Habilitation à Diriger des Recherches :

Contribution to statistical learning of complex data using generative models

Julien JACQUES

Université Lille 1, France & CNRS & Inria

28/11/2012
Statistical learning

**Statistical learning** aims to define and estimate a link between *features* variables $X$ (inputs) and a *response variable* $Y$ (output):

$$Y \in \mathcal{Y} \quad \overset{\text{link}}{\leftarrow} \quad X = (X_1, \ldots, X_p) \in \mathcal{X}.$$ 

**Y can be**

- quantitative (typically $\mathcal{Y} = \mathbb{R}$) $\rightarrow$ regression,
- categorical (typically $\mathcal{Y} = \{g_1, \ldots, g_K\}$),
  if moreover $Y$ is
  - observed $\rightarrow$ (supervised) classification,
  - *un*observed $\rightarrow$ clustering (*unsupervised* classification).
Research domain: Statistical learning

Focus of my work

Statistical learning for complex data using generative models.

Complex data?

<table>
<thead>
<tr>
<th>usual data</th>
<th>complex data</th>
</tr>
</thead>
<tbody>
<tr>
<td>categorical</td>
<td>ranking data</td>
</tr>
<tr>
<td>ordinal data</td>
<td></td>
</tr>
<tr>
<td>high-dimensional data ($p &gt;&gt; n$)</td>
<td></td>
</tr>
<tr>
<td>functional data</td>
<td></td>
</tr>
<tr>
<td>all data</td>
<td>learning pop. $\neq$ prediction pop.</td>
</tr>
<tr>
<td>population evolution</td>
<td>$p \rightarrow \infty$</td>
</tr>
</tbody>
</table>

$G$ structure on $G$
Goal of my research

**Goal**

To provide **statistical learning tools**
- density estimation,
- classification,
- clustering,
- regression,

for each kind of **complex data**.

For this, we proceed as follows:
- define generative probabilistic models (if needed),
- consider mixture models for classification and clustering,
- propose estimation procedures.
## Summary

<table>
<thead>
<tr>
<th>Complex data</th>
<th>model design</th>
<th>classification</th>
<th>clustering</th>
<th>regression</th>
<th>Ph.D. in progress</th>
</tr>
</thead>
<tbody>
<tr>
<td>ranking</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>ordinal</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>high-dimensional functional data from ≠ pop.</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
</tr>
</tbody>
</table>

**Legend:**
- ✓: done
- ✓: can be deduce easily
Plan

1. Ranking data
2. Ordinal data
3. Functional data
4. Future works
1 Ranking data
2 Ordinal data
3 Functional data
4 Future works
A rank datum is a **ranking of** $m$ **objects** by a judge according to a given preference order.

**Example:**
Three holidays destinations have to be ranked:

\[
O_1 = \text{Campaign}, \quad O_2 = \text{Mountain} \quad \text{and} \quad O_3 = \text{Sea}.
\]

A judge can prefer: first Sea, second Campaign, and last Mountain. The corresponding ranking can be quoted by:

\[
x = (3, 1, 2) = (1^{\text{st}} O_3, 2^{\text{nd}} O_1, 3^{\text{th}} O_2).
\]
Existing models for ranking data (1/2)

**Thurstone 1927**
- A note $Z_j$ is associated to each object $O_j$,
- $Z = (Z_1, \ldots, Z_m) \sim \mathcal{N}_m(\xi, \Sigma)$,
- Ranking data = ranking of the $Z_j$'s.

**Luce 1959, Plackett 1975**

Multi-stage models assume

$$p(x) = \prod_{j=1}^{m-1} \frac{v_j}{v_j + v_{j+1} + \ldots + v_m}$$

where $v_j$ the probability that $O_{x_j}$ is the preferred object.
Paired comparison models assume

\[ p(x) \propto \prod_{1 \leq i < j \leq m} p_{ij}, \quad \text{with } p_{ij} \text{ the probability that } O_{x_i} \text{ is preferred to } O_{x_j}. \]

Parsimony + re-parametrisation \( \Rightarrow \) Mallows \( \Phi \) model (\( \sim \)1950):

\[ p(x; \mu, \theta) \propto \exp(-\theta d_K(x, \mu)) \]

- \( \mu = (\mu^1, \ldots, \mu^m) \): reference/central ranking,
- \( \theta \in \mathbb{R}^+ \): dispersion parameter,
- \( d_K \): Kendall distance.
Defining a new model: ISR

What is the generative process of a rank datum?

