.4 – *Univariate covariate case* with errors-in-variables and function f from Section 3.4.1 - Confidence Intervals were computed through $u = 30$ replications for various sample sizes - noise parameter given by $\mu = 3.33\%$



(a) Evolution of Prediction Risk

(b) Evolution of the Estimation Risk

Figure 3.5 – *Univariate covariate case* with errors-in-variable and function f from Section 3.4.1 - Confidence Intervals are computed through $u = 30$ replications for various sample sizes - noise parameter given by $\mu = 6.66\%$

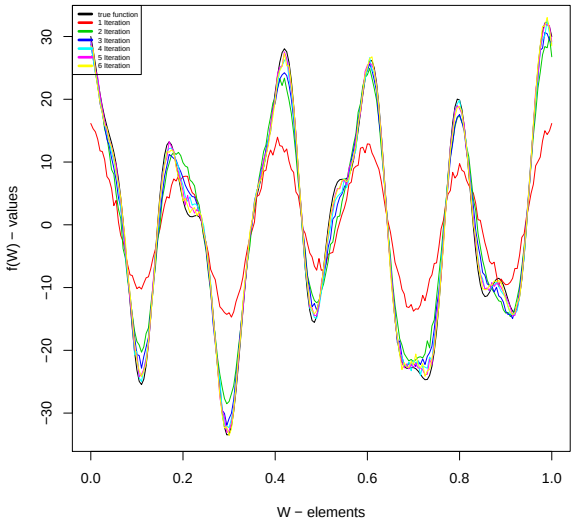
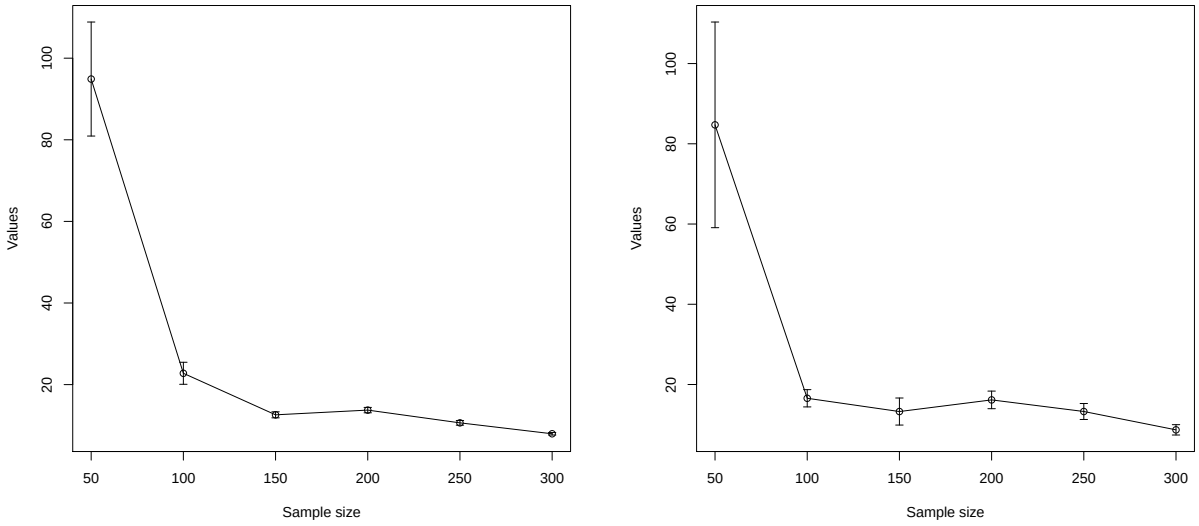


Figure 3.6 – *Univariate covariate* case with errors-in-variable - Evolution of the estimators of the function f (each iteration is linked to an estimator, which is an increase of the previous sample size by 50, with $n = 300$ for the last iteration) - noise parameter given by $\mu = 6.66\%$



(a) Evolution of *Prediction Risk*

(b) Evolution of the Estimation Risk

Figure 3.7 – *Univariate covariate case* with errors-in-variables - and function f from Section 3.4.1 - Confidence Intervals were computed through $u = 30$ replications for each sample size - noise parameter given by $\mu = 9.99\%$

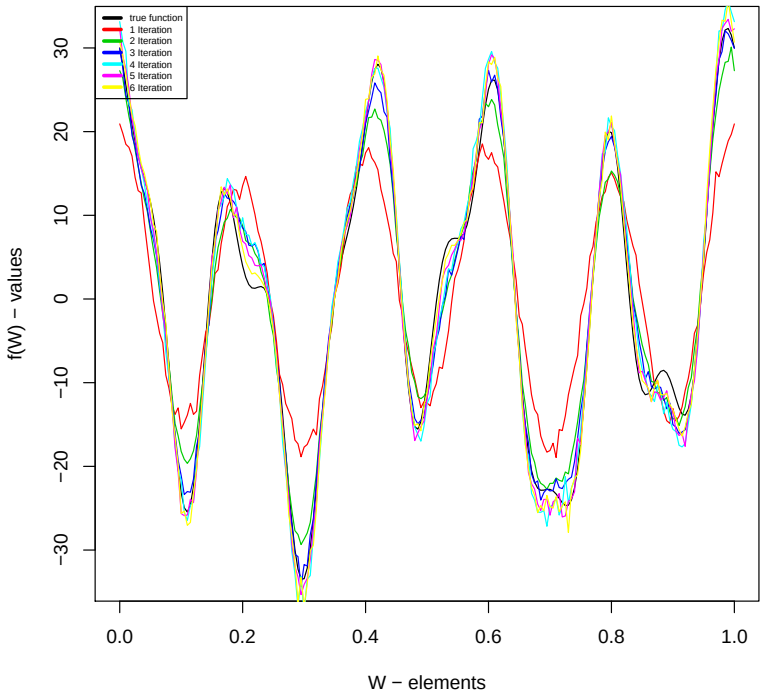


Figure 3.8 – *Univariate covariate case* with errors-in-variable - Evolution of the estimators of function f (each iteration is linked to an estimator, which is an increase of the previous sample size by 50, with $n = 300$ for the last iteration) - noise parameter given by $\mu = 9.99\%$

φ_{CT} (3.2) which is given by

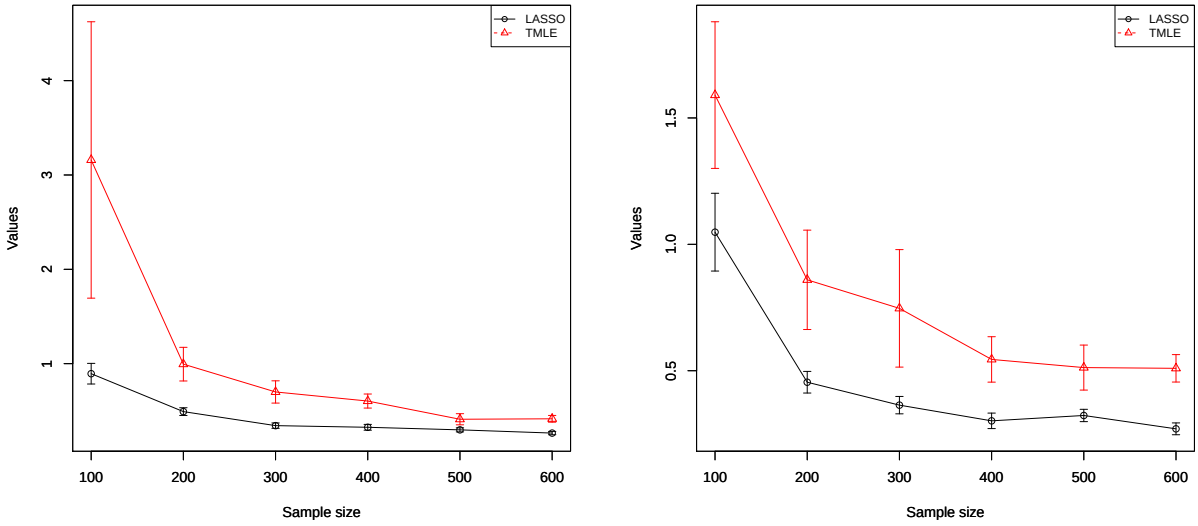
$$\varphi_{CT}(P) = \arg\min_{B\in\mathbb{R}^d} E_P\left[(Y - E_P(Y|X=0,W) - X\cdot f(B,W))^2\right],$$

where f is assumed known. This definition introduces a variable importance measure whose size is equal to d, the number of covariates since $\varphi_{CT}(P) \in \mathbb{R}^d$. Hence, it provides a granular view of the variable importance such that each element of the parameter of interest $\varphi_{CT}(P)$ can be linked to a covariate.

The TMLE procedure is an iterative method which relies on the computation of the efficient influence curve of the parameter of interest, in order to derive its consecutive estimators. The iterative procedure can be described in few steps. First, we define the parameter of interest as a smooth functional Ψ at P_0 (the true and unknown data-generating distribution). Second, using machine learning techniques, we compute the initial estimate, based on an empirical law P_n^0, which can be biased. Third, we define a submodel based on P_n^0 such that its score relies on the gradient of the functional Ψ. Fourth, we maximize the log-likelihood of the submodel such that we can build an update estimate P_n^k. In the final phase, the last two steps are iterated until the procedure converges to an updated $P_n^\star$. The final estimator is then given by $\varphi_{CT}^\star = \Psi(P_n^\star)$. It is shown, in [23], that the estimator $\varphi_{CT}^\star$ is consistent and asymptotically normal under mild conditions.

We would like in this section to evaluate the performance of a TMLE based estimator against the ones from the methodologies introduced in this document. We consider consecutively the model (3.3) and (3.9), while assuming that our function g is given by $g(W) = W$. Furthermore, the function f, is given by $f(W) = 20cos(2\pi W) + 5sin(2\pi 2W)$. We also ensure that (i) $P(X=0) > \delta$ with $\delta > 0$, in order to respect a constraint introduced in [22] and preserved in [23]. We note that such a direct comparison is possible because we chose a function f which is a linear combination of elements of a *Fourier* basis. As such, the TMLE estimator is then computing the coefficients associated with the function.

Through our tests, we have realized that the TMLE doesn't perform well in high dimensional settings and even more so when the parameter of interest is sparse. As such, we have restricted ourself to a compact case, which should ease the performance evaluation. We assume that the number of elements in our variable of interest is much lower than our sample size ($p \ll n$). We also use metrics previously introduced for our evaluation : *Prediction Risk* and *Estimation Risk*. We generate data samples of size $n = \{100, 200, 300, 400, 500, 600\}$ and decompose the function f and g using a *Fourier basis* such that $p = 10 (m = 5, k = 5)$. We note here that each data sample is used for both methods. Furthermore, similarly to the previous section, we repeat our procedure $u = 20$ times for each data sample size, in order to find error bounds for our metrics.



(a) Evolution of the Prediction Risk

(b) Evolution of the *Estimation Risk*

Figure 3.9 – Model without measurement errors using both TMLE and Lasso methodologies for function f

Model without measurement errors

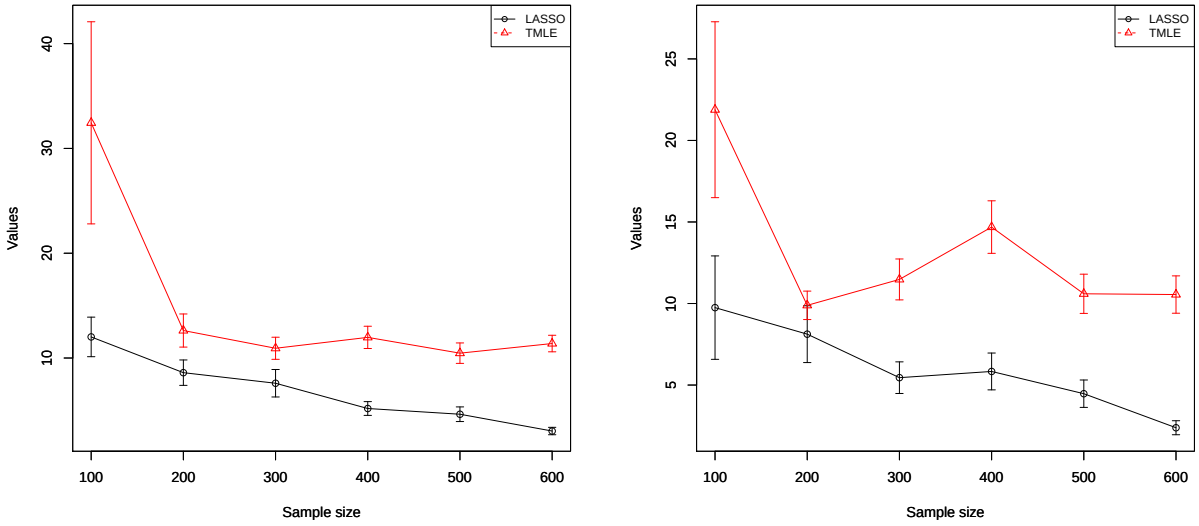
Let us consider the data sample named $\mathcal{O} = (X, W, Y)$. The variable W and X are drawn from a $[0,1]$-uniform distribution. We then force randomly 15% of values of X to be equal to 0, to be in line with condition (i). The variable Y is generated through the aforementioned functions f and g.

We see from figures (3.9b) and (3.9a) that the TMLE estimators perform relatively poorly compared to our estimators for very small sample size. However, as data increases, the performance of both methods tends to converge.

Performance evaluation for model with measurement errors

Let us consider the data sample named $\mathcal{O}_{err} = (Z, W, Y)$, where W is drawn from a $[0,1]$-uniform distribution. The variable Z is given by $Z = X + \nu$ where X is drawn from a $[0,1]$-uniform distribution, and ν is a centered Gaussian distribution with standard deviation $\mu = 20\%$. As before, we force 15% of the elements of Z to be equal to 0. The resulting variable Y will be generated through the function f and g.

We see from figures (3.10b) and (3.10a) that the TMLE performs poorly for small sample size and further struggle to improve even when we increase the sample. Clearly, it is biased for cases when measurement errors are added to the covariates. To the best of our knowledge, there are no bias corrections in the literature of TMLE for errors-in-variables



(a) Evolution of the *Prediction Risk*

(b) Evolution of the *Estimation Risk*

Figure 3.10 – Model with measurement errors using both TMLE and Lasso methodologies for the function f - noise parameter given by $\mu = 20\%$

models.

Discussion: The main objective in [22] is to find estimators of the variable of interest $\Psi(P)$ (3.2), using the TMLE methodology. The model developed in this document put more emphasis on finding estimator of the function f, defining the relationship that exists between the covariates W and the response variable Y. It was supposed known in [22, 23]. Assuming a limited amount of information on the characteristics of the function f, finding a good estimator of f, can lead to a good estimator of $\Psi(P)$. However, for an unknown function f, the estimator of $\Psi(P)$ can be derived through the metric *Prediction Risk*. As such, we made sure to include its characteristics in the simulation cases explored above.

3.5 Application to a real data set

It is very common in financial markets for an investor to build portfolios which attempt to track the return of an existing financial index. As such, the investor has to find a selection of securities which, bundle together, recover to a certain degree of accuracy the return and risk profile of the target index. This exercise doesn't have to be a full replication of the index in question. It should not, in most cases. In fact, a full replication, which consists of buying exactly the same amount of securities which are existing in the index, is impractical and not cost effective.

Let us consider a portfolio containing p securities, where the return of the i^{th} security at time t is denoted by $r_{\{i,t\}} = \frac{P_i^{t+1}}{P_i^t} - 1$, where P_i^t (respectively P_i^{t+1}) represents the closing price of the i^{th} security at time t (respectively $t+1$). Furthermore, let us consider a universe of n timestamps $\{t_1, \cdots, t_n\}$ where one is able to construct a $n \times p$ matrix of return, called R_p^n, of all the securities in the portfolio. We denote by w_i, with $i \in \{1, \cdots, p\}$, the amount of capital invested in the i^{th} security and by r_y, the $1 \times n$ vector of return of the target index. Thus, at time t, $r_{y,t} = \frac{P_{t+1}^y}{P_t^y} - 1$, where P_t^y (respectively P_{t+1}^y) is the price of the index at time t (respectively $t+1$). The optimal replicating portfolio is given by

$$\hat{w} = \arg\min_w \|r_y - R_p^n \cdot w\|_2^2,$$

where $w = (w_1, \cdots, w_p)$. It is a linear regression problem whose solution is known to be unstable. However, it can be improved by adding a $\mathbb{L}_1$ regularization to the problem. Indeed, it has several benefits among which the sparsity of the solution. When trying to replicate an index, investors are most often looking for portfolio with a low amount of securities. The reduction of all costs is key since they impact the overall return of the strategy.

As such, the replication problem is rewritten as

$$\hat{w} = \arg\min_w \|r_y - R_p^n \cdot w\|_2^2 + \lambda \|w\|_1 \tag{3.25}$$

such that $\lambda > 0$ and $\sum_{i=1}^p w_i = 1$. It can be further extended by taking into account errors-in-variables in the model.

An investor trying to replicate the return profile of an index needs to take into consideration not only the open and close prices of the constituents and the index, which are in most cases publicly known, but few other factors which could impact its net return. We can mention as an example the management fees, which are periodic payments paid to the fund investment advisor. We can also mention the transaction costs which are expenses incurred when buying or selling some constituents when attempting to rebalance a portfolio. Another point to consider is also the liquidity cost which is linked to the daily volatility of the financial instruments in the market. This price volatility can be substantial during periods of considerable market stress. As such, it is also important for investors to have access to live and historical intraday prices which sometimes come at a hefty cost. Some of these historical data are in some instances polluted. To incorporate the aforementioned elements in the regression model guiding the relationship between the return of an index and its constituents, one can assume that the model is subject to measurement errors. The article [62] used similar considerations.

As such, we observe a matrix $Z_p^n = R_p^n + \Sigma$, where Σ is considered to be a multivariate

Gaussian distribution. The replication problem can then be reformulated as

$$\begin{cases} \hat{w} & = \arg\min_w \|r_y - R_p^n \cdot w\|_2^2 + \lambda \|w\|_1 \\ Z_p^n & = R_p^n + \Sigma. \end{cases} \tag{3.26}$$

Let us consider that there exists a function g, element of a *Hilbert* space $\mathcal{H}$, such that $r_y = g(r) + \epsilon$, where ϵ is a white noise and $r = \{r_i\}_{i=1,\cdots,p}$ is the vector of return of all the constituents in the initial portfolio. The estimation of the function g, leads to the precise definition of the constituents we want to keep in the replicating portfolio, as well as their corresponding weight. We consider $\Phi = (\Phi_1, \cdots, \Phi_p, \cdots) = \{\Phi_j\}_{j=1}^{\infty}$, a complete *Legendre* orthonormal base in $\mathcal{H}$. Hence, following (3.5), there exists $m > 0$ and $\beta \in \mathbb{R}^m$ such that $g(w_i) = \sum_{j=0}^{m} \Phi_j(w_i) \cdot \beta_j$. Based on the matrix form developed in (3.6), we can rewrite (3.25) as

$$\hat{\beta} = \arg\min_{\beta} \|r_y - \widetilde{R}_m^n \cdot \beta\|_2^2 + \lambda \|\beta\|_1 \tag{3.27}$$

where $\widetilde{R}_m^n$ is a $m \times n$ matrix which correspond the expansion of R_p^n in the base Φ. We note that $m \gg n$. Similarly as above, (3.26) can then be rewritten as

$$\begin{cases} \hat{\beta} & = \arg\min_{\beta} \|y - \widetilde{R}_m^n \cdot \beta\|_2^2 + \lambda \|\beta\|_1 \\ \tilde{Z} & = \widetilde{R}_m^n + \Sigma. \end{cases} \tag{3.28}$$

The index of interest is called *Standard & Poor*'s 500 (also known as $S\&P500$). It is an American stock market index based on the capitalizations of 500 large companies having common stock listed on the *New York Stock Exchange* or the *NASDAQ* stock exchange. It is an extremely liquid equity index and certainly one of the most followed. Because of its broad constituency, it is considered as one of the best representation of the *U.S.* stock market. In order to categorize the majorly traded public companies, the term *GICS* was coined in 1999 by *MSCI* and *Standard & Poor's* for use in the financial community. *GICS*, which stands for *Global Industry Classification Standard*, consists of 11 sectors : *Energy, Materials, Industrials, Consumer Discretionary, Consumer Staples, Health Care, Financials, Information Technology, Telecommunication Services, Utilities* and *Real Estate*. The companies in the $S\&P500$ covered all 11 *GICS Sectors*. This point will be of further importance later on. The data used to perform the analysis was retrieved from the *yahoo* website (*https://finance.yahoo.com/quote/DATA/*). The list of constituents of the $S\&P500$ is also public. The dataset extracted only contained 375 of the full 500 that could be found in the aforementioned constituent's list. We now illustrate the notions developed above through three cases : synthetic target index, $S\&P500$ index with direct observations and with errors-in-variables, respectively.

Case 1: We start here by constructing synthetic target indices, that we aim to track. These synthetic indices are based on securities which are within the dataset we extracted.

We constructed a total of two synthetic indices. Each target index closing price S_t is the sum of its constituents closing prices. Using all S_t, we can infer the daily return of the synthetic target portfolio.

To produce replicating portfolios for both synthetic indices (see tables 3.2 and 3.3), we used the 375 constituents of the initial dataset, amongst them were the ones used to build the synthetic targets. We chose to only use two degree of freedom for the *Legendre Basis* decomposition of g. Two main reasons motivated this decision. First of all, it allows us to have a matrix $\tilde{R}^n_m$, which not only contains the return linked to each constituent but also to their cross product of size 2. We could then measure their impact, if any. Secondly, prior analysis taking into account higher degree of freedom showed us that is was not relevant for the study to go beyond that level. The above cross products allow one to quantify the role played, if any, by the interdependence between a pair of stocks. It is a feature mostly known in the literature as the *interaction term*, and has been used in econometrics (see [53]) and quantitative finance (see [33]).

We chose windows of one-year length for this study. The amount of data points available n, was then to remain well below the number of basis m. We extracted daily closing prices for our synthetic index and all the constituents of our portfolio and then deduced the daily discrete return rate for each one of them. Using *Legendre basis*, we built the matrix $\tilde{R}^n_m$, with $m = 375$, $n \approx 250$ (for each portfolio) and computed the measure associated with each element of our replicating portfolio. We emphasize here that the underlying *Lasso* methodology used also performs a variable selection. Hence, not only do we know what are the constituents to include in our replicating portfolio, but also the capital to associate to each of them. Furthermore, it outlined any interdependency between constituents, if any. The windows considered were: 2008 and 2014. The first one corresponds to the start of the financial crisis in which we experienced a lot of volatility in the financial markets. The second is considered as a *post crisis* period, where volatility was low. Those two different windows are important because they allow us to apply the method in two different regimes and hence help assess the quality of the estimation methodology.

For the period 2014, we obtained a replicating portfolio whose constituents are displayed in Table 3.5. It contains all the securities available in the synthetic portfolio (see Table 3.2). An intercept was produced with a β value of 0.00133, negligible then. For the period 2008, we obtained a replicating portfolio whose elements are in Table 3.4. The match, again, in this case is perfect as we retrieved the exact number of constituents which were in the synthetic portfolio (see Table 3.3). An intercept was produced with a β value of -0.00278, negligible as for the previous case. From the full dataset, we were able to identify the key drivers of our synthetic indices. Furthermore, we see that even in high volatility period, we are still very accurate in our reconstruction. These synthetic indices help demonstrate the correctness of the method. Thus we can conclude through simple yet convincing examples

Ticker Symbol	Security	GICS Sector
CSCO	Cisco Systems	Information Technology
HRS	Harris Corporation	Information Technology
XRX	Xerox Corporation	Information Technology

Table 3.2 – Synthetic Index 1 - based on securities from the same GICS Sector

Ticker Symbol	Security	GICS Sector
GPS	Gap Inc.	Consumer Discretionary
HST	Host Hotels & Resorts.	Real Estate
ORLY	O Ŕeilly Automative	Consumer Discretionary

Table 3.3 – Synthetic Index 2 - based on securities from different GICS Sector

that on a predefined framework, we can accurately identify the key drivers of a target index.

Case 2: We turn our attention to the $S\&P500$ index, which now represents our target index. Similarly to the prior case, two windows were considered: 2008 and 2014. The reasons for this choice are the same as previously discussed.

For the first period, results show a replicating portfolio containing 80 constituents (the full list is not shown here). All these elements can be grouped in 11 distinct *GICS Sectors*. They represent the exact number of unique *GICS sectors* in the original $S\&P500$ list of constituents. As such, the replicating portfolio does cover the full spectrum of market segments represented by the index. For the second period 2014, the replicating portfolio

Ticker Symbol	Security	GICS Sector	β value
GPS	Gap Inc.	Consumer Discretionary	0.04483074
HST	Host Hotels & Resorts	Real Estate	0.01420947
ORLY	O Ŕeilly Automative	Consumer Discretionary	0.1347531

Table 3.4 – Replicating Portfolio 2 - with their respective importance measure - year 2008

Ticker Symbol	Security	GICS Sector	β value
CSCO	Cisco Systems	Information Technology	0.06976
HRS	Harris Corporation	Information Technology	0.1707516
XRX	Xerox Corp.	Information Technology	0.1364902

Table 3.5 – Replicating Portfolio 1 - with their respective importance measure - year 2014

has 139 constituents. As for the prior period, it covers as well all 11 sectors present in the index list of constituents. We represent in Table 3.6 the number (in percentage) of securities linked to each *GICS Sector*. Thus, we show the distribution of our replicating portfolios.

Having constructed these portfolios, we were then interested in measuring their stability. Given the scaling profile of these replicating portfolios (β values corresponding to each constituent) provided by the decomposition, we wanted to know how the return profile performs using out of sample data relatively close to the time frame of the analysis windows.

Let us denote by $\{t_{n+1}, \cdots, t_m\}$ the out of sample time-stamps we want to use to gauge the portfolio performance. Let us denote by β_i, with $i \in \{1, \cdots, p\}$ where p represents the number of drivers infer from our methodology. If we denote by $\tilde{y}$ the vector containing the prediction values of our replicating portfolio, such that $\tilde{y}_j = \frac{P_j^{\tilde{y}}}{P_{j-1}^{\tilde{y}}} - 1$, where $P_j^{\tilde{y}} = \sum_{i=1}^{p} \beta_i \cdot P_i^j$, $j \in \{t_{n+1}, \cdots, t_m\}$ and P_i^j is the closing price of the i^{th} key driver at time j. We can see from the pictures 3.11 and 3.12 that the replicating portfolios performed well.

Case 3: The target index remains the $S\&P500$. We would like to replicate it within the framework defined in (3.26). We applied the same standard deviation μ to all constituents in the initial portfolio, hence defining the variable Σ. Like in prior cases, we have considered two periods for this analysis : the year 2008, at the heart of the financial crisis, characterized by a high volatility and the year 2014, which is more of a low volatility period. For each period, we used the following values for $\mu \in \{0.001; 0.003; 0.005; 0.007; 0.01\}$ which correspond to cases where the matrix of return, that we observed, is more and more polluted.

Our results (see Tables 3.7 and 3.8) show us that the greater the value of μ, which implies a heavily polluted matrix of return, the more difficult it is for the methodology to construct a well-behaved replicating portfolio. However, we can still note that even

GICS Sector Name	$S\&P500$	Rep. Port. - 2008	Rep. Port. - 2014
Industrials	13.09 %	12.658%	14.49%
Health Care	12.3 %	13.92%	13.7%
Information Technology	13.88%	16.455%	14.49%
Consumer Discretionary	16.66%	12.65%	13.76%
Utilities	5.55%	5.06%	4.34%
Financials	13.49%	15.18%	14.49%
Materials	4.96%	6.32%	6.52%
Real Estate	6.54%	5.06%	2.17%
Consumer Staples	6.74%	3.79%	8.69%
Energy	6.34%	7.59%	5.79%
Telecommunication Services	0.59%	1.26%	1.44%
Number of Securities in Portfolio	500	80	139

Table 3.6 – percentage of each sector present in the $S\&P500$ and the replicating portfolios for the period 2008 and 2014.

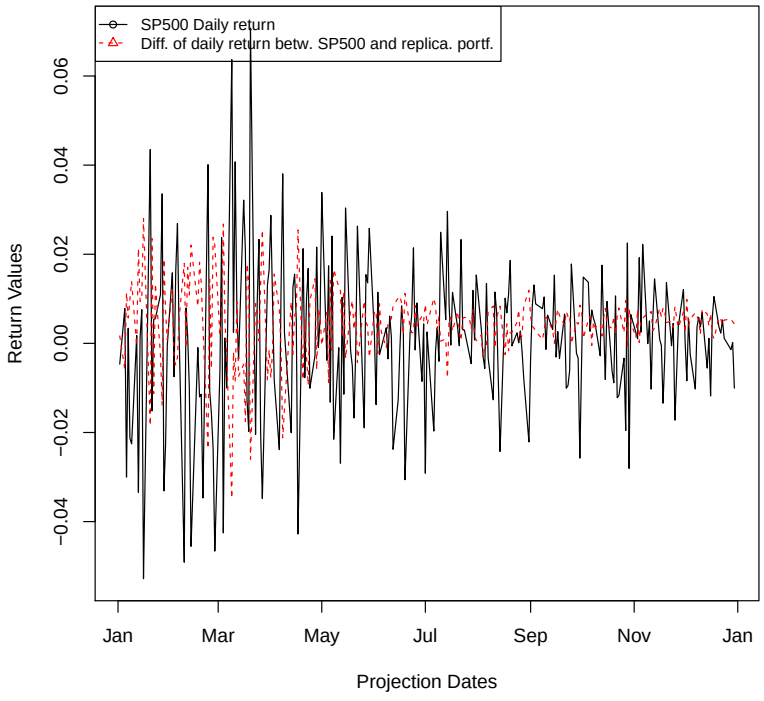


Figure 3.11 – Difference in daily return between $S\&P500$ and replication portfolio - projection dates are in 2009 - sample data covers the period 2008

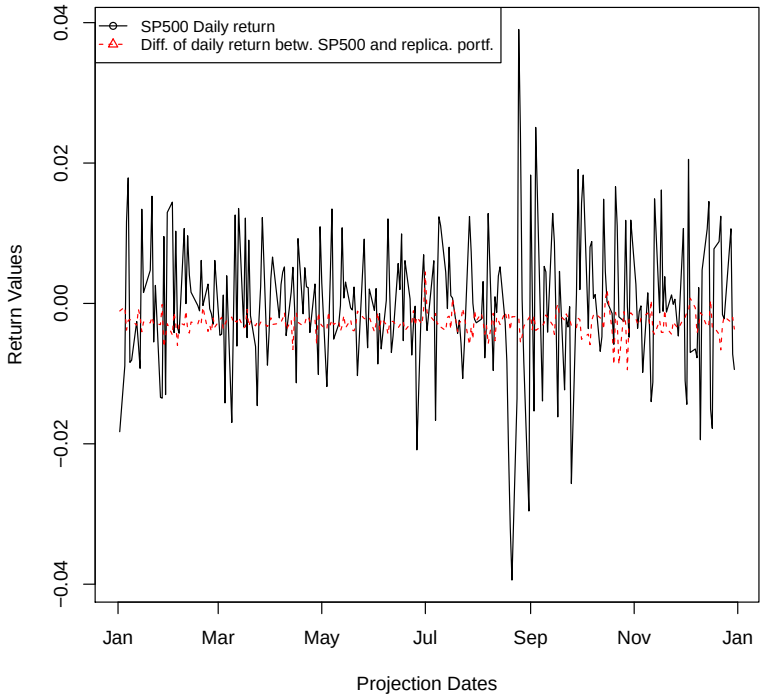


Figure 3.12 – Difference in daily return between $S\&P500$ and replication portfolio - projection dates are in 2015 - sample data covers the period 2014

μ	0.001	0.003	0.005	0.007	0.01
Industrials	5%	14.49%	10.6%	9.52%	5.55%
Health Care	13.75%	7.24%	10.6%	4.76%	8.33%
Information Technology	21.25%	13.04%	10.6%	11.9%	8.33%
Consumer Discretionary	20%	18.8%	22.72%	9.52%	22.22%
Utilities	6.25%	2.89%	3.03%	2.38%	5.55%
Financials	12.5%	23.18%	25.75%	30.95%	25%
Materials	2.5%	4.34%	0%	0%	2.77%
Real Estate	6.25%	4.34%	4.54%	4.76%	5.55%
Consumer Staples	2.5%	5.79%	3.03%	0%	0%
Energy	8.75%	5.79%	7.57%	23.81%	16.66%
Telecommunication Services	1.25%	0%	1.51%	2.38%	0%
Number of Securities in Portfolio	81	70	67	43	37

Table 3.7 – representation of each sector in the replicating portfolio with respect to the starting list, for each value of μ, on the period 2008

μ	0.001	0.003	0.005	0.007	0.01
Industrials	15.85%	12.5%	15.68%	21.73%	20.58%
Health Care	13.41%	17.85%	15.68%	13.04%	8.82%
Information Technology	13.41%	16.07%	25.49%	15.21%	11.76%
Consumer Discretionary	10.97%	5.35%	7.84%	10.86%	11.76%
Utilities	2.43%	3.57%	1.96%	2.17%	2.94%
Financials	20.73%	19.64%	23.52%	23.91%	23.52%
Materials	4.87%	8.92%	3.92%	6.52%	8.82%
Real Estate	2.43%	1.78%	0%	0%	0%
Consumer Staples	7.31%	7.14%	0%	0%	2.94%
Energy	7.31%	7.14%	5.88%	2.17%	8.82%
Telecommunication Services	1.22%	0%	0%	4.34%	0%
Number of Securities in Portfolio	83	57	52	47	35

Table 3.8 – representation of each sector in the replicating portfolio with respect to the starting list, for each value of μ, on the period 2014

with large amount of noise in the data, the methodology constructs a portfolio whose constituents are adequately distributed among *GICS sector*. Hence, even through the noise, the methodology is able to recognize the role played by each sector in defining the overall behavior of the target index.

3.6 Appendix

Proof of Proposition 3.1: Given (3.10), we get

$$\frac{1}{2n}\|Y - \mathbb{X}\hat{\beta}\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{2n}\|Y - \mathbb{X}\beta\|_2^2 + \lambda\|\beta\|_1$$

for all $\beta \in \mathbb{R}^p$. This can be rewritten as

$$\frac{1}{2n}\|\mathbb{X}(\hat{\beta} - \beta)\|_2^2 + \lambda\|\hat{\beta}\|_1 \leq \frac{1}{n}\epsilon^\top X(\hat{\beta} - \beta) + \lambda\|\beta\|_1.$$

We write, through Hölder inequality,

$$\frac{1}{n}\epsilon^\top X(\hat{\beta} - \beta) \leq \frac{1}{n}\|\epsilon^\top X\|_\infty \|\hat{\beta} - \beta\|_1$$

and consider the event $\mathcal{A} = \{\frac{1}{n}\|\epsilon^\top X\|_\infty \leq \lambda_0\}$. On $\mathcal{A}$, we get

$$\frac{1}{2n}\|\mathbb{X}(\hat{\beta} - \beta)\|_2^2 \leq \lambda_0\|\hat{\beta} - \beta\|_1 + \lambda\|\beta\|_1 - \lambda\|\hat{\beta}\|_1.$$

We denote by $S = \{j : \beta_j \neq 0\}$ and by $N = \{j : \beta_j = 0\}$ and we split the vector $\hat{\beta} - \beta$ and $\hat{\beta}$ according to the sets S and N. Note also that $\|\beta\|_1 = \|\beta_S\|_1$, thus

$$\frac{1}{2n}\|\mathbb{X}(\hat{\beta} - \beta)\|_2^2 \leq \lambda_0\|(\hat{\beta} - \beta)_S\|_1 + \lambda_0\|(\hat{\beta} - \beta)_N\|_1 + \lambda\|\beta_S\|_1 - \lambda\|\hat{\beta}_S\|_1 - \lambda\|\hat{\beta}_N\|_1.$$

We remark that $\|\hat{\beta}_S - \beta_S\|_1 \geq \|\beta_S\|_1 - \|\hat{\beta}_S\|_1$, which implies that $-\lambda\|\hat{\beta}_S\|_1 \leq -\lambda\|\beta_S\|_1 + \lambda\|(\hat{\beta} - \beta)_S\|_1$. Hence,

$$\begin{aligned}\frac{1}{2n}\|\mathbb{X}(\hat{\beta} - \beta)\|_2^2 &\leq (\lambda_0 + \lambda)\|(\hat{\beta} - \beta)\|_1 + \lambda_0\|(\hat{\beta} - \beta)_N\|_1 - \lambda\|\hat{\beta}_N\|_1 \\ &\leq (\lambda_0 + \lambda)\|(\hat{\beta} - \beta)_S\|_1 + (\lambda_0 - \lambda)\|\hat{\beta}_N\|_1 \\ &\leq \frac{3\lambda}{2}\|(\hat{\beta} - \beta)_S\|_1 - \frac{\lambda}{2}\|\hat{\beta}_N\|_1 \text{ for } \lambda_0 \leq \frac{\lambda}{2}.\end{aligned}$$

In particular, we obtained here that $\|(\hat{\beta} - \beta)_N\|_1 = \|\hat{\beta}_N\|_1 \leq 3\|(\hat{\beta} - \beta)\|_1$. We used here the compatibility condition (3.11) and apply it to $\hat{\beta} - \beta$ and get $\|(\hat{\beta} - \beta)_S\|_1 \leq \frac{\sqrt{s}}{\phi_0\sqrt{n}}\|\mathbb{X}(\hat{\beta} - \beta)\|_2$. We then have

$$\begin{aligned}\frac{1}{2n}\|\mathbb{X}(\hat{\beta} - \beta)\|_2^2 + \frac{\lambda}{2}\|\hat{\beta} - \beta\|_1 &\leq \frac{3\lambda}{2}\|(\hat{\beta} - \beta)_S\|_1 - \frac{\lambda}{2}\|\hat{\beta}_N\|_1 + \frac{\lambda}{2}\|(\hat{\beta} - \beta)_S\|_1 + \frac{\lambda}{2}\|\hat{\beta}_N - \beta_N\|_1 \\ &\leq 2\lambda\|(\hat{\beta} - \beta)_S\|_1 \leq 2\lambda\frac{\sqrt{s}}{\phi_0\sqrt{n}}\|\mathbb{X}(\hat{\beta} - \beta)\|_2.\end{aligned}$$

We use the inequality $2ab \leq \frac{a^2}{4} + 4b^2$ to get

$$\frac{1}{2n}\|\mathbb{X}(\hat{\beta} - \beta)\|_2^2 + \frac{\lambda}{2}\|\hat{\beta} - \beta\|_1 \leq \frac{1}{4n}\|\mathbb{X}(\hat{\beta} - \beta)\|_2^2 + 4\lambda^2\frac{s}{\phi_0^2}$$

and this gives

$$\frac{1}{2n}\|\mathbb{X}(\hat{\beta} - \beta)\|_2^2 + \lambda\|\hat{\beta} - \beta\|_1 \leq 8\lambda^2\frac{s}{\phi_0^2}.$$

For $\lambda = 2\sigma\|X\|_\infty\sqrt{\frac{t^2+2\log(p)}{n}}$, we can then conclude that

$$\frac{1}{2n}\|\mathbb{X}(\hat{\beta} - \beta)\|_2^2 + \lambda\|\hat{\beta} - \beta\|_1 \leq 4\sigma^2\|X\|_\infty^2\frac{8s}{\phi_0^2}\frac{t^2 + 2\log(p)}{n}$$

and

$$\|\hat{\beta} - \beta\|_1 \leq 2\sigma\|X\|_\infty \cdot \frac{8s}{\phi_0^2} \cdot \sqrt{\frac{t^2 + 2\log(p)}{n}}.$$

We note that the choice of λ is such that $\lambda \geq 2\lambda_0$ and that $\mathcal{A}$ holds with high probability

$$P(\mathcal{A}^c) = P(\max_{1\leq j\leq p}\frac{1}{n}|\epsilon^\top\mathbb{X}_j| > \lambda_0) \leq 2\exp(-\frac{t^2}{2})$$

where $\lambda_0 = 2\sigma\|X\|_\infty\sqrt{\frac{t^2+2\log(p)}{n}}$ with $t > 0$.

The above inequality relies on what follows. It is used at several occasions in the demonstrations below.

We note here that if we assume $Z \sim \mathcal{N}(0,1)$. Hence, for all $t > 0$

$$P(Z > t) = \frac{1}{2\sqrt{\pi}} \int_t^\infty \exp(-\frac{x^2}{2}) dx$$
$$\leq \frac{1}{2\sqrt{\pi}} \int_t^\infty \frac{x}{t} \exp(-\frac{x^2}{2}) dx = \frac{1}{2t\sqrt{\pi}} \exp(-\frac{t^2}{2}).$$

□

Proof of Proposition 3.2:

We repeat here classical arguments for proving convergence rates for (3.14). By definition, for all $\theta \in \mathbb{R}^k$, we have

$$\frac{1}{2n}\|V \cdot (Y - \widetilde{\Psi}_k \hat{\theta})\|_2^2 + \lambda \|\hat{\theta}\|_1 \leq \frac{1}{2n}\|V \cdot (Y - \widetilde{\Psi}_k \theta)\|_2^2 + \lambda \|\theta\|_1,$$

which can be rearranged into

$$\frac{1}{2n}(\|V\widetilde{\Psi}_k \hat{\theta}\|_2^2 - \|V\widetilde{\Psi}_k \theta\|_2^2) \leq \frac{2}{2n}\langle VY, V\widetilde{\Psi}_k(\hat{\theta} - \theta)\rangle + \lambda\|\theta\|_1 - \lambda\|\hat{\theta}\|_1$$

giving also

$$\frac{1}{2n}\|V\widetilde{\Psi}_k(\hat{\theta} - \theta)\|_2^2 \leq \frac{2}{2n}\langle V(Y - \widetilde{\Psi}_k \theta), V\widetilde{\Psi}_k(\hat{\theta} - \theta)\rangle + \lambda\|\theta\|_1 - \lambda\|\hat{\theta}\|_1$$

We know that $V(Y - \widetilde{\Psi}_k \theta) = V(\Psi_m \eta + \epsilon) = V\epsilon$. Indeed, $V\Psi_m = 0$. Therefore,

$$\frac{1}{2n}\|V\widetilde{\Psi}_k(\hat{\theta} - \theta)\|_2^2 \leq \frac{2}{2n}\epsilon^\top V\widetilde{\Psi}_k(\hat{\theta} - \theta) + \lambda\|\theta\|_1 - \lambda\|\hat{\theta}\|_1$$
$$\leq \frac{1}{n}\|\epsilon^\top V\widetilde{\Psi}_k\|_\infty \|\hat{\theta} - \theta\|_1 + \lambda\|\theta\|_1 - \lambda\|\hat{\theta}\|_1 \text{ through Hölder inequality.}$$

Let us assume that $\frac{1}{n}\|\epsilon^\top V\widetilde{\Psi}_k\|_\infty \leq \lambda_0$ on some event $\mathcal{A}$. However, we split the coordinates of θ in $S_\theta = \{j : \theta_j \neq 0\}$ and $N_\theta = \{j : \theta_j = 0\}$. Thus,

$$\frac{1}{2n}\|V\widetilde{\Psi}_k(\hat{\theta} - \theta)\|_2^2 \leq \lambda_0\|\hat{\theta}_S - \theta_S\|_1 + \lambda_0\|\hat{\theta}_N\|_1 + \lambda\|\theta_S\|_1 - \lambda\|\hat{\theta}_S\|_1 - \lambda\|\hat{\theta}_N\|_1$$