- Ranking data = result of a sorting algorithm,
- Elementary operation = comparison of paired of objects,
- Rank $x \neq \mu \leftarrow \text{error}$ in paired comparison.

The ISR model (Biernacki & Jacques 2012)

- Sorting algorithm: Insertion Sort algorithm,
- Error in paired comparison $\sim B(1 - \pi),$

$$p(x; \mu, \pi) = \frac{1}{m!} \sum_{y \in \mathcal{P}_m} \pi^{\text{good}(x,y,\mu)} (1 - \pi)^{\text{bad}(x,y,\mu)}$$

Where

- $y$ is the presentation order of the objects,
- $\text{good}(x, y, \mu)$: nb. of good paired comparisons during the sort,
- $\text{bad}(x, y, \mu)$: nb. of bad paired comparisons during the sort.
Defining a new model: ISR

Properties of ISR

- meaningful parameters:
  - $\mu$: central ranking (mode if $\pi > \frac{1}{2}$),
  - $\pi$: dispersion parameter (uniform for $\pi = \frac{1}{2}$),
  - $\mu$ uniformly more pronounced when $\pi$ grows,
- ...
Extension to multivariate partial rankings clustering

Multivariate rank

\[ \mathbf{x} = (x^1, \ldots, x^p): \text{multivariate rank}, \]
with \( x^j = (x^{j1}, \ldots, x^{jm_j}) \) a rank of \( m_j \) objects.

Mixture of ISR for multivariate rankings (Jacques & Biernacki 2012)

- population composed of \( K \) groups (proportions \( p_k \)),
- conditional independence assumption,

\[ \Rightarrow \quad p(\mathbf{x}; \theta) = \sum_{k=1}^{K} p_k \prod_{j=1}^{p} \frac{1}{m_j!} \sum_{y \in \mathcal{P}_{m_j}} p(x^j|y; \mu^j_k, \pi^j_k), \]

with \( \theta = (\pi^j_k, \mu^j_k, p_k)_{k=1,\ldots,K, j=1,\ldots,p} \).

Partial rank

Each \( x^j = (x^{j1}, \ldots, x^{jm_j}) \) can be full or partial.
Maximum likelihood

with missing data:
- presentation orders,
- group memberships,
- missing positions in partial rankings.

SEM-Gibbs algorithm

- SE-Gibbs step: generate missing data thanks to a Gibbs algorithm,
- M step: maximise the completed-data log-likelihood.
The Eurovision Song Contest

- annual competition in which European countries rank ten preferred song,
- voting from 2007 to 2012 are analysed,
- \( n = 34 \) countries voted to each contest between 2007 and 2012,
- \( m = 8 \) countries participated to the six finals:
  France (1), Germany (2), Greece (3), Romania (4), Russia (5), Spain (6), Ukraine (7) and United Kingdom (8),
- since none of the 34 countries ranked all of the 8 finalist countries in its 10 preferences, all rankings are partial.

Mixture of ISR estimation

Thanks to the RankClust package for R, soon available on the CRAN website.
Geographical repartition of the clusters suggests geographical alliances between countries:
- group 1: West European countries,
- group 2: some Northern countries,
- group 3: Mediterranean countries,
- group 4: maybe more dispersed,
- group 5: essentially East European.
Plan

1. Ranking data
2. Ordinal data
3. Functional data
4. Future works
An **ordinal** variable $X$ with $m$ modalities $\{1, \ldots, m\}$ is a **nominal** variable with **full ordered** modalities:

$$1 < \ldots < m$$

**Example: AÉRES evaluation**

- AÉRES: Agence d’Évaluation de la Recherche et de l’Enseignement Supérieur,
- Evaluation of licenses in March 2011,
- Academies in wave A: Bordeaux, Toulouse, Lyon, Montpellier, Grenoble,
- 23 universities evaluated through 4 criteria: Pilot training (PT), Educational project (EP), Support success (SS), Employability and further studies (EFS),
- Each criterion is evaluated by a letter score $\{A+, A, B, C\}$.