Using $\|\hat{\theta}_S - \theta_S\|_1 \geq \|\theta_S\|_1 - \|\hat{\theta}_S\|_1$, we then get

$$\frac{1}{2n}\|V\widetilde{\Psi}_k(\hat{\theta} - \theta)\|_2^2 \leq (\lambda_0 + \lambda)\|\hat{\theta}_S - \theta_S\|_1 + (\lambda_0 - \lambda)\|\hat{\theta}_N\|_1$$
$$\frac{3\lambda}{2}\|\hat{\theta}_S - \theta_S\|_1 - \frac{\lambda}{2}\|\hat{\theta}_N\|_1$$

for $\lambda_0 \leq \frac{\lambda}{2}$. Given the compatibility condition (3.11), we know that there exists $\phi_0 > 0$ and s_θ the sparsity of θ such that for all θ in the cone $\|\theta_N\|_1 \leq 3\|\theta_S\|_1$, we have

$$\|\theta_S\|_1^2 \leq \frac{s_\theta}{\phi_0^2}\frac{\|V\widetilde{\Psi}_k \theta\|_2^2}{n}.$$

We can then say that

$$\frac{1}{2n}\|V\widetilde{\Psi}_k(\hat{\theta}-\theta)\|_2^2+\frac{\lambda}{2}\|\hat{\theta}-\theta\|_1\leq 2\lambda\|\hat{\theta}_S-\theta_S\|_1\leq 2\lambda\frac{\sqrt{s_\theta}}{\phi_0\sqrt{n}}\|V\widetilde{\Psi}_k(\hat{\theta}-\theta)\|_2.$$

We use the inequality $2ab\leq\frac{a^2}{4}+4b^2$ to infer that

$$\frac{1}{2n}\|V\widetilde{\Psi}_k(\hat{\theta}-\theta)\|_2^2+\frac{\lambda}{2}\|\hat{\theta}-\theta\|_1\leq\frac{1}{4n}\|V\widetilde{\Psi}_k(\hat{\theta}-\theta)\|_2^2+4\lambda^2\frac{s_\theta}{\phi_0^2}$$

thus,

$$\frac{1}{2n}\|V\widetilde{\Psi}_k(\hat{\theta}-\theta)\|_2^2+\lambda\|\hat{\theta}-\theta\|_1\leq\lambda^2\frac{8s_\theta}{\phi_0^2}.$$

For $\lambda=2\sigma\|X\|_\infty\sqrt{\frac{t^2+2\log(k)}{n}}$, we can then conclude that

$$\frac{1}{2n}\|V\widetilde{\Psi}_k(\hat{\theta}-\theta)\|_2^2+\lambda\|\hat{\theta}-\theta\|_1\leq 4\sigma^2\|X\|_\infty^2\frac{8s_\theta}{\phi_0^2}\cdot\frac{t^2+2\log(k)}{n}.$$

Furthermore,

$$\|\hat{\theta}-\theta\|_1\leq 2\sigma\|X\|_\infty\frac{8s_\theta}{\phi_0^2}\cdot\sqrt{\frac{t^2+2\log(k)}{n}}.$$

This choice of $\lambda\geq 2\lambda_0$ is such that $P(\frac{1}{n}\|\epsilon^\top V\widetilde{\Psi}_k\|_\infty>\lambda_0)\leq 2\exp(-\frac{t^2}{2})$ for some $t>0$. We note that

$$\|\epsilon^\top V\widetilde{\Psi}_k\|_\infty=|(\epsilon^\top V\widetilde{\Psi}_k)_j|=\max_{j=1,\dots,k}|\sum_{i=1}^n\epsilon_i[V\widetilde{\Psi}_k]_{ij}|.$$

we have $\sum_{i=1}^n\epsilon_i[V\widetilde{\Psi}_k]_{ij}\sim\mathcal{N}(0,\sigma^2v_j^2)$ and

$$\begin{aligned}
v_j^2&=\sum_{i=1}^n[V\widetilde{\Psi}_k]_{ij}^2=\sum_{i=1}^nX_i^2[V\Psi_k]_{ij}^2\\
&\leq\max_i|X_i|^2(V\Psi_k\Psi_k^\top V^\top)_{ij}\\
&\leq\|X\|_\infty^2(\Psi_k^\top V^2\Psi_k)_{jj}\\
&\leq\|X\|_\infty^2(\Psi_k^\top V\Psi_k)_{jj}\quad\text{since } V^2=V\text{ , and } 0\leq V\leq I\text{ (the identity matrix)}\\
&\leq\|X\|_\infty^2(\Psi_k^\top\Psi_k)_{jj}=n\|X\|_\infty^2.
\end{aligned}$$

Hence, we have $\|\epsilon^\top V\widetilde{\Psi}_k\|_\infty\leq\sigma\|X\|_\infty\sqrt{2n(1+t)\log(k)}$ with probability larger than $1-\exp(-\frac{t^2}{2})$. □

Proof of Theorem 3.3:

We denote by $\widetilde{\mathcal{L}}=\frac{1}{2}\beta^\top\Gamma\beta-\gamma^T\beta+\lambda\|\beta\|_1$ the loss function to be minimized. We denote by β the true value of our variable of interest. Hence, we know that β is feasible but the estimator $\hat{\beta}$ is optimal for (3.19). Hence, $\widetilde{\mathcal{L}}(\hat{\beta})\leq\widetilde{\mathcal{L}}(\beta)$. Thus,

$$\hat{\beta}^\top\Gamma\hat{\beta}-\beta^\top\Gamma\beta\leq 2\langle\gamma,\hat{\beta}-\beta\rangle+2\lambda(\|\beta\|_1-\|\hat{\beta}\|_1).\tag{3.29}$$

Let us denote define the variable $\Delta = \hat{\beta} - \beta$. We have

$$\begin{aligned}\Delta^\top \Gamma \Delta &= \Delta^\top \Gamma \hat{\beta} - \Delta^\top \Gamma \beta \\ &= \hat{\beta}^\top \Gamma \hat{\beta} - \beta^\top \Gamma \hat{\beta} + \hat{\beta}^\top \Gamma \beta - \beta^\top \Gamma \beta - 2\Delta^\top \Gamma \beta.\end{aligned}$$

Then, $\Delta^\top \Gamma \Delta + 2\Delta^\top \Gamma \beta = \hat{\beta}^\top \Gamma \hat{\beta} - \beta^\top \Gamma \beta.$

Thus (3.29) can be rearranged into,

$$\Delta^\top \Gamma \Delta \leq 2\langle \Delta, \gamma - \Gamma\beta \rangle + 2\lambda(\|\beta\|_1 - \|\hat{\beta}\|_1). \tag{3.30}$$

Let us now turn our attention to the term $\langle \Delta, \gamma - \Gamma\beta \rangle$. Using Hölder inequality, it can be written,

$$\langle \Delta, \gamma - \Gamma\beta \rangle \leq \|\gamma - \Gamma\beta\|_\infty \|\Delta\|_1.$$

Furthermore, we know that

$$\|\gamma - \Gamma\beta\|_\infty = \|\frac{1}{n}\left(\widetilde{\mathbb{X}}^\top Y - \widetilde{\mathbb{X}}^\top \widetilde{\mathbb{X}}\beta + n\mu^2 \zeta\beta\right)\|_\infty, \quad \text{since } \zeta^2 = \zeta^\top \zeta = \zeta. \tag{3.31}$$

We note that,

$$\begin{aligned}\widetilde{\mathbb{X}}^\top \widetilde{\mathbb{X}} &= \widetilde{\mathbb{X}}^\top \mathbb{X} + \widetilde{\mathbb{X}}^\top \mathbb{K} \\ &= \mathbb{X}^\top \mathbb{X} + \mathbb{X}^\top \mathbb{K} + (\mathbb{X}^\top \mathbb{K})^\top + \mathbb{K}^\top \mathbb{K},\end{aligned}$$

moreover,

$$\mathbb{U}_1 = (\mathbb{X}^\top \mathbb{K})^\top = [0_{n\times m}, D_\nu \Psi_k]^\top [\Psi_m, D_X \Psi_k] = \begin{bmatrix} 0_{m\times m} & 0_{m\times k} \\ (D_\nu \Psi_k)^\top \Psi_m & (D_\nu \Psi_k)^\top D_X \Psi_k \end{bmatrix}$$

and

$$\mathbb{U}_2 = \mathbb{K}^\top \mathbb{K} = [0_{n\times m}, D_\nu \Psi_k]^\top [0_{n\times m}, D_\nu \Psi_k] = \begin{bmatrix} 0_{m\times m} & 0_{m\times k} \\ 0_{k\times k} & (D_\nu \Psi_k)^\top D_\nu \Psi_k \end{bmatrix}.$$

Thus,

$$\widetilde{\mathbb{X}}^\top \widetilde{\mathbb{X}} = \mathbb{X}^\top \mathbb{X} + \mathbb{U}_1 + \mathbb{U}_1^\top + \mathbb{U}_2.$$

Knowing that $Y = \mathbb{X}\beta + \epsilon$ and the previous expression of $\widetilde{\mathbb{X}}^\top \widetilde{\mathbb{X}}$, we can then write (3.31)

$$\begin{aligned}\|\gamma - \Gamma\beta\|_\infty &= \frac{1}{n}\|\mathbb{X}^\top \epsilon + \mathbb{K}^\top \epsilon - \mathbb{U}_1^\top \beta - \mathbb{U}_2^\top \beta + n\mu^2 \zeta\beta\|_\infty \\ &\leq \frac{1}{n}\|(\mathbb{X}^\top + \mathbb{K}^\top)\epsilon\|_\infty + \frac{1}{n}\|\mathbb{U}_1^\top \beta\|_\infty + \frac{1}{n}\|(\Psi_k^\top D_{\nu^2} \Psi_k - \Psi_k^\top D_{\mu^2} \Psi_k)\theta\|_\infty \\ &\leq \frac{1}{n}\|\mathbb{X}^\top \epsilon\|_\infty + \frac{1}{n}\|\mathbb{K}^\top \epsilon\|_\infty + \frac{1}{n}\|\mathbb{U}_1^\top \beta\|_\infty + \frac{1}{n}\|\Psi_k^\top D_{\nu^2 - \mu^2} \Psi_k \theta\|_\infty.\end{aligned}$$

In order to find an upper bound of $\|\gamma - \Gamma\beta\|_\infty$, we look individually at each of the four terms of the inequality above and find their corresponding upper bound.

(i) We know that $\|\mathbb{X}^\top\epsilon\|_\infty = \max_{j=1,\dots,p}\sum_{i=1}^n \epsilon_i X_i \Phi_j(W_i) \sim \mathcal{N}(0,\sigma^2 d_j^2)$. The variable d_j can be bounded as follows : $d_j^2 = \sum_{i=1}^n X_i\Phi_j^2 \leq n\|X\|_\infty^2$. Hence, we can conclude that, $\frac{1}{n}\|\mathbb{X}^\top\epsilon\|_\infty \leq \sigma\|X\|_\infty\sqrt{\frac{2(1+t)\log(p)}{n}}$ with probability greater than $1-\exp(-\frac{t^2}{2})$.

(ii) We have

$$\begin{aligned}\|\mathbb{K}^\top\epsilon\|_\infty &= \max_{j=1,\dots,k}\sum_{i=1}^n \mathbb{K}_{m+j}\epsilon_i \\ &\leq \max_{j=1,\dots,k}\sum_{i=1}^n |\epsilon_i\nu_i\Phi_j(W_i)| \\ &\leq \max_{j=1,\dots,k}\frac{1}{4}\sum_{i=1}^n |\widetilde{\epsilon}_i^2\Phi_j(W_i)| + \max_{j=1,\dots,k}\frac{1}{4}\sum_{i=1}^n |\widetilde{\nu}_i^2\Phi_j(W_i)| \end{aligned} \tag{3.32}$$

with $\widetilde{\nu}_i = \epsilon_i - \nu_i \sim \mathcal{N}(0,\sigma^2+\mu^2)$, $\widetilde{\epsilon}_i = \epsilon_i + \nu_i \sim \mathcal{N}(0,\sigma^2+\mu^2)$ and using the equality $xy = \frac{1}{4}\{(x+y)^2-(x-y)^2\}$. Let us introduce the variable $N_j = \frac{1}{4n}\sum_{i=1}^n |\widetilde{\epsilon}_i^2\Phi_j(W_i)| = \frac{1}{4n}\sum_{i=1}^n \widetilde{\epsilon}_i^2 B_{ij}$ with $B_{ij} = |\Phi_j(W_i)|$. Let us found from above the Laplace transform of N_j:

$$\begin{aligned} E\left[\exp\left(\frac{\lambda}{4n}\sum_{i=1}^n \widetilde{\epsilon}_i^2 B_{ij}\right)\right] &= \prod_{i=1}^n E\left[\exp\left(\frac{\lambda}{4n}\widetilde{\epsilon}_i^2 B_{ij}\right)\right] \\ &= \prod_{i=1}^n E\left[\exp\left(\frac{\lambda\alpha^2}{4n}B_{ij}\cdot\frac{\widetilde{\epsilon}_i^2}{\alpha^2}\right)\right] \quad \text{with } \alpha^2 = \sigma^2+\mu^2 \\ &\leq \prod_{i=1}^n \exp(2(\frac{\lambda\alpha^2}{4n}B_{ij})^2) \\ &\leq \exp\left(\frac{1}{2}\frac{\lambda^2\alpha^4}{4n^2}\sum_{i=1}^n B_{ij}^2\right) \quad \text{with } \lambda \textit{ subject to } |\lambda| \leq \frac{n}{\alpha^2 B_{ij}}\forall i = 1,\dots,n \\ &\leq \exp\left(\frac{1}{2}\frac{\lambda^2\alpha^4}{4n}\right) \quad \text{with } |\lambda| \leq \frac{n}{\alpha^2\delta} \end{aligned}$$

with the constant $\delta > 0$ which satisfies $|\phi_j(W_i)| \leq \delta$ for all i and j. Note that such a constant exists by construction. We know that if Z is a sub-exponential variable with non-negative parameters (κ, b), then

$$\mathbb{P}\left[|Z - E[Z]| \geq \max\{\kappa\sqrt{u}, bu\}\right] \leq \exp\left(-\frac{u}{2}\right), \quad \text{for all } u > 0. \tag{3.33}$$

Given the expression of the Laplace transform, we obtained that N_j is a sub-exponential random variable with parameter $\kappa = \frac{\alpha^2}{2\sqrt{n}}$ and $b = \frac{\alpha^2\delta}{n}$. Hence, for all j,

$$N_j \leq \max\left\{\frac{\alpha^2}{2}\sqrt{\frac{t}{n}}, \alpha^2\delta\frac{t}{n}\right\},$$

for $t > 0$, with probability greater than $1-\exp(-\frac{t}{2})$. The same analysis provides the same bounds for the second term in (3.32). We then have, $\frac{1}{n}\|\mathbb{K}^\top\epsilon\|_\infty \leq (\sigma^2+\mu^2)\max\left\{\sqrt{\frac{t}{n}}, 2\delta\frac{t}{n}\right\}$ with probability larger than $1-2\exp(-\frac{t}{2})$.

(iii) Let us denote by $T_j = \sum_{i=1}^n \nu_i \Phi_j(W_i)[\Psi_k\theta]_i$ and $S_j = \sum_{i=1}^n \nu_i X_i \Phi_j(W_i)[\Psi_k\theta]_i$, such that $\|\mathbb{U}_1^\top\beta\|_\infty = \max\{\max_{j=1,\dots,p}|T_j|, \max_{j=1,\dots,p}|S_j|\}$. We know that for all j, $T_j \sim \mathcal{N}(0, \mu^2 v_j^2)$. The variance v_j^2 can be bounded as follows :

$$\begin{aligned} v_j^2 &= \sum_{i=1}^n (\Phi_j(W_i)^2[\Psi_k\theta]_i)^2 \\ &= \Phi_j^\top \Psi_k \theta\theta^\top \Psi_k^\top \Phi_j \\ &\leq Tr(\Psi_k\theta\theta^\top\Psi_k^\top) Tr(\Phi_j\Phi_j^\top) \\ &\leq nL \quad \text{since } \sum_{l=1}^k (\Phi_l(W_i)\theta_l)^2 \leq L \quad \text{by hypothesis.} \end{aligned}$$

Similarly, we know that for all j, $S_j \sim \mathcal{N}(0, \mu^2 v_j^2)$. As before, we show that

$$v_j^2 \leq \|X\|_\infty^2 \max_i [\Phi_j(W_i)(\Psi_k\theta)_i]^2 = \|X\|_\infty^2 nL.$$

As such, we can write that (iii), $\frac{1}{n}\|\mathbb{U}_1^\top\beta\|_\infty \leq \mu(1+\|X\|_\infty)\sqrt{\frac{2(1+t)L\log(p)}{n}}$ with probability greater than $1-\exp(-\frac{t^2}{2})$.

(iv) Let us denote by M_j the j^{th} element of the vector $\frac{1}{n}[\Psi_k^\top D_{\nu^2-\mu^2}\Psi_k\theta]$ such that $M_j = \frac{1}{n}[\Phi_j D_{\nu^2-\mu^2}(\Psi_k\theta)]_j = \frac{1}{n}\sum_{i=1}^n(\nu_i^2-\mu^2)V_{ij}$ where $V_{ij} = \left(\sum_{l=1}^k \Phi_l(W_i)\theta_l\right)\Phi_j(W_i)$. Applying the Laplace transform to the variable M_j, we obtain

$$\begin{aligned} E[\exp(\frac{\lambda}{n}\sum_{i=1}^n(\nu_i^2-\mu^2)V_{ij})] &= \prod_{i=1}^n E[\exp(\frac{\lambda}{n}(\nu_i^2-\mu^2)V_{ij})] \\ &= \prod_{i=1}^n E[\exp(\frac{\lambda\mu^2}{n}(\frac{\nu_i^2}{\mu^2}-1)V_{ij})] \quad \text{with } \frac{\nu_i^2}{\mu^2} \sim \chi_1^2 \\ &= \prod_{i=1}^n \frac{\exp(\frac{\lambda\mu^2}{n}V_{ij})}{\sqrt{1-2V_{ij}\frac{\lambda\mu^2}{n}}} \\ &\leq \prod_{i=1}^n \exp(\frac{1}{2}\frac{4\lambda^2\mu^4}{n^2}V_{ij}^2) \\ &\qquad \text{with } \frac{|\lambda|\mu^2}{n}V_{ij} < \frac{1}{4}, \quad \forall i = 1,\dots,n \end{aligned}$$

since, $\max_i |V_{ij}| \leq \max_i |\phi_j(W_i)| \sum_{l=1}^k |\phi_l(W_i)\theta_l| \leq \delta\|\Psi_k(W)\theta\|_2^2 \leq \delta L,$

and $\frac{1}{n}\sum_{i=1}^n V_{ij}^2 \leq L$ by hypothesis .

We have,

$$E[\exp(\frac{\lambda}{n}\sum_{i=1}^n(\nu_i^2-\mu^2)V_{ij})] \leq \exp(\frac{1}{2}\frac{4\lambda^2\mu^4}{n}L) \quad \text{for } |\lambda| \leq \frac{n}{2\mu^2\delta L}$$

Given the expression of the Laplace Transform, we know that we are dealing with a sub-exponential random variable with parameter $\kappa = \frac{2\mu^2\sqrt{L}}{\sqrt{n}}$ and $b = \frac{4\mu^2\delta L}{n}$. Hence, from (3.33), we can conclude that, for all j, $|M_j| \leq \max\{\frac{2\mu^2\sqrt{L}}{\sqrt{n}}\sqrt{t}, \frac{4\mu^2 L}{n}t\}$ with probability greater than $1-\exp(-\frac{t}{2})$.

From the results $(i), (ii), (iii)$ and (iv), we can then conclude that, with probability greater than $1 - 5\exp(-\frac{t}{2})$, with t large enough (we note here that $P(A \cap B) \geq P(A) + P(B) - 1$). Hence, there exists $\lambda_0 \in \mathbb{R}$ such that $\|\gamma - \Gamma\beta\|_\infty \leq \lambda_0$, with

$$\lambda_0 = \left(\sigma\|X\|_\infty + \mu\sqrt{L}(1 + \|X\|_\infty)\right)\sqrt{\frac{2(1+t)\log(p)}{n}} + \max\left\{\frac{2\mu^2\sqrt{L}}{\sqrt{n}}\sqrt{t}, \frac{4\mu^2\delta L}{n}t\right\} + (\sigma^2 + \mu^2) \cdot \max\left\{\sqrt{\frac{t}{n}}, 2\delta\frac{t}{n}\right\}.$$

Thus (3.30) can be rewritten as,

$$\Delta^\top\Gamma\Delta \leq 2\lambda_0\|\Delta\|_1 + 2\lambda(\|\beta\|_1 - \|\hat{\beta}\|_1).$$

We know that $S = \{j : \beta_j \neq 0\}$ and by $N = \{j : \beta_j = 0\}$. Hence,

$$\begin{aligned}\lambda(\|\beta\| - \|\hat{\beta}\|_1) &\leq \lambda\|\beta_S\| - \lambda\|\hat{\beta}_S\|_1 - \lambda\|\hat{\beta}_N\|_1 \quad \text{since } \|\beta_N\| = 0 \\ &\leq \lambda\|\hat{\beta}_S - \beta_S\|_1 - \lambda\|\hat{\beta}_N\|_1\end{aligned}$$

therefore,

$$\begin{aligned}\Delta^\top\Gamma\Delta &\leq 2(\lambda_0 + \lambda)\|\hat{\beta}_S - \beta_S\|_1 + 2(\lambda_0 - \lambda)\|\hat{\beta}_N - \beta_N\|_1 \\ &\leq 3\lambda\|\hat{\beta}_S - \beta_S\|_1 - \lambda\|\hat{\beta}_N - \beta_N\|_1, \quad \text{assuming } \lambda_0 \leq \frac{\lambda}{2} \\ &\leq 3\lambda\|\hat{\beta}_S - \beta_S\|_1\end{aligned}$$

From the above, we can infer that Δ verifies the cone property with $c_0 = 3$. Given the Restricted Eigenvalues condition (3.20), we can write that $\Delta^\top\Gamma\Delta \geq t_0\|\Delta_S\|_2^2$. Since, $\|\Delta_S\|_1^2 \leq s_\beta\|\Delta_S\|_2^2$, then we have $\Delta^\top\Gamma\Delta \geq \frac{t_0}{s_\beta}\|\Delta_S\|_1^2$. Then, from the above, we have:

$$\frac{t_0}{s_\beta}\|\Delta_S\|_1^2 \leq 3\lambda\|\Delta_S\|_1, \qquad \text{then} \qquad \|\Delta_S\|_1 \leq 3\lambda\frac{s_\beta}{t_0}.$$

Furthermore, we know that $\|\Delta_N\|_1 \leq c_0\|\Delta_S\|_1$, thus $\|\Delta\|_1 \leq (c_0 + 1)\|\Delta_S\|_1$. We can then conclude that

$$\|\hat{\beta} - \beta\|_1 \leq \lambda \cdot s_\beta \cdot \frac{12}{t_0}$$

with probability larger than $1 - 5\exp(-\frac{t}{2})$, and t large enough.

□

Proof of Theorem 3.4:

Let us denote by θ the true value of our parameter of interest. We know that θ is feasible and $\hat{\theta}$ is the optimal parameter for (3.22). Hence, $\mathcal{L}(\hat{\theta}) \leq \mathcal{L}(\theta)$. Thus,

$$\hat{\theta}^\top\Gamma\hat{\theta} - \theta^\top\Gamma\theta \leq 2\langle\gamma, \hat{\theta} - \theta\rangle + 2\lambda(\|\theta\| - \|\hat{\theta}\|_1).$$

It can be rearranged into,

$$\Delta^\top \Gamma \Delta \leq 2\langle \Delta, \gamma - \Gamma\theta \rangle + 2\lambda(\|\theta\| - \|\hat{\theta}\|_1), \quad \text{similarly to (3.30) and with } \Delta = \hat{\theta} - \theta. \tag{3.34}$$

Through Hölder inequality, we can write,

$$\langle \Delta, \gamma - \Gamma\beta \rangle \leq \|\gamma - \Gamma\beta\|_\infty \|\Delta\|_1.$$

Furthermore, knowing that $V\epsilon = VY - VD_X\Psi_k\theta$ we have,

$$\begin{aligned}
\|\gamma - \Gamma\theta\|_\infty &= \frac{1}{n}\|(VD_Z\Psi_k)^\top Y - (VD_Z\Psi_k)^\top (VD_Z\Psi_k)\theta + \mu^2 (V\Psi_k)^\top (V\Psi_k)\theta\|_\infty \\
&= \frac{1}{n}\|(D_Z\Psi_k)^\top V\epsilon - (D_Z\Psi_k)^\top VD_\nu\Psi_k\theta + \mu^2\Psi_k^\top V\Psi_k\|_\infty \\
&= \frac{1}{n}\|\Psi_k^\top D_Z V\epsilon - \Psi_k^\top VD_X D_\nu\Psi_k\theta + \Psi_k^\top D_{\nu^2-\mu^2}V\Psi_k\|_\infty \\
&\leq \frac{1}{n}\|\Psi_k^\top D_Z V\epsilon\|_\infty + \frac{1}{n}\|\Psi_k^\top VD_X D_\nu\Psi_k\theta\|_\infty + \frac{1}{n}\|\Psi_k^\top D_{\nu^2-\mu^2}V\Psi_k\|_\infty \\
&\leq \frac{1}{n}\|\Psi_k^\top D_X V\epsilon\|_\infty + \frac{1}{n}\|\Psi_k^\top D_\nu V\epsilon\|_\infty + \frac{1}{n}\|\Psi_k^\top VD_X D_\nu\Psi_k\theta\|_\infty + \frac{1}{n}\|\Psi_k^\top D_{\nu^2-\mu^2}V\Psi_k\|_\infty
\end{aligned}$$

In order to find an upper bound of $\|\gamma - \Gamma\theta\|_\infty$, we look individually at each of the four terms of the inequality above and find their corresponding upper bound.

(i) We know that $\|\Psi_k^\top D_X V\epsilon\|_\infty \leq \max_{j=1,\dots,k}\sum_{i=1}^n \epsilon_i X_i \Phi_j(W_i) \sim \mathcal{N}(0, \sigma^2 d_j^2)$. The variable d_j can be bounded as follows : $d_j^2 = \sum_{i=1}^n X_i^2(\Phi_k^\top V)_{ij}^2 \leq n\|X\|_\infty^2$ since $V^2 = V$ and $0 \leq V \leq I$ (Identity matrix). Hence, we can conclude that $\frac{1}{n}|\Psi_k^\top D_X V\epsilon\|_\infty \leq \sigma\|X\|_\infty\sqrt{\frac{2(1+t)\log(k)}{n}}$ with probability greater than $1 - \exp(-\frac{t^2}{2})$.

(ii) We also have,

$$\begin{aligned}
\|\Psi_k^\top D_\nu V\epsilon\|_\infty &\leq \max_{j=1,\dots,k}|\sum_{i=1}^n \epsilon_i\nu_i\Phi_j(W_i)| \\
&\leq \max_{j=1,\dots,k}\sum_{i=1}^n |\epsilon_i\nu_i\Phi_j(W_i)| \\
&\leq \max_{j=1,\dots,k}\frac{1}{4}\sum_{i=1}^n |\widetilde{\epsilon}_i^2\Phi_j(W_i)| + \max_{j=1,\dots,k}\frac{1}{4}\sum_{i=1}^n |\widetilde{\nu}_i^2\Phi_j(W_i)|
\end{aligned} \tag{3.35}$$

with $\widetilde{\nu}_i = \epsilon_i - \nu_i \sim \mathcal{N}(0, \sigma^2 + \mu^2)$, $\widetilde{\epsilon}_i = \epsilon_i + \nu_i \sim \mathcal{N}(0, \sigma^2 + \mu^2)$ and using the equality $xy = \frac{1}{4}\{(x+y)^2 - (x-y)^2\}$. Let us introduce the variable $N_j = \frac{1}{4n}\sum_{i=1}^n |\widetilde{\epsilon}_i^2\Phi_j(W_i)| = \frac{1}{2n}\sum_{i=1}^n \widetilde{\epsilon}_i^2 B_{ij}$ with $B_{ij} = |\Phi_j(W_i)|$. Let us found from above the Laplace transform of N_j:

$$
\begin{aligned}
E\left[\exp\left(\frac{\lambda}{4n}\sum_{i=1}^{n}\widetilde{\epsilon}_i^2 B_{ij}\right)\right] &= \prod_{i=1}^{n} E\left[\exp\left(\frac{\lambda}{4n}\widetilde{\epsilon}_i^2 B_{ij}\right)\right] \\
&= \prod_{i=1}^{n} E\left[\exp\left(\frac{\lambda\alpha^2}{4n}B_{ij}\cdot\frac{\widetilde{\epsilon}_i^2}{\alpha^2}\right)\right] \quad \text{with } \alpha^2=\sigma^2+\mu^2 \\
&\leq \prod_{i=1}^{n}\exp(2(\frac{\lambda\alpha^2}{4n}))^2 \\
&\leq \exp\left(\frac{1}{2}\frac{\lambda^2\alpha^4}{4n^2}\sum_{i=1}^{n}B_{ij}^2\right) \quad \text{with } \lambda \text{ subject to } |\lambda|\leq\frac{n}{\alpha^2 B_{ij}}, \forall i=1,\ldots,n \\
&\leq \exp\left(\frac{1}{2}\frac{\lambda^2\alpha^4}{4n}\right) \quad \text{with } |\lambda|\leq\frac{n}{\alpha^2\delta}
\end{aligned}
$$

with the variable $\delta > 0$ an assumption of the theorem which satisfies $|\phi_j(W_i)| \leq \delta$ for all i and j. Note that such a constant exists by construction. Given the expression of the Laplace transform, we obtained N_j is a sub-exponential variable with parameter $\kappa = \frac{\alpha^2}{2\sqrt{n}}$ and $b = \frac{\alpha^2\delta}{n}$. Hence, given (3.33), for all j, $N_j \leq \max\left\{\frac{\alpha^2}{2}\sqrt{\frac{t}{n}}, \alpha^2\delta\frac{t}{n}\right\}$ for $t > 0$, with probability greater than $1-\exp(-\frac{t}{2})$. The same analysis provides the same bounds for the second term in (3.35). We then have $\|\Psi_k^\top D_\nu V\epsilon\|_\infty \leq (\sigma^2+\mu^2)\cdot\max\left\{\sqrt{\frac{t}{n}}, 2\delta\frac{t}{n}\right\}$ with probability greater than $1-2\exp(-\frac{t}{2})$.

(iii) We also know that $\|\Psi_k^\top V D_X D_\nu \Psi_k\theta\|_\infty = \max_{j=1,\ldots,k}\sum_{i=1}^{n}\nu_i X_i\Phi_j(W_i)[V\Psi_k\theta]_i = \max_{j=1,\ldots,k}|T_j|$. We can infer that $T_j \sim \mathcal{N}(0,\mu^2 v_j^2)$. The variable v_j can be bounded as follows

$$
\begin{aligned}
v_j^2 &= \sum_{i=1}^{n} X_i^2\Phi_j(W_i)^2[V\Psi_k\theta]_i^2 \\
&\leq \|X\|_\infty^2 \max_i[\Phi_j(W_i)(V\psi_k\theta)]^2 \\
&\leq \|X\|_\infty^2(\theta^\top\Psi_k^\top V^\top\Phi_j V\Psi_k\theta)_{jj} \\
&\leq \|X\|_\infty^2 nL \quad \text{where } (\sum_{l=1}^{k}\Phi_l(W_l)\theta_l)^2 \leq L \text{ by hypothesis and } 0\leq V\leq I(\text{Identity matrix}).
\end{aligned}
$$

Hence, we can conclude that, $\frac{1}{n}\|\Psi_k V D_X D_\nu\Psi_k\theta\|_\infty \leq \mu\sqrt{L}\|X\|_\infty\sqrt{\frac{2(1+t)\log(k)}{n}}$ with probability greater than $1-\exp(-\frac{t^2}{2})$.

(iv) Let us denote by M_j the j^{th} element of the vector $\frac{1}{n}[\Psi_k^\top D_{\nu^2-\mu^2}V\Psi_k\theta]$ such that $M_j = \frac{1}{n}[\Phi_j D_{\nu^2-\mu^2}(V\Psi_k\theta)]_j = \frac{1}{n}\sum_{i=1}^{n}(\nu_i^2-\mu^2)G_{ij}$ where $G_{ij} = (V\Psi_k\theta)_i\Phi_j(W_i)$. Applying

the Laplace transform to the equation above, we obtain

$$\begin{aligned}
E[\exp(\frac{\lambda}{n}\sum_{i=1}^{n}(\nu_i^2-\mu^2)G_{ij})] &= \prod_{i=1}^{n} E[\exp(\frac{\lambda}{n}(\nu_i^2-\mu^2)G_{ij})] \\
&= \prod_{i=1}^{n} E[\exp(\frac{\lambda\mu^2}{n}(\frac{\nu_i^2}{\mu^2}-1)G_{ij})] \quad \text{with } \frac{\nu_i^2}{\mu^2}\sim\chi_1^2 \\
&= \prod_{i=1}^{n} \frac{\exp(\frac{\lambda\mu^2}{n}G_{ij})}{\sqrt{1-2G_{ij}\frac{\lambda\mu^2}{n}}} \\
&\leq \prod_{i=1}^{n} \exp(\frac{1}{2}\frac{4\lambda^2\mu^4}{n^2}G_{ij}^2) \quad \text{with } \frac{\lambda\mu^2}{n}G_{ij}<\frac{1}{4}, \forall i=1,\dots,n
\end{aligned}$$

$$\textit{since, } \max_i |G_{ij}| \leq \delta\|V\Psi_k(W)\theta\|_2^2 \leq \delta L \text{ and } \frac{1}{n}\sum_{i=1}^{n} G_{ij}^2 \leq L \text{ by hypothesis}$$

$$\text{hence, } E[\exp(\frac{\lambda}{n}\sum_{i=1}^{n}(\nu_i^2-\mu^2)G_{ij})] \leq \exp(\frac{1}{2}\frac{4\lambda^2\mu^4}{n}L) \quad \text{since } \frac{1}{n}\sum_{i=1}^{n}(G_{ij})^2 \leq L.$$

Given the expression of the Laplace transform, we know that we are faced with a sub-exponential random variable with parameter $\kappa = \frac{2\mu^2\sqrt{L}}{\sqrt{n}}$ and $b = \frac{4\mu^2\delta L}{n}$. Hence, from (3.33), we can conclude that, for all j, $|M_j| \leq \max\left\{\frac{2\mu^2\sqrt{L}}{\sqrt{n}}\sqrt{t}, \frac{4\mu^2 L}{n}t\right\}$ with probability greater than $1-\exp(-\frac{t}{2})$.

From the results $(i), (ii), (iii)$ and (iv), we can conclude that with probability greater than $1-5\exp(-\frac{t}{2})$, with t large enough (we note here that $P(A\cap B)\geq P(A)+P(B)-1$). Hence, there exists $\lambda_0 \in \mathbb{R}$ such that $\|\gamma-\Gamma\theta\|_\infty \leq \lambda_0$, with

$$\begin{aligned}
\lambda_0 = \|X\|_\infty(\mu\sqrt{L}+\sigma)\sqrt{\frac{2(1+t)\log(k)}{n}} &+ \max\left\{\frac{2\mu^2\sqrt{L}}{\sqrt{n}}\sqrt{t}, \frac{4\mu^2 L}{n}t\right\} \\
&+ (\sigma^2+\mu^2)\max\left\{\sqrt{\frac{t}{n}}, 2\delta\frac{t}{n}\right\}.
\end{aligned}$$

Thus (3.34) can be rewritten as,

$$\Delta^\top\Gamma\Delta \leq 2\lambda_0\|\Delta\|_1 + 2\lambda(\|\theta\|-\|\hat{\theta}\|_1).$$

We know that $S=\{j:\theta_j\neq 0\}$ and by $N=\{j:\theta_j=0\}$. Hence,

$$\begin{aligned}
\lambda(\|\theta\|-\|\hat{\theta}\|_1) &\leq \lambda\|\theta_S\| - \lambda\|\hat{\theta_S}\|_1 - \lambda\|\hat{\theta_S}\|_1 \quad \text{since } \|\theta_N\|=0 \\
&\leq \lambda\|\hat{\theta}_S-\theta_S\|_1 - \lambda\|\hat{\theta}_N\|_1
\end{aligned}$$

therefore,

$$\begin{aligned}
\Delta^\top\Gamma\Delta &\leq 2(\lambda_0+\lambda)\|\hat{\theta}_S-\theta_S\|_1 + 2(\lambda_0-\lambda)\|\hat{\theta}_N-\theta_N\|_1 \\
&\leq 3\lambda\|\hat{\theta}_S-\theta_S\|_1 - \lambda\|\hat{\theta}_N-\theta_N\|_1, \quad \text{assuming } \lambda_0 \leq \frac{\lambda}{2} \\
&\leq 3\lambda\|\hat{\theta}_S-\theta_S\|_1
\end{aligned}$$

From the above, we can infer that Δ verifies the cone property with $c_0 = 3$. Hence, given the Restricted Eigenvalues condition (3.20), we can write that $\Delta^\top \Gamma \Delta \geq t_0 \|\Delta_S\|_2^2$. Since, $\|\Delta_S\|_1^2 \leq s_\theta \|\Delta_S\|_2^2$, we have $\Delta^\top \Gamma \Delta \geq \frac{t_0}{s_\theta} \|\Delta_S\|_1^2$. Then, from the above, we have,

$$\frac{t_0}{s_\theta} \|\Delta_S\|_1^2 \leq 3\lambda \|\Delta_S\|_1, \qquad \text{then} \qquad \|\Delta_S\|_1 \leq 3\lambda \frac{s_\theta}{t_0}.$$

Furthermore, we know that $\|\Delta_N\|_1 \leq c_0 \|\Delta_S\|_1$, thus $\|\Delta\|_1 \leq (c_0 + 1)\|\Delta_S\|_1$. We can then conclude that

$$\|\hat{\theta} - \theta\|_1 \leq \lambda \cdot s_\theta \cdot \frac{12}{t_0}$$

with probability larger than $1 - 5\exp(-\frac{t}{2})$, and t large enough.

□

Chapter 4

Slope for high dimensional linear models with measurement errors

This chapter is based on the manuscript [25].

Abstract

We study the linear regression model $Y = X \cdot \beta^\star + \epsilon$, while considering that the $n \times p$ design matrix X is subject to an additive noise given by $W = X + U$, where n is the number of elements of our dataset, p is the number of covariates and $\beta^\star$ is our true parameter of interest. The response $Y \in R^n$ and the matrix of covariates $W \in R^{n \times p}$ are observed. The $n \times p$ matrix U is drawn from a Gaussian distribution with a known covariance matrix C_U. We consider a Slope based optimization procedure to estimate our parameter of interest. A correction of the least-squares criterion is inevitable to take into account measurement errors. The corrected risk is no longer convex and we project it on the space of convex quadratic functions. We give sufficient conditions on the errors' distribution in order to attain the optimal rates and quantify the loss in the rate otherwise.

4.1 Introduction

We consider the errors-in-variables linear regression model where we observe Y and W satisfying

$$\begin{cases} Y &= X\beta^\star + \epsilon \\ W &= X + U \end{cases} \tag{4.1}$$

where $Y = (Y_1, \cdots, Y_n)^\top$ is the vector of responses, $X = [X_{ij}]_{1\le i\le n, 1\le j\le p}$ is the $n \times p$ design matrix of covariates and $\epsilon = (\epsilon_1, \cdots, \epsilon_n)^\top$ is a Gaussian vector, having centered independent and identically distributed elements with variance σ^2. Furthermore, $W = [W_{ij}]_{1\le i\le n, 1\le j\le p}$ is a $n \times p$ matrix of noisy covariates X, that is X being polluted by additive measurement errors $U \in \mathbb{R}^{n\times p}$. We will assume that the rows of U are drawn independently from a centered Gaussian multivariate distribution such that its $p\times p$ covariance matrix is denoted by C_U.

We want to estimate the unknown p-dimensional parameter $\beta^\star$. It is assumed *s-sparse*, hence $\|\beta^\star\|_0 \le s$. The model is considered to be high dimensional which usually implies that the number of covariates p is assumed (much) larger than the number of elements n of the dataset $(p > n)$. The variables ϵ and U will be considered independent.

We call the direct model the linear regression model where Y and X are observed. The Slope estimator $\hat\beta$ of $\beta^\star$ is defined through the following optimization problem

$$\hat\beta \in \arg\min_{\beta\in\mathbb{R}^p}\left\{\frac{1}{\sqrt{n}}\|Y - X\beta\|_2^2 + \|\beta\|_\star\right\} \tag{4.2}$$

with $\|\beta\|_\star = \sum_{i=1}^p \lambda_i|\beta|_{(i)}$, known as the *Slope* norm, where $\lambda_1 \ge \cdots \ge \lambda_p > 0$ are tuning parameters.

In this paper, we introduce a Slope based estimator for errors-in-variables regression model (4.1) and study its properties.

There is a vast literature on high dimensional regression methods. Most popular and immensely studied are the Lasso estimator (*Least Absolute Shrinkage and selection operator*), which was introduced in [72] and the Dantzig selector [19]. Its full theoretical study was developed through multiple subsequent articles and several books such as [39] and [36] that provide an exhaustive list of its properties. Recently, it was noted that the Lasso attains a suboptimal rate that can be improved by the Slope *Sorted L-One Penalized Estimation*, see [69] and [50]. The authors introduced a convex optimization problem with a different penalty whose algorithm solution has a complexity similar to most common ℓ_1 penalization procedures. They provided a path for solving the question of variable selection through a non-orthogonal design. The properties, namely the convergence rate, of both aforementioned methods were improved to $\sqrt{s\log(p/s)/n}$ instead of $\sqrt{s\log(p)/n}$ for the

Lasso. In [4], the authors were able to prove that both Lasso (with known or estimated sparsity s) and Slope estimators achieve the minimax prediction and l_2 estimation rate of $\sqrt{s\log(p/s)/n}$. The minimax optimal bounds were also obtained for an l_q estimation error, with $1 \leq q \leq 2$.

Many derivations of the Lasso were thoroughly studied. Let us mention the *Square Root Lasso*, which was introduced in [5]. The authors showed that by minimizing the squared root of the least-squares criterion penalized with the usual ℓ_1 norm, the Square Root Lasso estimator did not rely on the noise of the model, and was able to achieve minimax optimal rate of $\sigma\sqrt{s\log(p)/n}$, where σ is the variance of the noise of the model. Based on some of the results developed in [4], the author in [30] improved the estimation rate of the Square Root Lasso and introduced the *Square Root Slope* estimator. He showed that both estimators can achieve optimal minimax rate of $\sqrt{s\log(p/s)/n}$. The author was able to preserve the adaptivity to σ.

In many examples, measurement errors are inevitable. Errors-in-variables (EIV) regression models (4.1) have been largely considered in the literature, mostly in parametric or semi-parametric setups. More recently, in high dimensional EIV linear regression models, [62] have proved that the regular Lasso or Dantzig estimator is unstable. They then introduced matrix uncertainty selectors and their improvements in [6], which are stable, consistent and capable of reproducing a sparsity pattern common to most high dimensional settings, see also [7]. We can also mention the authors in [67] who introduced a corrected Lasso estimator and were able to prove its sign consistency, among other properties. Similarly, the authors in [27] introduced a corrected convex optimization problem to tackle the errors-in-variables model. They developed a convex corrected Lasso estimator and showed that it was sign consistent. Lastly, the authors in [48] opted for a non-convex optimization problem from which they derived the estimator's error bounds and showed that a gradient descent based algorithm allowed the estimator to converge in polynomial time.

In this paper, we consider a corrected least-square criterion that we project on the space of convex quadratic functions. We introduce a Slope penalized procedure for estimating the errors-in-variables linear regression model (4.1) and study its properties. We describe sufficient conditions on the parameters of the model in order to attain the rates known optimal in the direct case. As expected, these conditions are quite stringent, however we also give upper bounds on the slower rates that are attained in the opposite case. As discussed at the beginning of Section 4.6.2, these results can be extended to sub-Gaussian random rows under an additional assumption, following [16].