<table>
<thead>
<tr>
<th>University</th>
<th>PT</th>
<th>EP</th>
<th>SS</th>
<th>EFS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bordeaux 1</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Bordeaux 2</td>
<td>A+</td>
<td>A</td>
<td>A+</td>
<td>A</td>
</tr>
<tr>
<td>Bordeaux 3</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>Bordeaux 4</td>
<td>B</td>
<td>A</td>
<td>A+</td>
<td>A</td>
</tr>
<tr>
<td>Pau</td>
<td>C</td>
<td>B</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>Toulouse 1</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>Toulouse 2</td>
<td>B</td>
<td>B</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Toulouse 3</td>
<td>A</td>
<td>A</td>
<td>A+</td>
<td>A</td>
</tr>
<tr>
<td>Champollion</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>Lyon 1</td>
<td>A</td>
<td>A+</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>Lyon 2</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>Lyon 3</td>
<td>B</td>
<td>A+</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Existing models for ordinal data

**Standard choices**
- Ordinal data $\sim$ continuous data $\Rightarrow$ artificial distance information,
- Ordinal data $\sim$ nominal data $\Rightarrow$ lost order information.

**Gouget 2006, Jollois & Nadif 2007**
- Extended continuous: latent discretization of a Gaussian.
- Restrained nominal: order constraints on a multinomial.

**D’Elia & Piccolo 2005**
CUB model is defined as a mixture of Binomial + uniform + Dirac $\Rightarrow$ artificial construction to obtain natural properties:
- The distribution decreases with distance from the mode,
- Uniform or Dirac distribution available.
Defining a new model: DSO

**Goal**
To build a new model following the same strategy as for ranking data.

**What is the generative process of an ordinal datum?**
- **Ordinal data** = result of a search algorithm in an ordered list,
- **Elementary operation** = comparison with a current modality,
- **Stochastic distribution** ↔ error in comparison.
Defining a new model: DSO

The DSO model (Biernacki & Jacques 2012)

- **search algorithm**: Dichotomic search, relying on comparison 
  \{<, =, >\},
- **error** in comparison: \( z_j \sim B(1 - \pi) \)

\[
\Rightarrow p(x; \mu, \pi) = \sum_{e_{m-1}, \ldots, e_1} \prod_{j=1}^{m-1} p(e_{j+1} | e_j; \mu, \pi)p(e_1),
\]

with \( p(e_{j+1} | e_j; \mu, \pi) = \sum_{c_j \in e_j} p(e_{j+1} | e_j, c_j; \mu, \pi)p(c_j | e_j), \)

and \( p(e_{j+1} | e_j, c_j; \mu, \pi) = \pi p(e_{j+1} | c_j, e_j, z_j = 1; \mu) + (1 - \pi)p(e_{j+1} | c_j, e_j, z_j = 0), \)

where

- \( e_j \) is the current interval of search,
- \( c_j \) is the cut point in \( e_j \sim \) uniform.
Defining a new model: DSO

Properties of DSO

- meaningful parameters:
  - $\mu$: position parameter (mode if $\pi > 0$),
  - $\pi$: dispersion parameter (uniform if $\pi = 0$ and Dirac if $\pi = 1$),
- illustration of the distribution:

\[
\begin{align*}
\pi &= 0 & \pi &= .1 & \pi &= .2 & \pi &= .5 \\
\mu &= 1 \\
\mu &= 2 \\
\mu &= 3 
\end{align*}
\]
Extension to clustering of multivariate ordinal data

Mixture of multivariate DSO
- clustering $\rightarrow$ mixture model,
- multivariate ordinal data $\rightarrow$ conditional independence assumption.

Estimation
Double nested EM algorithm with two levels of missing variables:
- the cut points $c_j$, the search intervals $e_j$ and the comparison accuracy $z_j$,
- the group memberships.
Application: AÉRES evaluation (wave A)

- AÉRES: Agence d’Évaluation de la Recherche et de l’Enseignement Supérieur,
- Evaluation of licenses in March 2011,
- Academies in wave A: Bordeaux, Toulouse, Lyon, Montpellier, Grenoble,
- 23 universities evaluated through 4 criteria:
  - Pilot training (PT),
  - Educational project (EP),
  - Support success (SS),
  - Employability and further studies (EFS).
- Each criterion is evaluated by a letter score \{A+, A, B, C\}.
## AÉRES evaluation: Data