Notation and organization of the article

For a given vector $v \in \mathbb{R}^p$, we use the notation $\|v\|_q$ for the l_q norm, with $1 \leq q \leq \infty$, and $\|v\|_0$ for the number of non-zero coordinates of v. For any set of coordinates $S \subset \{1, \cdots, p\}$,

we denote by v_S, the vector $(v_i \mathbb{1}\{i \in S\})_{i=1,\cdots,p}$. Furthermore, we denote by $v_{(j)}$, the j-th largest component of v. As an example, $|v|_{(k)}$ is the k-th largest component of the vector $|v|$, whose components are the absolute value of the components of v. We will use the notation $\langle \cdot, \cdot \rangle$ for the inner product with respect to the Euclidean norm and $(e_j)_{j=1,\cdots,p}$ for the canonical basis in $\mathbb{R}^p$. We will also use the notation $(a)_+ = \max\{a, 0\}$, $a \vee b = \max\{a, b\}$ and $a \wedge b = \min\{a, b\}$ for all $a, b \in \mathbb{R}$. Similarly, for all $v \in \mathbb{R}^p$ and $A \in \mathbb{R}^{p\times p}$, we will have $v_+ = (\max\{v_i, 0\})_{i=1,\cdots,p}$. The transpose of a matrix X is noted $X^\top$. If we consider a random variable z, the symbol $E[z]$ denotes its expected value. Let us denote by M a $p \times p$ matrix. We will consider that $\|M\|_F = \sqrt{trace(M^\top M)} = \sqrt{\Sigma_{i=1}^p \eta_i(M)}$ is the Frobenius norm, where η_i is the i-th eigenvalue of $M^\top M$. Similarly, $\|M\|_2 = \sup_{\beta\neq 0} \frac{\|M\beta\|_2}{\|\beta\|_2}$, is the spectral or operator norm. Furthermore, $\|M\|_1$ (respectively $\|M\|_\infty$) is the l_1 induced norm (respectively the l_∞ induced norm). In general, an l_q induced norm is given by $\|M\|_q = \sup_{\beta\neq 0} \frac{\|M\beta\|_q}{\|\beta\|_q}$.

The article will be presented as follows. Section 4.2 presents the adjustments performed on the optimization problem (4.2) in order to construct our estimator $\hat{\beta}$ of $\beta^\star$. Section 4.3 gives upper bounds of the estimation risk of $\hat{\beta}$. A numerical analysis of the estimation procedure is provided in section 4.4. All proofs are developed in Section 4.6.

4.2 Estimation procedure

Penalized methods are based on unbiased estimators of $E\left(\frac{1}{n}\|Y - X\beta\|_2^2\right)$. In order to take into account the errors-in-variables, a correction for the errors is necessary and we consider

$$F(\beta) := E\left(\frac{1}{n}\|Y - W\beta\|_2^2\right) - \beta^\top C_U \beta, \quad \text{where } C_U = \frac{1}{n}E[U^\top U],$$

C_U is the covariance matrix of the p-dimensional vector U.

The empirical version given by

$$\frac{1}{n}\|Y - W\beta\|_2^2 - \beta^\top C_U \beta = \beta^\top \left(\frac{1}{n}W^\top W - C_U\right)\beta - \frac{2}{n}Y^\top W\beta + \frac{1}{n}Y^\top Y \tag{4.3}$$

has the major inconvenient of being a quadratic form which is not necessarily convex. Indeed, the symmetric matrix

$$\hat{\Sigma} := \frac{1}{n}W^\top W - C_U = \frac{1}{n}X^\top X + \frac{1}{n}\left(X^\top U + U^\top X\right) + \frac{1}{n}U^\top U - C_U$$

may have negative eigenvalues. Therefore, we project $\hat{\Sigma}$ on the set $\mathcal{S}_{\geq 0}$ of symmetric positive semi-definite matrices in Frobenius norm. We denote by

$$\widetilde{\Sigma} = \arg\min_{M\in\mathcal{S}_{\geq 0}} \|\hat{\Sigma} - M\|_F. \tag{4.4}$$

The following shows that $\widetilde{\Sigma}$ defined in (4.4) also minimizes the operator norm over $\mathcal{S}_{\geq 0}$.

Lemma 1. *The matrix $\widetilde{\Sigma}$ defined in (4.4) is such that*

$$\widetilde{\Sigma} \in \arg\min_{M \in \mathcal{S}_{\geq 0}} \|\hat{\Sigma} - M\|_2. \tag{4.5}$$

The proof is in Section 4.6.3.

Let us see that $\widetilde{\Sigma}$ is easily constructed from $\hat{\Sigma}$. We note here that $\hat{\Sigma}$ is a $p \times p$ symmetric matrix and thus, it allows a spectral decomposition. Let us denote by D a $p \times p$ diagonal matrix whose elements are the eigenvalues $(\hat{\eta}_j)_{j=1,\dots,p}$ of the matrix $\hat{\Sigma}$. We also denote by V a $p \times p$ orthogonal matrix of eigenvectors of $\hat{\Sigma}$. As such, we can write $\hat{\Sigma} = V^\top D V$. From the equation (4.4), we can then deduce that

$$\widetilde{\Sigma} = V^\top D_+ V, \tag{4.6}$$

with D_+ the diagonal matrix whose elements are the positive parts $(\hat{\eta}_j)_+$ of $\hat{\eta}_j$, $j = 1, \dots, p$. We define the estimator $\hat{\beta}$ of β^* by

$$\hat{\beta} \in \arg\min_{\beta \in \mathbb{R}^p} \beta^\top \widetilde{\Sigma} \beta - \frac{2}{n} Y^\top W \beta + \|\beta\|_\star \tag{4.7}$$

where $\|\beta\|_\star = \sum_{i=1}^p \lambda_i |\beta|_{(i)}$, with $\lambda_1 \geq \cdots \geq \lambda_p > 0$.

The non-convex character of the optimization problem (4.3) has been addressed through different means in the literature. For example, the authors in [48] introduced a constraint on the space of possible solutions through the equation $\|\beta\|_1 \leq b_0 \sqrt{s}$ where b_0 is a suitable constant which must verify $b_0 \geq \|\beta^\star\|_2$. This assumption is restrictive and relies on a constant which can not be necessarily implied from the dataset, since $\beta^\star$ is an unknown parameter. We can also mention the authors in [27] which have relied on a positive semidefinite projection of the matrix $\hat{\Sigma}$ in order to derive a convex optimization problem. However, the projection is an element-wise one which was then complemented by element-wise concentration inequalities in order to derive desirable error bounds. These inequalities help define suitable convergence rates but yet again it is not obvious under which global assumptions they are satisfied. In this paper, we consider a matrix projection of $\hat{\Sigma}$ in Frobenius norm on the space $\mathcal{S}_{\geq 0}$ of positive and semi-definite matrices. Using this convex criterion, we derive sufficient conditions for optimal convergence rate of our estimator, as well as quantify the loss in the rate when these conditions are not satisfied.

4.3 Sufficient conditions for optimal convergence rate

In this section, we derive sufficient conditions to attain the optimal rate for the *Slope* estimator $\hat{\beta}$. In most cases, the rates attained by our procedure are much slower due to

the high-dimensional noise that interferes with the estimation. We will illustrate a few examples where the optimal rates from the direct problem can still be attained. We recall the *Weighted Restricted Eigenvalues* (WRE) condition introduced in [4].

Definition We say that the design matrix X satisfies the $WRE(s, c_0)$ condition if

$$\max_{j=1,\cdots,p} \|Xe_j\|_n \leq 1$$

and

$$\kappa := \min_{\delta \in C_{WRE(s,c_0)}} \frac{\|X\delta\|_n^2}{\|\delta\|_2^2} > 0 \tag{4.8}$$

where $C_{WRE(s,c_0)} = \{\delta \in \mathbb{R}^p : \|\delta\|_\star \leq c_0 \|\delta\|_2 \sqrt{\sum_{j=1}^s \lambda_j^2}\}$ is a cone in $\mathbb{R}^p$.

Note that the $WRE(s, c_0)$ condition is written with respect to X which is unknown. It is possible as we will see below to deduce an analogous condition with respect to $\delta^\top \widetilde{\Sigma} \delta$ with high probability. Indeed, $\delta^\top \widetilde{\Sigma} \delta$ is the natural quadratic risk associated to our optimization problem (4.7).

Let us introduce the following notation

$$R_{n,p}(C_U, \|\beta^\star\|_2, t) = t^2 \sqrt{\frac{\|C_U\|_2}{n}} + \|\beta^\star\|_2 \left(2A_{n,p}(C_U) + 16tp \sqrt{\frac{1}{n}\|C_U\|_2} \right) \tag{4.9}$$

$$A_{n,p}(C_U) = c \cdot \max \left\{ Tr(C_U) \frac{\log(pn)}{n}, \sqrt{Tr(C_U)\|C_U\|_2 \frac{\log(pn)}{n}} \right\}, \tag{4.10}$$

with $t > 1$ and $c > 0$.

Assumption. We assume that

$$A_{n,p}(C_U) + 2t \sqrt{\frac{p}{n}\|C_U\|_2} \leq \frac{\kappa}{2}, \tag{4.11}$$

when n, p, s are large enough and for $t > 0$ large enough. In the rest of the article, A_{n,p,C_U} will be used interchangeably with A.

Lemma 2. *Under the assumption (4.11) and if X verifies the $WRE(s, c_0)$ condition then*

$$\frac{\kappa}{2} = \min_{\delta \in C_{WRE(s,c_0)}} \frac{\delta^\top \widetilde{\Sigma} \delta}{\|\delta\|_2^2} > 0 \tag{4.12}$$

with probability at least $1 - \frac{4}{n} - 2\exp(-\frac{t^2}{2})$.

The proof is in Section 4.6.3.

The next theorem gives upper bounds for the $\mathbb{L}_2$-risk. These bounds allow to deduce sufficient conditions on the parameters of the noise distribution in order to attain the known optimal rates for estimating the sparse parameter β. They also quantify the loss in the rate when these conditions are not verified.

Theorem 4.1. *Let $s \in \{1, \cdots, p\}$ and assume that the unknown design X satisfies the $WRE(s, c_0)$ condition (4.12) and that the assumption (4.11) holds. We choose the following tuning parameters*

$$\lambda_j = v \cdot \max\left\{\sqrt{\frac{1}{n}}, R_{n,p}(C_U, B, t)\right\} \cdot \log(2p/j), \tag{4.13}$$

where $t = \sqrt{2\log(2/\delta_0)}$, B is an upper bound of $\|\beta^*\|_2$ and we assume that $\delta_0 < 2e^{-1/2}$ and that

$$v \geq 2\sigma(4 + \sqrt{2}). \tag{4.14}$$

Then, with probability at least $1 - \frac{8}{n} - 4\exp(-2p(2\log(2/\delta_0) - 1)) - \delta_0$, we have

$$\|\hat{\beta} - \beta^\star\|_2 \leq \max\left\{C_1\sqrt{\frac{s}{n}\log(\frac{2ep}{s})}, C_2\sqrt{\frac{\log^2(1/\delta_0)}{sn\log(2ep/s)}}, C_3 R_{n,p}(C_U, B, t)\sqrt{s\log(\frac{2ep}{s})}\right\}, \tag{4.15}$$

where $C_1 = \frac{(3+4\sigma)\gamma}{\kappa}$, $C_2 = 4(4 + \sqrt{2})^2$ and $C_3 = \frac{28}{\kappa}$.

The proof is in Section 4.6.3.

As expected, the presence of measurement errors affects the overall convergence rates in most cases. If the covariance matrix has small enough rank and spectral norm, it is possible to attain the rates of convergence known optimal in noiseless case. Let us illustrate with the following example sufficient conditions for optimal rates of convergence. Suppose the noise W has a diagonal covariance matrix $C_U = v \cdot diag(1, ..., 1, 0, ...0)$ where the rank is $r < p$. We get $A_{n,p}(C_U) = c \cdot v\sqrt{r \cdot \log(pn)/n}$ that needs to be bounded. Moreover,

$$R_{n,p} = \sqrt{\frac{s\log(pn)}{n}}(v\sqrt{r} + p\sqrt{v})$$

is smaller than $\sqrt{\frac{s}{n}\log(\frac{2ep}{s})}$ provided that

$$\max\{v^2 r, vp^2\} \ll \frac{\log(2ep/s)}{\log(pn)}.$$

For example, $v = p^{-2-\epsilon}$ and any $r \in \{1, ..., p\}$ check the constraints, that is a full rank covariance matrix $C_U = p^{-2-\epsilon} I_p$ or $p^{-2-\epsilon} diag(1, ..., 1, 0, ...0)$.

4.4 Simulation Analysis

We highlight below experiments that were conducted to illustrate the optimization methodology, on a predefined dataset. The implementation is based on the *coordinate descent algorithm.*

4.4.1 Construction of datasets

The $n \times p$ elements of the matrix X are drawn from i.i.d $[0,1]$- uniform distribution. The elements of ϵ are drawn from a centered Gaussian distribution with standard deviation given by σ. The matrix C_U is built using a centered Gaussian distribution, with standard deviation μ (we assumed that it exists a $c > 0$ such that $\mu = c \cdot \sigma$) such that C_U is symmetric.

Two cases are considered in order to ensure that C_U is symmetric. In the first one, we only draw the diagonal term from the aforementioned Gaussian distribution and we further set all of the non-diagonal term to be equal to 0. In the second case, we draw all the $p \times p$ elements of a matrix M from the same Gaussian distribution. The matrix C_U is implied from M using the formula $C_U = \frac{1}{p} M^\top M$. The elements of the parameter of interest $\beta^\star$ are drawn from a $[0,1]$–uniform distribution such that $\|\beta^\star\|_0 = s$. Hence, for known n, p, σ, s and c, we can infer using the above methodology a predefined dataset. In fact, having drawn the parameters X, $\beta^\star$, C_U, we then draw the elements of the $n \times p$ matrix of U from a centered multivariate Gaussian distribution with matrix of covariance given by C_U. The elements of W are given by $W = X + U$. Similarly, the elements of Y, will be given by $Y = X \cdot \beta^\star + \epsilon$.

4.4.2 Convergence metrics

Let us now considered a predefined dataset given by $O = (W, Y)$ of size n which is derived using the methodology described above. We then compute an estimator of our parameter of interest $\hat{\beta}$, defined by the optimization problem (4.7). To assess the optimization methodology of our estimator, we compute its *Estimate Risk* value, for several sample sizes n, for a given p, through a two-step cross-validation process.

In fact, we partitioned the sample in l distinct subsets O^{subset}, of equal size. In the first step, we set aside one of the $O^{subset}_{\{i\}}$, which will be used as the test set, and we use the $l-1$ remaining subsets as training sets. Using the latter, we then compute the estimator $\hat{\beta}$. In the second step, we use the subset $O^{subset}_{\{i\}}$ (test set left out from the initial estimation), to determine how close the estimator $\hat{\beta}$ is to its real value. These two steps are repeated successively l times. For each iteration, we calculate the difference between the estimator and the true oracle value, using the test set. The empirical average of these l distances give us the desired *Estimation Risk* measure. In brief, the *Estimation Risk (ER)* value of a sample of size n is given by

$$ER = \frac{1}{l} \sum_{i=1}^{l} \frac{1}{|O_{\{i\}}|} \|\beta^\star - \hat{\beta}\|_2^2,$$

where $|O_{\{i\}}|$ is the number of elements of the subset $O^s_{\{i\}}ubset$. For a single sample size n, we computed the *Estimation Risk* $u(>= 20)$ times. It allowed us to imply the confidence interval associated with the measure.

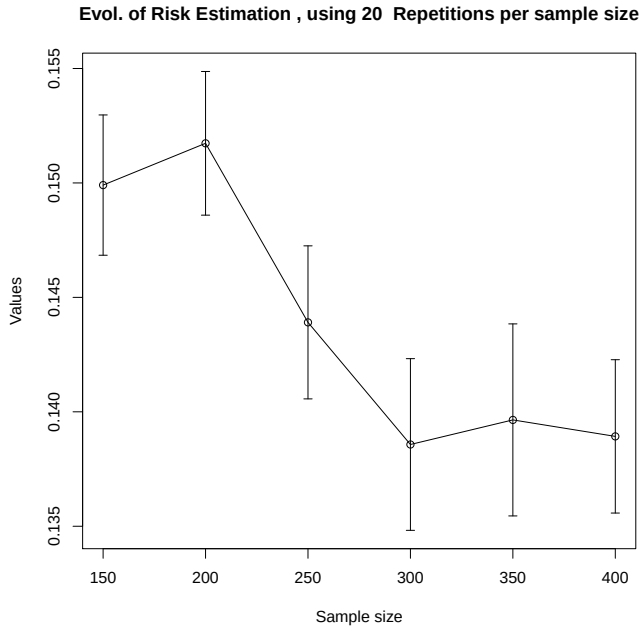


Figure 4.1 – Evolution of the *Estimation Risk* for different value of n - First case.

The exact same methodology was used to imply the *Prediction Risk* with our estimate given by $X \cdot \hat{\beta}$. As such the *Prediction Risk (PR)* value of a sample size n was given by

$$PR = \frac{1}{l} \sum_{i=1}^{l} \frac{1}{|O_{\{i\}}|} \|X \cdot \beta^{\star} - X \cdot \hat{\beta}\|_2^2.$$

Similarly, for a single sample size, we computed the *Prediction Risk* u times. It allowed us to imply the confidence interval of the measure. It is important to note that for a convergent estimation procedure, the values PR and ER should converge to 0, as n increases, for a given p.

4.4.3 Results

We show below two set of results corresponding to the evolution of the key metrics: *Estimation Risk, Prediction Risk*. We also show for completeness at each step the value taken by key noise parameters such as $Tr(C_u)$ and $\|C_U\|_2$. In each of these two cases, the value p is fixed while n increases. It is important to note that the synthetic data construction methodology follows the step described in section 4.4.1. For both cases, $u = 20$.

First case. $p = 500$, $\sigma = 1$, $\mu = .05$ and $n = \{150, 200, 250, 300, 350, 400\}$.

As the value of n increases, we realize that the *Prediction Risk* and *Estimation Risk* converges steadily towards 0. Furthermore, $Tr(C_U)$ and $\|C_U\|_2$ are well below critical levels.

Second case. $p = 500$, $\sigma = 1$, $\mu = .1$ and $n = \{150, 200, 250, 300, 350, 400\}$.

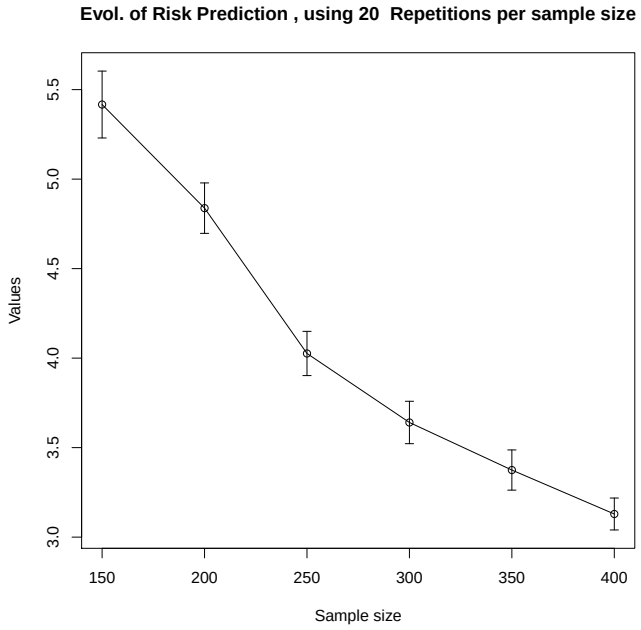


Figure 4.2 – Evolution of the *Prediction Risk* for different value of n - First case.

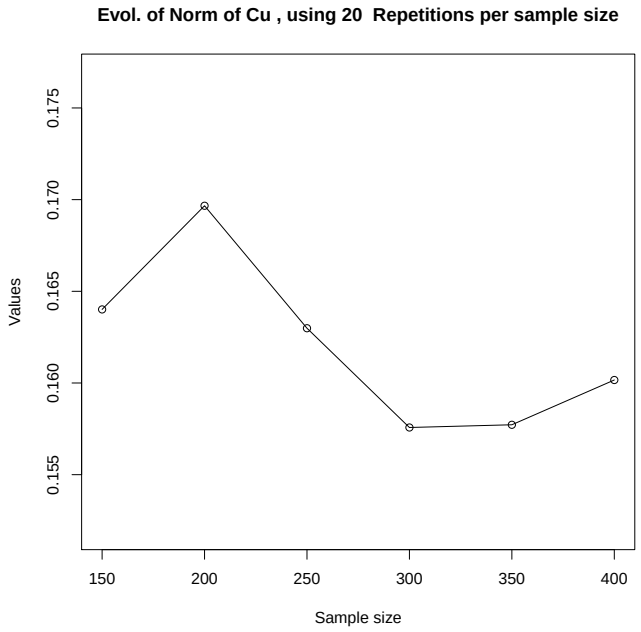


Figure 4.3 – Levels taken by $\|C_U\|_2$ at each step of the iteration process - First case.

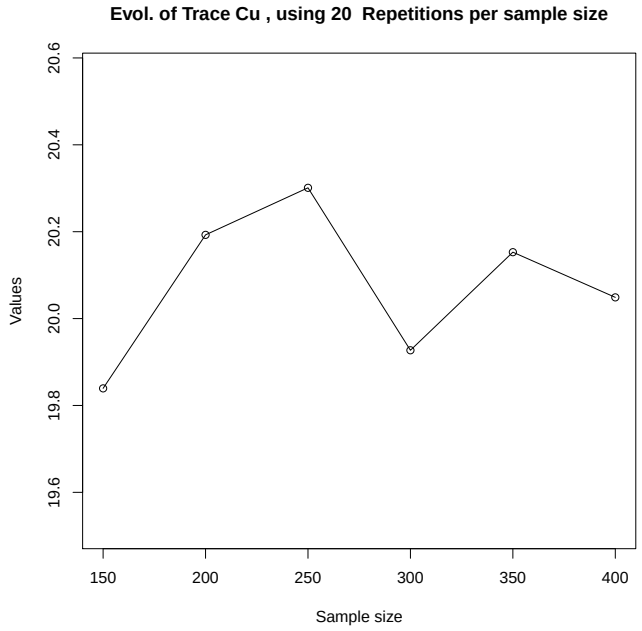


Figure 4.4 – Levels taken by $Tr(C_U)$ at each step of the iteration process - First case. - First case.

Similarly as the first case, the overall conclusion is the same. However, we see a deterioration of the *estimation risk*, even though the values themselves remain very small.

Remark. As the value of μ further increases, the deterioration of the overall estimation metrics become more and more flagrant. The results are not shown here.

4.5 Conclusion

We have established sufficient conditions under which we can reach optimal rates for our *Slope* based estimators, given a polluted design matrix. The impact on rates is also outlined when these conditions are not fulfilled. Our approach is novel since it relies on a Frobenius based projection in order to derive a convex optimization problem. The study was completed by a simulation study which illustrated the accuracy and correctness of our optimization methodology, with an implementation based on the *coordinate descent algorithm*.

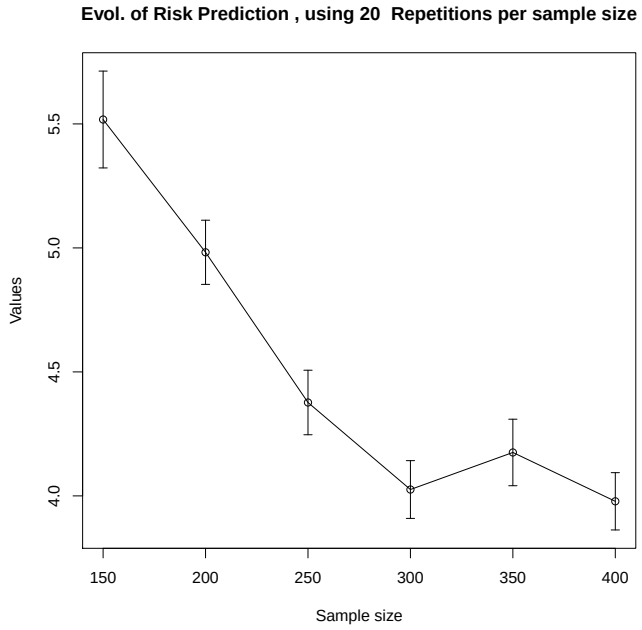


Figure 4.5 – Evolution of the *Prediction Risk* for different value of n - Second case.

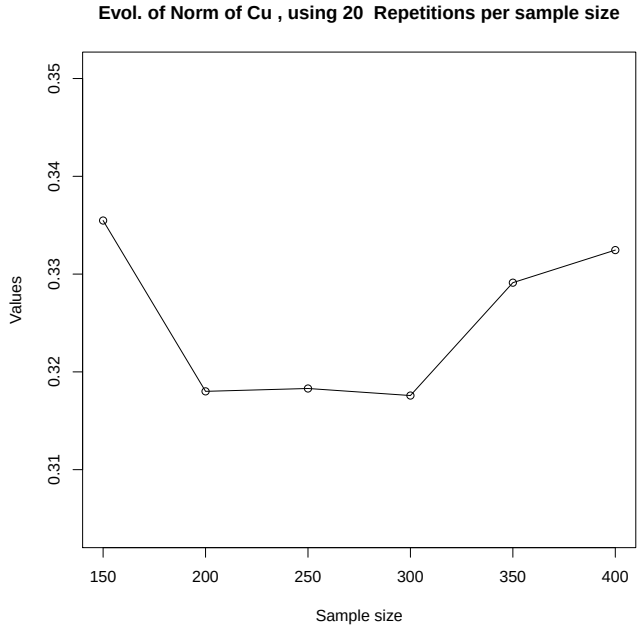


Figure 4.6 – Levels taken by $\|C_U\|_2$ at each step of the iteration process - Second case.

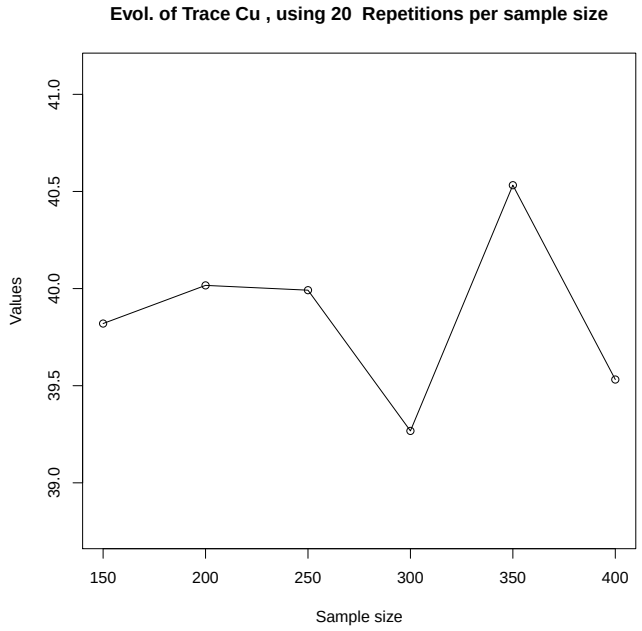


Figure 4.7 – Levels taken by $Tr(C_U)$ at each step of the iteration process - Second case.

4.6 Proofs

4.6.1 Preliminary results

The following result is demonstrated in [36, p. 112]. It has been used as a lemma in [30] and is reproduced here for the reader's convenience.

Lemma 3. *With probability at least* $1-(1+e^2)e^{-n/24}$*, we have*

$$\frac{\sigma}{\sqrt{2}} \leq \frac{\|\epsilon\|_2}{\sqrt{n}} \leq 2\sigma.$$

Similarly, the following results are cited for reader's convenience. In the following Lemmas 4 and 5, we use

$$\lambda_j = \gamma\sqrt{\frac{\log(2p/j)}{n}}, \quad \text{for } \gamma \geq 4(4+\sqrt{2}) > 0 \text{ large enough.}$$

Lemma 4 ([30])**.** *We have,*

$$\gamma\sqrt{(s/n)\log(2p/s)} \leq \sqrt{\sum_{j=1}^{s}\lambda_j^2} \leq \gamma\sqrt{(s/n)\log(2ep/s)}.$$

Lemma 5 ([4, Theorem 4.1])**.** *Let* $0<\delta_0<1$ *and let* $X \in \mathbb{R}^{n\times p}$ *be a matrix such that*

$$\max_{j=1,\cdots,p}\|Xe_j\|_n \leq 1.$$

For any $u = (u_1, \cdots, u_p) \in \mathbb{R}^p$, *we define :*

$$G(u) := (4+\sqrt{2})\sigma\sqrt{\frac{\log(1/\delta_0)}{n}}\|Xu\|_n,$$

$$H(u) := (4+\sqrt{2})\sum_{j=1}^{p}|u|_{(j)}\sigma\sqrt{\frac{\log(2p/j)}{n}},$$

$$F(u) := (4+\sqrt{2})\sigma\sqrt{\frac{\log(2p/s)}{n}}\left(\sqrt{s}\|u\|_2 + \sum_{j=s+1}^{p}|u|_{(j)}\right).$$

If $\epsilon \sim \mathcal{N}(0, \sigma^2 I_{n\times n})$, *then the random event*

$$\left\{\frac{1}{n}\epsilon^\top Xu \le \max\left(H(u), G(u)\right), \forall u \in \mathbb{R}^p\right\}$$

is of probability at least $1-\delta_0/2$. *Moreover, by the Cauchy-Schwarz inequality, we have* $H(u) \le F(u)$, *for all* $u \in \mathbb{R}^p$.

4.6.2 Auxiliary results

We say that a real valued random variable V is ν^2-subGaussian if $E[e^{tV}] \le \exp(t^2\nu^2/2)$, for all t. We denote by $\|\cdot\|_{\psi_2}$ the sub-Gaussian norm defined by $\|X\|_{\psi_2} = \sup_{k\le 1} k^{-\frac{1}{2}}(E[|X|^k])^{\frac{1}{k}}$. Following [16], we say that a random vector V of $\mathbb{R}^p$ is subGaussian if $V^\top u$ is subGaussian for all p-dimensional vector u and define the subGaussian norm using the Orlicz norm $\|V\|_{\psi_2} = \sup_{u:\|u\|_2=1}\|V^\top u\|_{\psi_2}$. We also make the additional assumption that there exists a constant $c_0 > 0$ such that

$$c_0\|V^\top u\|_{\psi_2}^2 \le u^\top Cu, \quad \text{for all } u \in \mathbb{R}^p. \tag{4.16}$$

This implies that $V^\top u$ is $\|u\|_2^2\|C\|_2/c_0$-subGaussian. Note that, if in particular V has a Gaussian distribution $\mathcal{N}_p(0, C)$, then $V^\top u$ is $\|u\|_2^2\|C\|_2$-subGaussian. We say that a real valued random variable V is (ν^2, b)-subexponential with parameters (ν^2, b), with $b > 0$, if

$$E[e^{tV}] \le \exp\left(\frac{t^2\nu^2}{2}\right), \quad \text{for all } |t| \le \frac{1}{b}.$$

If V is (ν^2, b)-subexponential then the Bernstein inequality gives that

$$|V| \le (\nu\sqrt{u}) \vee (bu), \quad \text{with probability larger than } 1-2e^{-u/2}, \text{ for all } u > 0.$$

Lemma 6. *Let us consider two arbitrary vectors* $v, w \in \mathbb{R}^p$. *If* U *has independent rows, identically distributed as a subGaussian vector satisfying (4.16) we have, with probability at least* $1-\frac{4}{n}$ *:*

$$|v^\top(\frac{1}{n}U^\top U - C_U)w| \le c\|v\|_2\cdot\|w\|_2\cdot\max\left\{Tr(C_U)\frac{\log(pn)}{n}, \sqrt{Tr(C_U)\|C_U\|_2\frac{\log(pn)}{n}}\right\},$$

where $c > 0$ *depends only on* c_0. *Note that, if the rows* $\{U_{i\cdot}\}_{i=1,\ldots,n}$ *are Gaussian, then* $c_0 = 1$.

Proof of Lemma 6. The proof of the lemma is a direct consequence of the application of *Cauchy-Schwarz* inequality and the Theorem 2.1 in [16]. Hence,

$$|v^\top(\frac{1}{n}U^\top U - C_U)w| \le \|v\|_2 \cdot \|w\|_2 \cdot \|\frac{1}{n}U^\top U - C_U\|_2$$
$$\le \|v\|_2 \cdot \|w\|_2 \cdot c \cdot \max\left\{Tr(C_U)\frac{\log(pn)}{n}, \sqrt{Tr(C_U)\|C_U\|_2\frac{\log(pn)}{n}}\right\}$$

with probability at least $1 - \frac{4}{n}$. □

Lemma 7. *For all $t > 1$ and $\beta \in \mathbb{R}^p$, we will have*

$$|\frac{1}{n}\epsilon^\top U\beta| \le \frac{\sigma}{\sqrt{n}}\|\beta\|_2\|C_U\|_2^{1/2}t,$$

with probability larger than $1 - 2e^{-t/2}$.

Proof of Lemma 7. The scalar product $\frac{1}{n}\epsilon^\top U\beta$ can be rewritten as:

$$\frac{1}{n}\epsilon^\top U\beta = \frac{1}{n}\sum_{i=1}^n \epsilon_i U_i \cdot \beta, \quad \text{where } U_{i\cdot}\cdot\beta = \sum_{j=1}^p U_{ij}\beta_j.$$

The Laplace transform of the above random variable writes, for all $t \in \mathbb{R}$ and $\beta \in \mathbb{R}^p$

$$E\left[\exp\left(\frac{t}{n}\epsilon^\top U\beta\right)\right] = E\left[\exp\left(\sum_{i=1}^n \frac{t}{n}\epsilon_i U_{i,\cdot}\cdot\beta\right)\right]$$
$$= E\left(\prod_{i=1}^n E\left[\exp\left(\frac{t}{n}\epsilon_i(U_{i,\cdot}\beta)\right)|U_{i,\cdot}\right]\right)$$
$$\le E\left[\prod_{i=1}^n \exp\left(\frac{t^2}{2n^2}\sigma^2(U_{i,\cdot}\beta)^2\right)\right]$$
$$= E\left[\exp\left(\frac{t^2}{2n}\sigma^2(U_{1,\cdot}\beta)^2\right)\right],$$

where we used that ϵ_i are Gaussian i.i.d. and that the rows $U_{i,\cdot}$ are independent. Now, since $U_{1,\cdot}$ is a $\|C_U\|_2$-subGaussian vector, then $(U_{1,\cdot}\cdot\beta)^2$ is subexponential and that implies

$$E\left[\exp\left(\frac{t^2\sigma^2}{2n}(U_{1,\cdot}\beta)^2\right)\right] \le E\left[\exp\left(\frac{t^2\sigma^2}{2n}\cdot\|\beta\|_2^2\|C_U\|_2(\frac{U_{1,\cdot}\beta}{\sqrt{\beta^\top C_U\beta}})^2\right)\right] = E\exp(\lambda Z^2),$$

where Z is a standard Gaussian random variable and $\lambda = \frac{t^2\sigma^2}{2n}\cdot\|\beta\|_2^2\|C_U\|_2$. Now,

$$E\exp(\lambda Z^2) = e^\lambda E\exp(\lambda(Z^2-1)) \le \frac{1}{\sqrt{1-2\lambda}} \le e^\lambda,$$

for $\lambda < 1/2$. In conclusion,

$$E\left[\exp\left(\frac{t}{n}\epsilon^\top U\beta\right)\right] \le \exp\left(\frac{t^2}{2n}\sigma^2\|\beta\|_2^2\|C_U\|_2\right), \quad \text{for all } \frac{t^2}{n}\sigma^2\|\beta\|_2^2\|C_U\|_2 \le 1$$

which means that $\frac{1}{n}\epsilon^\top U\beta$ is $(\frac{\sigma^2}{n}\|\beta\|_2^2\|C_U\|_2, \frac{\sigma}{\sqrt{n}}\|\beta\|_2\|C_U\|_2^{1/2})$-subexponential, and we get the result. □

Lemma 8. *For all $t \in \mathbb{R}$, and $\beta, u \in \mathbb{R}^p$, we will have*

$$\left|\frac{1}{n}\beta^\top U^\top X u\right| \le t \cdot \|u\|_2 \|\beta\|_2 \sqrt{\frac{p}{n} \cdot \|C_U\|_2}$$

with probability $1 - 2e^{-\frac{t^2}{2}}$

Proof of Lemma 8. Let us bound from above the Laplace transform,

$$\begin{aligned}
&E\left[\exp(\frac{t}{n}\beta^\top U^\top X u)\right] \\
&= E\left[\exp\left(\frac{t}{n}\sum_{k=1}^{n}\beta^\top U_{k.}^\top X_{k.} u\right)\right] \\
&= \prod_{k=1}^{n} E\left[\exp\left(\frac{t}{n} X_{k.} u \cdot U_{k.}\beta\right)\right] \\
&= \prod_{k=1}^{n} \exp(\frac{t^2}{2n^2}(X_{k.}u)^2 \cdot \beta^\top C_U \beta) \\
&\le \exp\left(\frac{t^2}{2n^2}\sum_{k=1}^{n}\sum_{j=1}^{p} X_{kj}^2 \|u\|_2^2 \|\beta\|_2^2 \|C_U\|_2\right) \\
&\le \exp\left(\frac{t^2 p}{2n}\|u\|_2^2\|\beta\|_2^2\|C_U\|_2\right) \quad \text{, since} \quad \frac{1}{n}\sum_{k=1}^{n} X_{kj}^2 \le 1.
\end{aligned}$$

By the Bernstein inequality,

$$\left|\frac{1}{n}\beta^\top U^\top X u\right| \le t \cdot \|u\|_2 \|\beta\|_2 \sqrt{\frac{p}{n} \cdot \|C_U\|_2},$$

with probability larger than $1 - 2e^{-\frac{t^2}{2}}$. □

Lemma 9. *For all $t > 1$, $c > 0$ and $\beta, u \in \mathbb{R}^p$, we will have*

$$\begin{aligned}
\left|u^\top\left(\widetilde{\Sigma} - \hat{\Sigma}\right)\beta\right| \quad \le \quad & \|\beta\|_2 \cdot \|u\|_2 \cdot \left\{c \cdot \max\left\{Tr(C_U)\frac{\log(pn)}{n}, \sqrt{Tr(C_U)\|C_U\|_2\frac{\log(pn)}{n}}\right\}\right. \\
& \left. + \|\frac{1}{n}(X^\top U + U^\top X)\|_2\right\},
\end{aligned}$$

with probability $1 - \frac{4}{n}$. Moreover,

$$\|\frac{1}{n}(X^\top U + U^\top X)\|_2 \le 3tp\sqrt{\frac{2\log(p)}{n}\max_{1\le j\le p} C_U^{jj}}$$

with probability larger than $1 - 4\exp(-(t^2-1)\log(p))$, and

$$\|\frac{1}{n}(X^\top U + U^\top X)\|_2 \le 16tp\sqrt{\frac{1}{n}\|C_U\|_2}$$

with probability larger than $1 - 2\cdot\exp(-2p(t^2-1))$.

Proof of Lemma 9. We know that through Cauchy-Schwarz inequality we can write

$$|\delta^\top(\widetilde{\Sigma}-\hat{\Sigma})\beta^\star| \leq \|\delta\|_2 \|\widetilde{\Sigma}-\hat{\Sigma}\|_2 \|\beta^\star\|_2.$$

We also know that $\widetilde{\Sigma}$ is the positive semi-definite projection of $\hat{\Sigma}$. As such, for all $\Sigma_0 \in \mathcal{S}_{\geq 0}$, we will have $\|\widetilde{\Sigma}-\hat{\Sigma}\|_2 \leq \|\Sigma_0-\hat{\Sigma}\|_2$. Let us choose $\Sigma_0 = \frac{1}{n}X^\top X$ which is positive semi-definite.

$$\begin{aligned}
\|\Sigma_0-\hat{\Sigma}\|_2 &= \|\frac{1}{n}X^\top X - \frac{1}{n}W^\top W + C_U\|_2 \\
&= \|\frac{1}{n}X^\top X - \frac{1}{n}X^\top X - \frac{1}{n}U^\top U - \frac{1}{n}X^\top U - \frac{1}{n}U^\top X + C_U\|_2 \\
&= \|\left(C_U - \frac{1}{n}U^\top U\right) - \left(\frac{1}{n}X^\top U + \frac{1}{n}U^\top X\right)\|_2 \\
&\leq \|C_U - \frac{1}{n}U^\top U\|_2 + \|\frac{1}{n}X^\top U + \frac{1}{n}U^\top X\|_2. \qquad (4.17)
\end{aligned}$$

We give 2 different bounds on this last quantity. Firstly, the squared matrix $\frac{1}{n}(X^\top U + U^\top X)$ is symmetric. Therefore, its spectral norm is the smallest of all matrix induced norms. Hence, we can write

$$\|\frac{1}{n}(X^\top U + U^\top X)\|_2 \leq \|\frac{1}{n}X^\top U + \frac{1}{n}U^\top X\|_1 \leq \|\frac{1}{n}X^\top U\|_1 + \|\frac{1}{n}U^\top X\|_1. \qquad (4.18)$$

On the one hand, we have

$$\|\frac{1}{n}U^\top X\|_1 = \max_{1\leq j\leq p}\sum_{i=1}^{p}\frac{1}{n}(U^\top X)_{ij} = \max_{1\leq j\leq p}\sum_{i=1}^{p}\sum_{k=1}^{n}\frac{1}{n}U_{ki}X_{kj} =: \max_{1\leq j\leq p} V_j.$$

Let us bound from above the Laplace transform on V_j. We will have

$$\begin{aligned}
E\left[\exp(tV_j)\right] &= E\left[\exp(\frac{t}{n}\sum_{i=1}^{p}\sum_{k=1}^{n}U_{ki}X_{kj})\right] \\
&= \prod_{k=1}^{n} E\left[\exp(\frac{t}{n}\sum_{i=1}^{p}U_{ki}X_{kj})\right] \\
&\leq \prod_{k=1}^{n}\exp\left(\frac{t^2}{2n^2}X_{kj}^2 Var(\sum_{i=1}^{p}U_{ki})\right) \\
&\leq \exp\left(\frac{t^2}{2n}\sum_{i,j=1}^{p}C_U^{ij}\right), \quad \text{since } \frac{1}{n}\sum_{i=1}^{n}X_{kj}^2 \leq 1.
\end{aligned}$$

Moreover, $\sum_{i,j=1}^{p}C_U^{ij} = \sum_i C_U^{ii} + 2\sum_{i<j}C_U^{ij} \leq 3Tr(C_U)$. Indeed, $C_U^{ij} \leq \sqrt{C_U^{ii}C_U^{jj}} \leq \frac{1}{2}(C_U^{ii}+C_U^{jj})$.