<table>
<thead>
<tr>
<th>University</th>
<th>PT</th>
<th>EP</th>
<th>SS</th>
<th>EFS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bordeaux 1</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Bordeaux 2</td>
<td>A+</td>
<td>A</td>
<td>A+</td>
<td>A</td>
</tr>
<tr>
<td>Bordeaux 3</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>Bordeaux 4</td>
<td>B</td>
<td>A</td>
<td>A+</td>
<td>A</td>
</tr>
<tr>
<td>Pau</td>
<td>C</td>
<td>B</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>Toulouse 1</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>Toulouse 2</td>
<td>B</td>
<td>B</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Toulouse 3</td>
<td>A</td>
<td>A</td>
<td>A+</td>
<td>A</td>
</tr>
<tr>
<td>Champollion</td>
<td>A</td>
<td>B</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>Lyon 1</td>
<td>A</td>
<td>A+</td>
<td>A</td>
<td>A</td>
</tr>
<tr>
<td>Lyon 2</td>
<td>B</td>
<td>A</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>Lyon 3</td>
<td>B</td>
<td>A+</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>St Etienne</td>
<td>A</td>
<td>B</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Montpellier 1</td>
<td>B</td>
<td>A</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Montpellier 2</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Montpellier 3</td>
<td>B</td>
<td>B</td>
<td>A</td>
<td>B</td>
</tr>
<tr>
<td>Nîmes</td>
<td>C</td>
<td>B</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>Perpignan</td>
<td>B</td>
<td>B</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>Grenoble 1</td>
<td>B</td>
<td>B</td>
<td>A+</td>
<td>A</td>
</tr>
<tr>
<td>Grenoble 2</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>B</td>
</tr>
<tr>
<td>Grenoble 3</td>
<td>C</td>
<td>B</td>
<td>B</td>
<td>C</td>
</tr>
<tr>
<td>Savoie</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>B</td>
</tr>
</tbody>
</table>
AÉRES evaluation: Global analysis

Resume of the whole data set \((g = 1)\)

\[ \hat{\mu} = (B,A,B,B) \quad \text{and} \quad \hat{\pi} = (0.37, 0.39, 0.27, 0.59) \]

- PT and EP have similar dispersion around B and A resp.,
- SS has higher dispersion around B.
- EFS has lower dispersion around B.

<table>
<thead>
<tr>
<th>Model</th>
<th>(g = 1)</th>
<th>(g = 2)</th>
<th>(g = 3)</th>
<th>(g = 4)</th>
<th>(g = 5)</th>
<th>(g = 6)</th>
</tr>
</thead>
<tbody>
<tr>
<td>DSO</td>
<td>-111.90</td>
<td>-109.14</td>
<td>-107.80</td>
<td><strong>-104.25</strong></td>
<td>-108.49</td>
<td>-114.28</td>
</tr>
</tbody>
</table>
Cluster 1: $\hat{\mu}_1 = (A, A, A, B)$ “homogeneous high score”,
Cluster 2: $\hat{\mu}_2 = (B, A, A+, C)$ “contrasted score”,
Cluster 3: $\hat{\mu}_3 = (B, B, B, B)$ “homogeneous middle score”,
Cluster 4: $\hat{\mu}_4 = (C, B, B, C)$ “lower score”.
Plan

1. Ranking data
2. Ordinal data
3. Functional data
4. Future works
We consider functional data

$$X = \{X(t), \ t \in [0, T]\}$$

with values in $L^2([0, T])^p$

- $p = 1$: a sample path of $X$ is a single curve,
- $p > 1$: a path of $X$ is a set of $p$ curves.

Goal

To define model-based clustering for such data.
Main difficulty
Functional data live in an infinite-dimensional space.

Standard choices
1. infinite $\rightarrow$ finite problem
   - time discretization,
   - basis approximation (including Functional PCA).
2. use of classical clustering techniques.

Non-parametric method (Ferraty & Vieu 2006, Ieva et al. 2012...)
k-means or hierarchical clustering with distance between functional data.
Model-based approaches

Modelling of

- basis expansion coefficients:
  James & Sugar 2003, Giacofci et al. 2012...
- principal component scores:
New clustering algorithms for functional data

\( \mathbf{X} = \{ \mathbf{X}(t), \ t \in [0, T] \} \) a \( L_2 \)-cont. stoch. proc. with values in \( L_2([0, T])^p \)