We can then deduce that

$$|V_j| \leq t\sqrt{\frac{3}{n}Tr(C_U)}, \quad \text{with probability larger than } 1-2e^{-\frac{t^2}{2}},$$

and then, for $t > 1$,

$$\|\frac{1}{n}U^\top X\|_1 \leq t\sqrt{\frac{6\log(p)}{n}Tr(C_U)} \quad \text{, with probability } 1-2\exp(-(t^2-1)\log(p)). \qquad (4.19)$$

On the other hand,

$$\|\frac{1}{n}X^\top U\|_1 = \max_{1\le j\le p}\sum_{i=1}^{p}\frac{1}{n}(X^\top U)_{ij} = \max_{1\le j\le p}\sum_{i=1}^{p}\sum_{k=1}^{n}\frac{1}{n}X_{ki}U_{kj} =: \max_{1\le j\le p} W_j.$$

The Laplace transform of W_j

$$\begin{aligned}
E\left[\exp(tW_j)\right] &= E\left[\exp(\frac{t}{n}\sum_{i=1}^{p}\sum_{k=1}^{n}X_{ki}U_{kj})\right] \\
&= \prod_{k=1}^{n}E\left[\exp(\frac{t}{n}\sum_{i=1}^{p}X_{ki}U_{kj})\right] \\
&\le \prod_{k=1}^{n}\exp\left(\frac{t^2}{2n^2}(\sum_{i=1}^{p}X_{ki})^2C_U^{jj}\right) \\
&\le \exp\left(\frac{t^2}{2n^2}\sum_{k=1}^{n}(\sum_{i=1}^{p}X_{ki})^2C_U^{jj}\right) \\
&\le \exp\left(\frac{t^2p}{2n}C_U^{jj}\right) \quad \text{since } \frac{1}{n}\sum_{i=1}^{n}X_{kj}^2 \le 1.
\end{aligned}$$

Thus,

$$|W_j| \le t\frac{p}{\sqrt{n}}\sqrt{C_U^{jj}}, \quad \text{with probability larger than } 1-2\exp(-\frac{t^2}{2})$$

and, by a union bound,

$$\|\frac{1}{n}X^\top U\|_1 \le tp\sqrt{2\frac{\log(p)}{n}\max_{1\le j\le p}C_U^{jj}}, \tag{4.20}$$

with probability larger than $1-2\exp(-(t^2-1)\log(p))$. Putting together (4.18), (4.19) and (4.20) we get that

$$\begin{aligned}
\|\frac{1}{n}(X^\top U+U^\top X)\|_2 &\le t\sqrt{\frac{2\log(p)}{n}}(\sqrt{3Tr(C_U)}+p\max_{1\le j\le p}\sqrt{C_U^{jj}}) \\
&\le 3tp\sqrt{\frac{2\log(p)}{n}\max_{1\le j\le p}C_U^{jj}},
\end{aligned}$$

with probability larger than $1-4\exp(-(t^2-1)\log(p))$. This result and *Lemma* 6 conclude the proof.

Secondly, a slightly different bound can be obtained for $\|\frac{1}{n}(X^\top U+U^\top X)\|_2$ using the definition of the norm and the metric entropy of the sphere, as follows. Denote by $W = \frac{1}{n}(X^\top U+U^\top X)$ and recall that

$$\|W\|_2 = \max_{v:\|v\|_2\le 1}|v^\top Wv|.$$

For any point v, there exists a point v_k belonging to an $1/2$-net $\{v_1,...,v_N\}$ such that $\|v-v_k\|_2 \le 1/2$. It is known that $N \le 5^p$, we use $\log(N) \le 2p$. We write for any v such that

$\|v\|_2 \leq 1$:

$$\begin{aligned} |v^\top W v| &\leq 2|v_k^\top W v_k| + 2|(v-v_k)^\top W (v-v_k)| \\ &\leq 2|v_k^\top W v_k| + 2 \max_{u:\|u\|_2 \leq 1/2} |u^\top W u| \\ &\leq 2|v_k^\top W v_k| + \frac{1}{2} \max_{u:\|u\|_2 \leq 1} |u^\top W u|. \end{aligned}$$

We deduce that $\max_{v:\|v\|_2\leq 1} |v^\top W v| \leq 4 \max_{k=1,\dots,N} |v_k^\top W v_k|$. We get

$$\begin{aligned} P(\|W\|_2 \geq 16tp\sqrt{\frac{1}{n}\|C_U\|_2}) &\leq P(4 \max_{k=1,\dots,N} |v_k^\top W v_k| \geq 16t\sqrt{p}\sqrt{\frac{p}{n}\|C_U\|_2}) \\ &\leq \sum_{k=1}^{N} P(\frac{2}{n}|v_k^\top X^\top U v_k| \geq 4t\sqrt{p}\sqrt{\frac{p}{n}\|C_U\|_2}) \\ &\leq 2N \exp(-(2t)^2 p/2) \leq 2\exp(-2pt^2 + 2p) \\ &= 2\exp(-2p(t^2-1)). \end{aligned}$$

□

Lemma 10. *Let us recall the definition $R_{n,p}$ in (4.9). We have*

$$|\langle K(\beta^\star), \delta \rangle| \leq \left| \frac{1}{n} \varepsilon^\top X \delta \right| + R_{n,p}(C_U, \beta^\star, t) \cdot \|\delta\|_2, \tag{4.21}$$

with probability larger than $1 - \frac{8}{n} - 4\exp(-\frac{t^2}{2}) - 4\exp(-2p(t^2-1))$.

Proof of Lemma 10. We know that

$$|\langle K(\beta^\star), \delta \rangle| \leq \left| \left\langle \left[\left(\widetilde{\Sigma} - \hat{\Sigma} \right) \beta^\star + \left(\frac{1}{n} U^\top U - C_U \right) \beta^\star - \frac{1}{n} X^\top \epsilon - \frac{1}{n} U^\top \epsilon + \frac{1}{n} X^\top U \beta^\star \right], \delta \right\rangle \right|.$$

Given Lemma 9 and the definition (4.10) of $A_{n,p}$, we have

$$\left| \delta^\top \left(\widetilde{\Sigma} - \hat{\Sigma} \right) \beta^\star \right| \leq \left(A_{n,p} + 16tp\sqrt{\frac{1}{n}\|C_U\|_2} \right) \|\delta\|_2 \|\beta^\star\|_2 \quad (i)$$

with probability at least $1 - \frac{4}{n} - 4\exp(-2p(t^2-1))$. Furthermore, from Lemma 6, we have

$$\left| \delta^\top \left(\frac{1}{n} U^\top U - C_U \right) \beta^\star \right| \leq \|\delta\|_2 \|\frac{1}{n} U^\top U - C_u\|_2 \|\beta^\star\|_2 \leq A_{n,p} \|\delta\|_2 \|\beta^\star\|_2 \qquad (ii)$$

with probability at least $1 - \frac{4}{n}$. Moreover, from Lemma 7 we will have

$$\left| \frac{1}{n} \delta^\top U^\top \epsilon \right| \leq t^2 \sqrt{\frac{\|C_U\|_2}{n}} \|\delta\|_2 \quad (iii)$$

with probability at least $1 - 2\exp(-\frac{t^2}{2})$. Finally, from Lemma 8 we have

$$\left| \frac{1}{n} \delta^\top X^\top U \beta^\star \right| \leq t\sqrt{\frac{p}{n}\|C_U\|_2} \|\delta\|_2 \|\beta^\star\|_2 \quad (iv)$$

with probability at least $1-2\exp(-\frac{t^2}{2})$. Given the results $(i)-(iv)$, we can conclude that

$$|\langle K(\beta^\star),\delta\rangle| \le \left|\frac{1}{n}\varepsilon^\top X\delta\right| + R_{n,p}(C_U,\beta^\star,t)\cdot\|\delta\|_2$$

with probability larger than $1-\frac{8}{n}-4\exp(-\frac{t^2}{2})-4\exp(-2p(t^2-1))$. □

4.6.3 Additional Proofs

Proof of Lemma 1. By definition of $\widetilde{\Sigma}$, for any symmetric, positive definite matrix M, we know that

$$\|\widetilde{\Sigma}-M\|_F \le \|\widetilde{\Sigma}-\hat{\Sigma}\|_F + \|M-\hat{\Sigma}\|_F \le 2\|M-\hat{\Sigma}\|_F.$$

The matrix $\widetilde{\Sigma}$ can be rewritten as $\widetilde{\Sigma}=\Sigma_{i=1}^p \hat{\eta}_+^i v_i v_i^\top$ where η_i is the i-th eigenvalue of the matrix $\hat{\Sigma}$ and $(v_i)_{i=1,\cdots,p}$ are its eigenvectors. Similarly, $\hat{\Sigma}=\Sigma_{i=1}^p \hat{\eta}^i v_i v_i^\top$.

For convenience, let us assume that $\hat{\eta}_1 \ge \hat{\eta}_2 \ge \cdots \ge \hat{\eta}_p$. Furthermore, we will also assume that $\hat{\eta}_p < 0$, otherwise $\widetilde{\Sigma}=\hat{\Sigma}$ and the result is trivial. We would like to show that $\widetilde{\Sigma}$ is also a solution of $\arg\min_{M\in\mathcal{S}_{\succeq 0}}\|\hat{\Sigma}-M\|_2$. We know that $\hat{\Sigma}=\Sigma_{i=1}^p \eta^i v_i v_i^\top$ then $v_i^\top\hat{\Sigma}v_i = v_i^\top\left(\Sigma_{j=1}^p\hat{\eta}_j v_j v_j^\top\right)v_i = \Sigma_{j=1}^p\hat{\eta}_j v_i^\top v_j v_j^\top v_i = \hat{\eta}_i$.
Let Σ be any positive semidefinite matrix, we will have

$$\begin{aligned}\|\Sigma-\hat{\Sigma}\|_2 &= \sup_{\|v\|_2=1} v^\top(\Sigma-\hat{\Sigma})v \ge v_p^\top(\Sigma-\hat{\Sigma})v_p \\ &= v_p^\top\Sigma v_p - v_p^\top\hat{\Sigma}v_p \ge -\hat{\eta}_p\end{aligned}$$

If we define $\widetilde{\Sigma}=\Sigma_{i=1}^p\max\{\hat{\eta}_i,0\}v_i v_i^\top$, we will have $\|\widetilde{\Sigma}-\hat{\Sigma}\|_2=-\hat{\eta}_p$. Hence,

$$\widetilde{\Sigma}\in\arg\min_{\Sigma\in\mathcal{S}_{\succeq 0}}\|\Sigma-\hat{\Sigma}\|_2.$$

□

Proof of Lemma 2. Let us consider $\delta^\top\widetilde{\Sigma}\delta$. We write that

$$\begin{aligned}\delta^\top\widetilde{\Sigma}\delta &= \delta^\top\widetilde{\Sigma}\delta - \delta^\top\hat{\Sigma}\delta + \delta^\top\hat{\Sigma}\delta \\ &= \delta^\top\left(\widetilde{\Sigma}-\hat{\Sigma}\right)\delta + \delta^\top\left(\frac{1}{n}W^\top W - C_u\right)\delta \\ &= \delta^\top\left(\widetilde{\Sigma}-\hat{\Sigma}\right)\delta + \delta^\top\left(\frac{1}{n}(X+U)^\top(X+U) - C_u\right)\delta \\ &= \delta^\top\left(\widetilde{\Sigma}-\hat{\Sigma}\right)\delta + \frac{1}{n}\delta^\top X^\top X\delta + \frac{1}{n}\delta^\top\left(X^\top U + XU^\top\right)\delta + \delta^\top\left(\frac{1}{n}U^\top U - C_u\right)\delta.\end{aligned}$$

By the definition given in (4.4), we know that $\delta^\top\left(\widetilde{\Sigma}-\hat{\Sigma}\right)\delta \geq 0$, hence using *Lemma* 6, *Lemma* 8 and the *assumption* (4.11) we have

$$\begin{aligned}
\delta^\top\widetilde{\Sigma}\delta &\geq \frac{1}{n}\|X\delta\|_2^2 - \left|\frac{2}{n}\delta^\top X^\top U\delta\right| - \|\delta\|_2^2\|\frac{1}{n}U^\top U - C_u\|_2 \\
&\geq \kappa\|\delta\|_2^2 - 2t\|\delta\|_2^2\sqrt{\frac{p}{n}\|C_U\|_2} - \|\delta\|_2^2\cdot A_{n,p} \\
&\geq \|\delta\|_2^2\left(\kappa - 2t\sqrt{\frac{p}{n}\|C_U\|_2} - A_{n,p}\right) \\
&\geq \|\delta\|_2^2\frac{\kappa}{2}
\end{aligned}$$

with probability at least $1-\frac{4}{n}-2\exp(-\frac{t^2}{2})$. □

Proof of Theorem 4.1. Let us consider the function $f:\beta\to\beta^\top\widetilde{\Sigma}\beta-2\gamma^\top\beta$, where $\gamma := \frac{1}{n}W^\top Y$, and denote by

$$K(\beta) := \nabla f(\beta) = \widetilde{\Sigma}\beta - \frac{1}{n}W^\top Y. \tag{4.22}$$

The function f is convex, hence we will have

$$f(\hat{\beta})-f(\beta^\star)\geq K(\beta^\star)\cdot(\hat{\beta}-\beta^\star)$$

that can also be written as

$$\left(\hat{\beta}^\top\widetilde{\Sigma}\hat{\beta}-2\gamma^\top\hat{\beta}\right)-\left(\beta^{\star\top}\widetilde{\Sigma}\beta^\star-2\gamma^\top\beta^\star\right)\geq\langle K(\beta^\star),\delta\rangle \quad \text{with } \delta=\hat{\beta}-\beta^\star. \tag{4.23}$$

The estimator $\hat{\beta}$ is solution of the optimization problem (4.7):

$$\hat{\beta}^\top\widetilde{\Sigma}\hat{\beta}-2\gamma^\top\hat{\beta}+\|\hat{\beta}\|_\star\leq\beta^{\star\top}\widetilde{\Sigma}\beta^\star-2\gamma^\top\beta^\star+\|\beta^\star\|_\star,$$

then

$$\left(\hat{\beta}^\top\widetilde{\Sigma}\hat{\beta}-2\gamma^\top\hat{\beta}\right)-\left(\beta^{\star\top}\widetilde{\Sigma}\beta^\star-2\gamma^\top\beta^\star\right)\leq\|\beta^\star\|_\star-\|\hat{\beta}\|_\star.$$

We follow here the lines of proof in [30]. We recall that the sorted l_1 norm can be written as

$$\|v\|_\star=\max\sum_{j=1}^{p}\lambda_j|v_{\phi(j)}|$$

where the maximum is taken over all permutations $\phi=(\phi(1),\cdots,\phi(p))$ of $\{1,\cdots,p\}$.

Let us choose the permutation ψ such that

$$\|\beta^\star\|_\star=\sum_{j=1}^{s}\lambda_j|\beta^\star_{\psi(j)}| \text{ and } |\delta_{\psi(s+1)}|\geq|\delta_{\psi(s+s)}|\geq\cdots\geq|\delta_{\psi(p)}|.$$

Thus

$$\begin{aligned}
\|\beta^\star\|_\star-\|\hat{\beta}\|_\star &\leq \sum_{j=1}^{s}\lambda_j\left(|\beta^\star_{\psi(j)}|-|\hat{\beta}_{\psi(j)}|\right)-\sum_{j=s+1}^{p}\lambda_j|\hat{\beta}_{\psi(j)}| \\
&\leq \sum_{j=1}^{s}\lambda_j\left(|\beta^\star_{\psi(j)}|-|\hat{\beta}_{\psi(j)}|\right)-\sum_{j=s+1}^{p}\lambda_j\left(|\beta^\star_{\psi(j)}-\hat{\beta}_{\psi(j)}|\right).
\end{aligned} \tag{4.24}$$

Hence, given the permutation mentioned above, we can then write

$$\|\beta^\star\|_\star - \|\hat{\beta}\|_\star \leq \sum_{j=1}^{s} \lambda_j \left(|\beta^\star - \hat{\beta}|_{(j)}\right) - \sum_{j=s+1}^{p} \lambda_j \left(|\beta^\star - \hat{\beta}|_{(j)}\right) \tag{4.25}$$

Putting together (4.23)-(4.25), we can conclude that

$$\sum_{j=s+1}^{p} \lambda_j |\delta|_{(j)} \leq \sum_{j=1}^{s} \lambda_j |\delta|_{(j)} - |\langle K(\beta^\star), \delta \rangle|. \tag{4.26}$$

From now on, the calculations are specific to our model and estimator. Given Lemma (10), we can conclude that

$$\sum_{j=s+1}^{p} \lambda_j |\delta|_{(j)} \leq \sum_{j=1}^{s} \lambda_j |\delta|_{(j)} + \left|\frac{1}{n}\epsilon^\top X\delta\right| + \|\delta\|_2 R_{n,p}(C_U, \beta^\star, t),$$

then, by Lemma 5,

$$\|\delta\|_\star \leq 2\sum_{j=1}^{s} \lambda_j |\delta|_{(j)} + \max\{H(\delta), G(\delta)\} + \|\delta\|_2 R_{n,p}(C_U, \beta^\star, t). \tag{4.27}$$

From now on, we keep the notation λ_j for the case where $R_{n,p} = O(1)/\sqrt{n}$ and use $\tilde{\lambda}_j$ otherwise (see Case III below).

Case I. $\max\{H(\delta) + \|\delta\|_2 \cdot R_{n,p}, G(\delta)\} \leq H(\delta) + \sigma\|\delta\|_2\sqrt{\sum_{j=1}^{s}\lambda_j^2}$

From (4.27) and (4.14), we infer that

$$\begin{aligned}
\|\delta\|_\star &\leq 2\sum_{j=1}^{s} \lambda_j |\delta|_{(j)} + \left|\frac{1}{n}\epsilon^\top X\delta\right| + R_{n,p}\|\delta\|_2 \\
&\leq 2\sum_{j=1}^{s} \lambda_j |\delta|_{(j)} + H(\delta) + 2\sigma\|\delta\|_2\sqrt{\sum_{j=1}^{s}\lambda_j^2} \\
&\leq 2\sum_{j=1}^{s} \lambda_j |\delta|_{(j)} + 2\sigma\|\delta\|_2\sqrt{\sum_{j=1}^{s}\lambda_j^2} + \sigma\frac{4+\sqrt{2}}{\gamma'}\|\delta\|_\star \\
&\leq 2(1+\sigma)\|\delta\|_2\sqrt{\sum_{j=1}^{s}\lambda_j^2} + \frac{1}{2}\|\delta\|_\star
\end{aligned}$$

hence,

$$\|\delta\|_\star \leq 4(1+\sigma)\|\delta\|_2\sqrt{\sum_{j=1}^{s}\lambda_j^2}$$

We can then conclude that $\delta \in C_{WRE}(c_0)$ with $c_0 = 4(1+\sigma)$.

By the definition of $\hat{\beta}$ as the solution of the optimization problem (4.7), there exists w such that

$$\nabla f(\beta) + w = 0.$$

We also know that

$$\begin{aligned}\delta^\top\widetilde{\Sigma}\delta &= (\hat{\beta}-\beta^\star)^\top\widetilde{\Sigma}(\hat{\beta}-\beta^\star)\\ &= \langle\widetilde{\Sigma}\hat{\beta}-\frac{1}{n}Y^\top W,\delta\rangle-\langle\widetilde{\Sigma}\beta^\star,\delta\rangle+\langle\frac{1}{n}Y^\top W,\delta\rangle\\ &= \langle K(\hat{\beta}),\delta\rangle-\langle K(\beta^\star),\delta\rangle\\ &= -\langle w,\delta\rangle+\left|\langle K(\beta^\star),\delta\rangle\right|\\ &\le \|\delta\|_\star+\left|\langle K(\beta^\star),\delta\rangle\right|, \quad \text{since } \|w\|_{dual}\le 1. \end{aligned} \tag{4.28}$$

We will then have, with probability at least $1-\frac{8}{n}-4\exp(-\frac{t^2}{2})-4\exp(-2p(t^2-1))-\frac{\delta_0}{2}$,

$$\begin{aligned}\delta^\top\widetilde{\Sigma}\delta &\le \|\delta\|_\star+\left|\langle K(\beta^\star),\delta\rangle\right|\\ &\le \|\delta\|_\star+\left|\frac{1}{n}\epsilon^\top X\delta\right|+R_{n,p}\|\delta\|_2\le\|\delta\|_\star+H(\delta)+2\sigma\|\delta\|_2\sqrt{\sum_{j=1}^s\lambda_j^2}.\end{aligned}$$

Let us recall at this point that $H(\delta)=(4+\sqrt{2})\sum_{j=1}^p|\delta|_{(j)}\sigma\sqrt{\frac{\log(2p/j)}{n}}\le\|\delta\|_*/2$. This gives

$$\begin{aligned}\delta^\top\widetilde{\Sigma}\delta &\le \frac{3}{2}\|\delta\|_\star+2\sigma\|\delta\|_2\sqrt{\sum_{j=1}^s\lambda_j^2}\le(\frac{3}{2}+2\sigma)\|\delta\|_2\sqrt{\sum_{j=1}^s\lambda_j^2}\\ &\le \frac{1}{2}(3+4\sigma)\|\delta\|_2\sqrt{\sum_{j=1}^s\lambda_j^2}.\end{aligned} \tag{4.29}$$

Given the WRE condition (4.12)

$$\delta^\top\widetilde{\Sigma}\delta\ge\frac{\kappa}{2}\|\delta\|_2^2$$

and using Lemma 4, we have

$$\begin{aligned}\frac{\kappa}{2}\|\delta\|_2 &\le \frac{1}{2}(3+4\sigma)\sqrt{\sum_{j=1}^s\lambda_j^2}\\ &\le \frac{1}{2}(3+4\sigma)\gamma\sqrt{\frac{s}{n}\log(\frac{2ep}{s})}\end{aligned}$$

hence,

$$\|\delta\|_2\le\frac{(3+4\sigma)\gamma}{\kappa}\sqrt{\frac{s}{n}\log(\frac{2ep}{s})}.$$

Case II. $\max\left\{H(\delta)+\|\delta\|_2\cdot R_{n,p},H(\delta)+\sigma\|\delta\|_2\sqrt{\sum_{j=1}^s\lambda_j^2}\right\}\le G(\delta)$

From the assumption of this case, we infer that

$$(4+\sqrt{2})\sum_{j=1}^p|\delta|_{(j)}\sigma\sqrt{\frac{\log(2p/j)}{n}}+\sigma\|\delta\|_2\sqrt{\sum_{j=1}^s\lambda_j^2}\le\sigma(4+\sqrt{2})\sqrt{\frac{\log(1/\delta_0)}{n}}\|X\delta\|_n$$

then

$$(4+\sqrt{2})\frac{\sigma}{\gamma}\|\delta\|_\star + \sigma\|\delta\|_2\sqrt{\sum_{j=1}^{s}\lambda_j^2} \le \sigma(4+\sqrt{2})\sqrt{\frac{\log(1/\delta_0)}{n}}\|X\delta\|_n.$$