- \( \mu = \mathbb{E}[\mathbf{X}] \) the mean function of \( \mathbf{X} \),
- covariance operator of \( \mathbf{X} = \) integral operator \( \mathbf{C} \) with kernel
  \[
  C(t, s) = \mathbb{E}[(\mathbf{X}(t) - \mu(t)) \otimes (\mathbf{X}(s) - \mu(s))].
  \]

Multivariate Functional Principal Component Analysis (MFPCA) of \( \mathbf{X} \)

\[
\begin{align*}
\mathbf{X}(t) &= \mu(t) + \sum_{j \geq 1} C_j f_j(t),
\end{align*}
\]

- \( f_j \) form an orthonormal basis of eigen-functions (principal factors), solutions of \( \mathbf{C} f_j = \lambda_j f_j \),
- \( C_j \) are zero-mean uncorrelated random variables (principal components) with variance \( \lambda_j \), \( \lambda_1 \geq \lambda_2 \geq \ldots \)

\[
C_j = \int_0^T \langle \mathbf{X}(t) - \mu(t), f_j(t) \rangle_{\mathbb{R}^p} \, dt.
\]
New clustering algorithms for functional data

- \( X = (X_1, ..., X_n) \) be an i.i.d sample of size \( n \) of \( X \),
- for each \( X_i, Z_i = (Z_{i1}, \ldots, Z_{iK}) \in \{0, 1\}^K \) is such that \( Z_{ik} = 1 \) if \( X_i \) belongs to the cluster \( k \),

Jacques & Preda 2012

We assume that \( X \) has the following density approximation,

\[
f_X^{(\mathbf{q})}(x; \theta) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{q_k} f_{C_j|Z_k=1}(c_{jk}(x); \lambda_{jk})
\]

where

- \( C_j \) is assumed to be Gaussian (true if \( X \) is a Gaussian process),
- \( \theta = (\pi_k, \lambda_{1k}, \ldots, \lambda_{q_kk})_{1 \leq k \leq K} \) have to be estimated,
- \( c_{jk}(x) \) have to be computed,
- \( \mathbf{q} = (q_1, \ldots, q_K) \) have to be selected.
New clustering algorithms for functional data

Bouveyron & Jacques 2011

We assume that $X$ has the following density approximation,

$$f_{X}^{(p)}(x; \theta) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{p_k} f_{C_{j|Z_k=1}}(c_{jk}(x); \lambda_{jk})$$

where

- $C_{j}$ is assumed to be Gaussian (true if $X$ is a Gaussian process),
- $\theta = (\pi_k, \lambda_{1k}, \ldots, \lambda_{q_kk})_{1 \leq k \leq K}$ have to be estimated,
- $c_{jk}(x)$ have to be computed,
- $p_k$ is the maximum number of $\lambda_{jk} > 0$,
- with $\lambda_{jk} = b_k$ or $b$ for all $j \geq q_k$ and $\lambda_{jk} = a_{jk}, a_j, a_k$ or $a$ for all $j < q_k$. 

New clustering algorithms for functional data

**Estimation procedure**

For each model we consider an EM-like algorithm, in which the M step consists of:

- computing the **cluster specific principal component** $c_{jk}$ (FPCA with curves weight depending on conditional probabilities computed at the E step),
- selecting the **regularization** parameter ($q_k$),
- maximizing the approximated completed log-likelihood.
Application: Canadian weather data

Temperature and precipitation curves are first normalized

[Graphs showing original and reduced curves for temperature and precipitation over time]
Clustering Canadian Weather data

Clustering visualization (Jacques & Preda 2012)
Plan

1. Ranking data
2. Ordinal data
3. Functional data
4. Future works
And now?

Further exploitation of the proposed models

- ranking and ordinal data:
  - new model with search algo. stopped before the end (ordinal), \( \pi \) non constant (both)...
  - new model for ranking with known presentation order,
  - multivariate data without conditional independence assumption,

- functional data:
  - convergence of the pseudo-EM algorithm,
  - qualitative functional data,

- ...

Dissemination of research results

- \textbf{R} package for each kind of data.
Heterogeneous data

How to work simultaneously with continuous, binary, ranking, ordinal, functional data... ?

- simple solution:
  - conditional independence
    → Mixmod hetero. soon available for conti. + categor.),

- to go further:
  - expert models (Gormley & Murphy 2008, Ng & McLachlan 2010),
  - latent variables (Browne & McNicholas 2012),
  - kernel methods (Bouveyron et al. 2012),
  - ...