We can imply from the above that both

$$\begin{aligned}(4+\sqrt{2})\frac{\sigma}{\gamma}\|\delta\|_\star &\le (4+\sqrt{2})\sigma\sqrt{\frac{\log(1/\delta_0)}{n}}\|X\delta\|_n \quad \text{and}\\ \sigma\|\delta\|_2\sqrt{\sum_{j=1}^{s}\lambda_j^2} &\le (4+\sqrt{2})\sigma\sqrt{\frac{\log(1/\delta_0)}{n}}\|X\delta\|_n,\end{aligned}$$

then,

$$\|\delta\|_\star \le \gamma\sqrt{\frac{\log(1/\delta_0)}{n}}\|X\delta\|_n \quad \text{and} \quad \|\delta\|_2\sqrt{\sum_{j=1}^{s}\lambda_j^2} \le (4+\sqrt{2})\sqrt{\frac{\log(1/\delta_0)}{n}}\|X\delta\|_n.$$

We know that, with probability at least $1-\frac{8}{n}-4\exp(-\frac{t^2}{2})-4\exp(-2p(t^2-1))-\frac{\delta_0}{2}$,

$$\begin{aligned}\delta^\top\widetilde{\Sigma}\delta &\le \|\delta\|_\star + R_{n,p}\cdot\|\delta\|_2 + \left|\frac{1}{n}\epsilon^\top X\delta\right| \le \|\delta\|_\star + 2G(\delta)\\ &\le \gamma\sqrt{\frac{\log(1/\delta_0)}{n}}\|X\delta\|_n + 2(4+\sqrt{2})\sigma\sqrt{\frac{\log(1/\delta_0)}{n}}\|X\delta\|_n\\ &\le 2\gamma\sqrt{\frac{\log(1/\delta_0)}{n}}\|X\delta\|_n \qquad (4.30)\end{aligned}$$

We also know that, using Lemma 6 and Lemma 8

$$\begin{aligned}\delta^\top\widetilde{\Sigma}\delta &= \delta^\top\left[\left(\widetilde{\Sigma}-\hat{\Sigma}\right)+\frac{1}{n}X^\top X+\left(\frac{1}{n}U^\top U-C_U\right)+\frac{1}{n}\left(X^\top U+U^\top X\right)\right]\delta\\ &\ge \|X\delta\|_n^2 + \langle\left(\frac{1}{n}U^\top U-C_U\right)\delta,\delta\rangle + \frac{1}{n}\langle\left(X^\top U+U^\top X\right)\delta,\delta\rangle \quad \text{since} \quad \delta^\top\left(\widetilde{\Sigma}-\hat{\Sigma}\right)\delta\ge 0\\ &\ge \|X\delta\|_n^2 - \left|\langle\left(\frac{1}{n}U^\top U-C_U\right)\delta,\delta\rangle + \frac{1}{n}\langle X^\top U+U^\top X,\delta\rangle\right|\\ &\ge \|X\delta\|_n^2 - \|\delta\|_2^2\left(A_{n,p}+2t\sqrt{\frac{p}{n}\|C_U\|_2}\right)\end{aligned}$$

with probability larger than $1-\frac{4}{n}-2\exp(-\frac{t^2}{2})$. Hence,

$$\begin{aligned}\|X\delta\|_n^2 &\le \delta^\top\widetilde{\Sigma}\delta + \|\delta\|_2^2\left(A_{n,p}+2t\sqrt{\frac{p}{n}\|C_U\|_2}\right)\\ &\le \delta^\top\widetilde{\Sigma}\delta + \frac{(4+\sqrt{2})^2}{\sum_{j=1}^{s}\lambda_j^2}\frac{\log(1/\delta_0)}{n}\left(A_{n,p}+2t\sqrt{\frac{p}{n}\|C_U\|_2}\right)\|X\delta\|_n^2\\ &\le \delta^\top\widetilde{\Sigma}\delta + \frac{(4+\sqrt{2})}{4\frac{s}{n}\log(\frac{2p}{s})}\frac{\log(1/\delta_0)}{n}\left(A_{n,p}+2t\sqrt{\frac{p}{n}\|C_U\|_2}\right)\|X\delta\|_n^2. \qquad (4.31)\end{aligned}$$

Recall that we assumed

$$A_{n,p}+2t\sqrt{\frac{p}{n}\|C_U\|_2} \le \frac{\kappa}{2}$$

when n, s, p are large enough. Hence, we deduce that for such large enough values of n, s, p

$$\frac{(4+\sqrt{2})}{4\frac{s}{n}\log(\frac{2p}{s})}\frac{\log(1/\delta_0)}{n}\left(A_{n,p}+2t\sqrt{\frac{Tr(C_U)}{n}}\right)\leq\frac{1}{2}.$$

Together with (4.31) this gives that

$$\frac{1}{2}\|X\delta\|_n^2\leq\delta^\top\widetilde{\Sigma}\delta,$$

and using (4.30) we get

$$\|X\delta\|_n\leq 4\gamma\sqrt{\frac{\log(1/\delta_0)}{n}}.$$

Furthermore, this leads to

$$\|\delta\|_\star\leq 4(\gamma)^2\frac{\log(1/\delta_0)}{n}$$

and to

$$\|\delta\|_2\leq\frac{(4+\sqrt{2})\sigma}{\sqrt{\sum_{j=1}^s\lambda_j^2}}4\gamma\frac{\log(1/\delta_0)}{n}\leq\frac{4(4+\sqrt{2})\sigma}{\sqrt{sn\log(2p/s)}}\log(1/\delta_0).$$

Case III. $\max\left\{G(\delta), H(\delta)+\sigma\|\delta\|_2\sqrt{\sum_{j=1}^s\lambda_j^2}\right\}\leq H(\delta)+\|\delta\|_2\cdot R_{n,p}$

Given (4.27), we deduce that

$$\begin{aligned}\|\delta\|_\star&\leq 2\sum_{j=1}^s\tilde{\lambda}_j|\delta|_{(j)}+H(\delta)+2R_{n,p}\cdot\|\delta\|_2\\&\leq 2\|\delta\|_2\sqrt{\sum_{j=1}^s\tilde{\lambda}_j^2}+\frac{1}{2}\|\delta\|_\star+2R_{n,p}\|\delta\|_2\leq\frac{1}{2}\|\delta\|_\star+4\|\delta\|_2\sqrt{\sum_{j=1}^s\tilde{\lambda}_j^2}\end{aligned}$$

giving

$$\|\delta\|_\star\leq 8\|\delta\|_2\sqrt{\sum_{j=1}^s\tilde{\lambda}_j^2}\leq 8\|\delta\|_2R_{n,p}\sqrt{s\log(2pe/s)}.$$

As already seen in the previous two cases, we write that with probability at least $1-\frac{8}{n}-4\exp(-\frac{t^2}{2})-4\exp(-2p(t^2-1))-\frac{\delta_0}{2}$

$$\begin{aligned}\delta^\top\widetilde{\Sigma}\delta&\leq\|\delta\|_\star+H(\delta)+2R_{n,p}\|\delta\|_2\\&\leq\frac{3}{2}\|\delta\|_\star+2\|\delta\|_2R_{n,p}\sqrt{s\log(2pe/s)}\\&\leq 14R_{n,p}\|\delta\|_2\sqrt{s\log(2pe/s)},\end{aligned}$$

hence,

$$\|\delta\|_2\leq\frac{28}{\kappa}R_{n,p}\sqrt{s\log(2pe/s)}.$$

□

Manuscripts

1. Cabral Amilcar Chanang Tondji, *Inference of a non-parametric covariate adjusted variable importance measure of a continuous exposure*, *https://hal.archives-ouvertes.fr/-hal-01336324v2*, June 2017

2. Cabral Amilcar Chanang Tondji, *Linear regression with functional coefficients and errors-in-variables*, Statistica Neerlandica, *In revision*

3. Cabral Amilcar Chanang Tondji, *Slope for high dimensional linear models with measurement errors*, 2020

Bibliography

[1] N.S. Altman. An introduction to kernel and nearest-neighbor nonparametric regression. *The Amerian Statistician*, 46(3):175–185, 1991.

[2] J. Andrews, W. Kennette, J. Pilon, A. Hodgson, A. B. Tuck, A.F. Chambers, and D. I. Rodenhiser. Multi-platform whole-genome microarray analyse refine the epigenetic signature of breast cancer metastasis with gene expresion and copy number. *PLos ONE*, 5(1):e8665, 2010.

[3] J.M. Begun, W.J. Hall, W.M. Huang, and J.A. Wellner. Information and asymptotic efficiency in parametric-nonparametric models. *The Annals of Statistics*, 11(2):432–452, 1983.

[4] P.C. Bellec, G. Lecue, and A.B. Tsybakov. Slope meets lasso: improved oracle bounds and optimality. *The Annals of Statistics*, 46:3603–3642, 2018.

[5] A. Belloni, V. Chernozhukov, and L. Wang. Pivotal estimation via square-root lasso in nonparametric regression. *The Annals of Statistics*, 42(2):757–788, 2014.

[6] A. Belloni, M. Rosenbaum, and A. B. Tsybakov. An $\{\ell_1, \ell_2, \ell_\infty\}$-regularization approach to high-dimensional errors-in-variables models. *Electron. J. Stat.*, 10(2):1729–1750, 2016.

[7] A. Belloni, M. Rosenbaum, and A. B. Tsybakov. Linear and conic programming estimators in high dimensional errors-in-variables models. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 79(3):939–956, 2017.

[8] P. J. Bickel, Y. Ritov, and A. B. Tsybakov. Simultaneous analysis of lasso and dantzig selector. *The Annals of Statistics*, 37(4):1705–1732, 2009.

[9] P.J. Bickel, C.A.J. Klassen, Y. Ritov, and J.A. Wellner. *Efficient and adaptive estimation for semiparametric models.* Johns Hopkins Series in the Mathematical Sciences., 1993.

[10] M. Birkner, M.J. Van Der Laan, and A. Hubbard. Data adaptive pathway testing. *Technical report, Division of Biostatistics, Universtiy of California, Berkeley*, page www.bepress.com/ucbbiostat/, 2005.

[11] R. Bolado-Lavin, W. Castings, and S. Tarantola. Contribution to the sample mean plot for graphical and numerical sensitivity analysis. *Reliability Engineering and System Safety*, 94(6):1041–1049, 2009.

[12] E. Borgonovo. A new uncertainty importance measure. *Reliability Engineering and System Safety*, 92(6):771–784, 2007.

[13] L. Breiman. Bagging predictors. *Technical Report 421, University of California Berkeley*, 2, 1994.

[14] L. Breiman. Random forests - random features. *Technical Report 567, University of California Berkeley*, 2, 1999.

[15] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[16] F. Bunea and L. Xiao. On the sample covariance matrix estimator of reduced effective rank population matrices, with applications to fpca. *Bernoulli*, 21(2):1200–1230, 2015.

[17] C. Butucea and F. Comte. Adaptive estimation of linear functionals in the convolution model and applications. *Bernoulli*, 15(1):69–98, 2009.

[18] C. Butucea and M.-L. Taupin. New m-estimators in semi-parametric regression with errors in variables. *Annales de l'institut Henri Poincaré - Probabilités et Statistiques*, 44(3):393–421, 2008.

[19] E. Candes and T Tao. The dantzig selector: Statistical estimation when p is much larger than n. *The Annals of Statistics*, 35(6):2313–2351, 2007.

[20] A. Chambaz and P. Neuvial. TMLE.NPVI. *http://CRAN.R-project.org/package=tmle.npvi*, 2015. R package 0.10.0.

[21] A. Chambaz and P. Neuvial. tmle.npvi: targeted, integrative search of associations between DNA copy number and gene expression, accounting for DNA methylation. *Bioinformatics*, 31(18):3054–3056, 2015.

[22] A. Chambaz, P. Neuvial, and M. J. van der Laan. Estimation of a non-parametric variable importance measure of a continuous exposure. *Electronic Journal of Statistics*, 6:1059–1099, 2012.

[23] C.A. Chanang Tondji. Inference of a non-parametric covariate-adjusted variable importance measure of a continuous exposure. *https://hal.archives-ouvertes.fr/hal-01336324v2*, June 2017.

[24] C.A. Chanang Tondji. Linear regression model with functional coefficients and errors-in-variables. *Statistica Neerlandica*, September 2018. submitted.

[25] C.A. Chanang Tondji. Slope for high dimensional linear models with measurement errors. *In progress*, 2020.

[26] K. Chang, C. Hsieh, and C. Lin. Coordinate descent method for large-scale l2-loss linear support vector machines. *Journal of Machine Learning Research*, 9:1369–1398, 2008.

[27] A. Datta and H. Zou. Cocolasso for high-dimensional error-in-variables regression. *The Annals of Statistics*, 45(6):3603–3642, 2017.

[28] I. Daubechies, M. Defrise, and C. De Mol. An iterative thresholding algorithm for linear inverse problems with a sparsity constraint. *Communications on Pure and Applied Mathematics*, 57:1413–1457, 2004.

[29] K. De Brabanter, J De Brabanter, and B. De Moor. Nonparametric derivatives estimation. *ftp://ftp.esat.kuleuven.be/sista/kdebaban/11-117.pdf*.

[30] A. Derumigny. Improved bounds for square-root lasso and square-root slope. *The Annals of Statistics*, 12:741–766, 2018.

[31] T. Dietterich. An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting and randomization. *Machine Learning*, pages 1–22, 1998.

[32] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.

[33] E. Fricke. Capital investment and stock returns: An alternative test of investment frictions. *SSRN Electronics Journal*, December 2010.

[34] J. Friedman, T. Hastie, H. Höfling, and R. Tibshirani. pathwise coordinate optimization. *The Annals of Applied Statistics*, 1(2):302–322, 2007.

[35] W.J. Fu. Penalized regressions: The bridge versus the lasso. *Journal of Computational and Graphical Statistics*, 7(3):397–416, 1998.

[36] C. Giraud. *Introduction to high-dimensional statistics*. CRC Press, 2015.

[37] L. Gyorfi, M. Krzyzak, and H. Walk. *A distribution free theory of nonparametric regression*. Springer, 2002.

[38] P. Hall, H.-G. Muller, and F. Yao. Estimation of functional derivatives. *The Annals of Statistics*, 37(6A):3307–3329, 2009.

[39] H. Hastie, R. Tibshirani, and M Wainwright. *The Lasso and generalizations.* CRC Press, 2015.

[40] T. Hastie, R. Tibshirani, and M. Wainwright. *Statistical Learning with Sparsity - The Lasso and generalizations.* CRC Press, 2015.

[41] R.R. Hocking. The analysis and selection of variables in linear. *Biometrics*, 32(1):1–49, 1976.

[42] T. Homma and A. Saltelli. Importance measures in global sensitivity analysis of nonlinear models. *Reliability Engineering and System Safety*, 52(1):1–17, April 1996.

[43] J. Huang, S. Ma, and C-H. Zhang. Adaptative lasso for sparse high dimensional regression. *Statistica Sinica*, 18:1603–1618, 2008.

[44] J. Koshevnik and B. J. Levit. On a nonparametric analogue of the information matrix. *Teor. Verojatnost. i Primenen*, 21(4):759–774, 1976.

[45] H. Liang and R. Li. Variable selection for partially linear models with measurement errors. 104(485):234–248, 2009.

[46] H. Liu, M. Palatucci, and J. Zhang. Blockwise coordinate descent procedures for multi-task lasso, with applications to neural semantic basis discovery. *Proceedings of the* 26^{th} *annual international conference on Machine Learning, ICML*, pages 649–656, 2009.

[47] X. Liu, L. Wang, and H. Liang. Estimation and variable selection for semiparametric additive partial linear models. *Statistica Sinica*, 21:1225–1248, 2011.

[48] P. O. Loh and M. J. Wainwright. High-dimensional regression with noisy and missing data : provable guarantees with nonconvexity. *The Annals of Statistics*, 40(3):1637–1664, 2012.

[49] R. Louhimo and S. Hautaniemi. CNAmet: an R package for integrating copy number, methylation and expression data. *Bioinformatics*, 27(6):887, 2011.

[50] B. Malgorzata, E. Van Den Berg, C. Sabatti, S. Weijie, and E.J. Candes. Slope - adaptative variable selection via convex optimization. *The Annals of Applied Statistics*, 9(3):1103–1140, 2015.

[51] S. Mallat. *A wavelet tour of signal processing.* Academic Press, 2008.

[52] N. Meinshausen and P. Bülhmann. High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics*, 34(3):1436–1462, 2006.

[53] H. Ozer-Balli and B.E. Sorensen. Interaction effects in econometrics. *Empirical Economics*, 45:583–603, 2013.

[54] J. Pearl. https://www.quantamagazine.org/to-build-truly-intelligent-machines-teach-them-cause-and-effect-20180515/. *quantamagazine*, 2018.

[55] J. Pfanzagl. Contributions to a general asymptotic statistical theory. *Springer Science and Business Media*, 13, 1982.

[56] J.R. Pollack, T. Sorlie, C. A. Rees, S.S. Jeffrey, P.E. Lonnin, R. Tbishirnani, D. Botstein, A.-L. Borresen-Dale, and P.O. Brown. Microarray analysis reveals a major direct role of DNA copy number alteration in the transcriptional program of human breast tumors. *Proc Natl Acad Sci USA*, 99(20):12963–12968, 2002.

[57] V. Ramanathan and G. Carmichael. Global and regional climate changes due to black carbon. *Nature Geoscience*, 1:221–227, 2008.

[58] J. M. Robins, S. D. Mark, and W. K. Newey. Estimating exposue effects by modelling the expectation of exposure conditional on confounders. *Biometrics*, 48(4):479–495, 1992.

[59] J. M. Robins and A. Rotnitzky. Comment on inference for semiparametric models : some questions and an answer by Bickel, P. J. and Kwon, J. *Statistica Sinica*, 11:920–935, 2001.

[60] J.M. Robins, M.A. Hernan, and B. Brumback. Marginal structural models and causal inference in epidemiology. *Epidemiology*, 11(5):550–560, September 2000.

[61] V. Rondonotti, J.S. Marron, and C. Park. Sizer for time series : a new approach to the analysis of trends. *Electronic Journal of Statistics*, 1:268–289, 2007.

[62] M. Rosenbaum and A. Tsybakov. Sparse recovery under matrix uncertainty. *The Annals of Statistics*, 38(5):2620–2651, 2010.

[63] R.J. Samworth. Optimal weighted nearest neighbour classifiers. *The Annals of statistics*, 40(5):2733–2763, 2012.

[64] I.J. Schoenberg. Spline functions and the problem of graduation. *Proceedings of the National Academy of Science of the USA*, 52(4):947–950, 1964.

[65] I.M. Sobol and S. Kucherenko. A new derivative based importance criterion for groups of variables and its link with the global sensitivity indices. *Computer physics communication*, 181(7):1212–1217, 2010.

[66] S. Solomon, G-K. Plattner, R. Knutti, and P. Friedlingstein. Irreversible climate change due to carbon dioxide emissions. *PNAS*, 106(6):1704–1709, 2008.

[67] O. Sorensen, A. Frigessi, and M. Thoresen. Measurement error in lasso : Impact and likelihood bias correction. *Statistica Sinica*, 25:809–829, 2015.

[68] C. Stein. Efficient nonparametric testing and estimation. *In Proceedings of the Third Berkeley, Symposium on Mathematical Statistics and Probability*, 1:1187–195, 1956.

[69] Weijie Su and Emmanuel Candès. SLOPE is adaptive to unknown sparsity and asymptotically minimax. *The Annals of Statistics*, 44(3):1038–1068, 2016.

[70] Z. Sun, Y.W. Asmann, K.R. Kalari, B. Bot, J.E. Eckel-Passow, T.R. Baker, J.M. Carr, I. Khrebtukova, S. Luo, and L. et al. Zhan. Integrated analysis of gene expression, CpG island methylation, and gene copy number in breast cancer cells by deep sequencing. *PLoS One*, 6(2):e17490, 2011.

[71] Z. Tang, Z. Lu, B. Jiang, P. Wang, and F. Zhang. Entropy-based importance measure for uncertainty model inputs. *AIAA Journal*, 51(10):2319–2334, 2013.

[72] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society*, 58:267–288, 1996.

[73] A. Tsybakov. *An Introduction to Nonparametrics Estimation.* Springer - New York, 2009.

[74] S. van de Geer. The deterministic lasso. *American Statistical Association*, 2007.

[75] M. J. van der Laan. Statistical inference for variable importance. *Journal of Biostatistics*, 2(1), 2006.

[76] M. J. van der Laan and D. Rubin. Targeted maximum likelihood learning. *International Journal of Biostatistics.*, 2(1), December 2006. Article 11.

[77] M.J. van der Laan. Statistical inference for variable importance. *International Journal of Biostatistics*, 2, 2006.

[78] M.J. van der Laan and S. Rose. *Targeted Learning : Causal Inference for Observational and Experimental Data.* Springer Series in Statistics. Springer Verlag, 2011.

[79] A. W. van der Vaart. *Asymptotic Statistics.* Cambridge University Press, Cambridge, 1998.

[80] W. N. van Wieringen and M. A. van de Wiel. Non parametric testing for DNA copy number induced differential mRNA gene expression. *Biometrics*, 65(1):19–29, 2008.

[81] P. Wei, Z. Lu, and J. Song. Variable importance analysis: a comprehensive review. *Reliability Engineering and System Safety*, (142):399–432., 2015.

[82] Tong T. Wu and K. Lange. Coordinate descent algorithm for lasso penalized regression. *The Annals of Applied Statistics*, 2(1):224–244, 2008.

[83] P. Zhao and B. Yu. On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563, 2006.

[84] S. Zhou and D.A. Wolfe. On derivative estimation in spline regression. *Statistica Sinica*, 10:93–108, 2000.

[85] H. Zou. The adaptative lasso and its oracle properties. *Journal of the American Statistical Association*, 101(476), December 2006.