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

Research domain: Statistical learning

Statistical learning

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

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

Y can be

- quantitative (typically $\mathcal{Y} = \mathbb{R}$) $\longrightarrow$ regression,
- categorical (typically $\mathcal{Y} = \{g_1, \ldots, g_K\}$), if moreover Y is
  - observed $\longrightarrow$ (supervised) classification,
  - *un*observed $\longrightarrow$ clustering (*un*supervised classification).

Research domain: Statistical learning

Focus of my work

Statistical learning for complex data using generative models.

Complex data?

usual data		complex data
categorical $\mathcal{G}$	$\xrightarrow[\text{on } \mathcal{G}]{\text{structure}}$	ranking data
		ordinal data
continuous $\mathbb{R}^p$	$\xrightarrow{p \to \infty}$	high-dimensional data ($p >> n$)
		functional data
all data	$\xrightarrow[\text{evolution}]{\text{population}}$	learning pop. $\neq$ prediction pop.

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 of my works

Summary

Complex data	model design	classification	clustering	regression
ranking	✓	✓	✓	
ordinal	✓	✓	✓	
high-dimensional				Ph.D. in progress
functional	✓	✓	✓	
data from $\neq$ pop.		✓		✓

Legend:

- ✓: done
- ✓: can be deduce easily

Plan

Plan

Ranking data

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:

$$\mathcal{O}_1 = \text{Campaign}, \mathcal{O}_2 = \text{Mountain and } \mathcal{O}_3 = \text{Sea}.$$

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

$$x = (3, 1, 2) = (\overset{1^{st}}{\mathcal{O}_3}, \overset{2^{nd}}{\mathcal{O}_1}, \overset{3^{th}}{\mathcal{O}_2}).$$

Existing models for ranking data (1/2)

Thurstone 1927

- a note Z_j is associated to each object $\mathcal{O}_j$,
- $Z = (Z_1, \ldots, Z_m) \sim \mathcal{N}_m(\xi, \Sigma)$,
- 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 $\mathcal{O}_{xj}$ is the preferred object.

Existing models for ranking data (2/2)

Kendall & Smith 1940, Mallows 1950

Paired comparison models assume

$$p(x) \propto \prod_{1 \leq i < j \leq m} p_{ij}, \qquad \text{with } p_{ij} \text{ the probability that } \mathcal{O}_{xi} \text{ is preferred to } \mathcal{O}_{xj}.$$

Parsimony + re-parametrisation $\Rightarrow$ Mallows Φ 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 $x \neq \mu \Leftarrow$ error in paired comparison.

The ISR model (Biernacki & Jacques 2012)

- sorting algorithm: Insertion Sort algorithm,
- error in paired comparison $\sim \mathcal{B}(1-\pi)$,

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

where

- y is the presentation order of the objects,
- good(x,y,μ): nb. of good paired comparisons during the sort,
- bad(x,y,μ): nb. of bad paired comparisons during the sort.

Defining a new model: ISR

Properties of ISR

- meaningful parameters:
  - μ: central ranking (mode if $\pi > \frac{1}{2}$),
  - π: dispersion parameter (uniform for $\pi = \frac{1}{2}$),
- μ uniformly more pronounced when π grows,
- ...

Extension to multivariate partial rankings clustering

Multivariate rank

$\boldsymbol{x} = (x^1, \ldots, x^p)$: *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,

$$\Rightarrow \quad p(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^{K} p_k \prod_{j=1}^{p} \frac{1}{m_j!} \sum_{y \in \mathcal{P}_{m_j}} p(x^j | y; \mu_k^j, \pi_k^j),$$

with $\boldsymbol{\theta} = (\pi_k^j, \mu_k^j, p_k)_{k=1,\ldots,K\ ,j=1,\ldots,p}$.

Partial rank

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

Estimation

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.

Application: The Eurovision Song Contest

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:

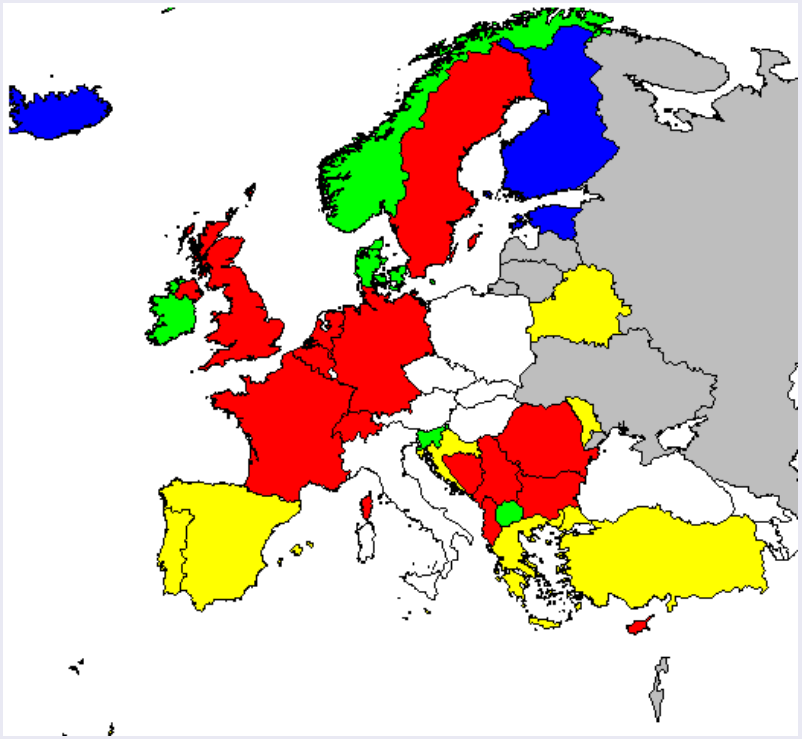
  France (1), Germany (2), Greece (3), Romania (4), Russia (5), Spain (6), Ukraine (7) and United Kingdom (8),
- since none of the 34 countries ranked all of the 8 finalist countries in its 10 preferences, all rankings are partial.

Mixture of ISR estimation

Thanks to the **RankClust** package for **R**, soon available on the CRAN website.

Application: The Eurovision Song Contest

Clustering visualization



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

Ordinal data

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}.

University	PT	EP	SS	EFS
Bordeaux 1	A	A	A	B
Bordeaux 2	A+	A	A+	A
Bordeaux 3	B	A	B	B
Bordeaux 4	B	A	A+	A
Pau	C	B	B	C
Toulouse 1	B	B	B	B
Toulouse 2	B	B	A	B
Toulouse 3	A	A	A+	A
Champollion	A	B	B	B
Lyon 1	A	A+	A	A
Lyon 2	B	A	B	B
Lyon 3	B	A+	B	B
...				

Existing models for ordinal data

Standard choices

- ordinal data $\simeq$ continuous data $\Rightarrow$ artificial distance information,
- ordinal data $\simeq$ 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 define as a mixture of Binomial + uniform + Dirac
$\Rightarrow$ artificial construction to obtain natural properties:

- the distribution decrease 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 $\Leftarrow$ error in comparison.

Defining a new model: DSO

The DSO model (Biernacki & Jacques 2012)

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

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

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

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

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

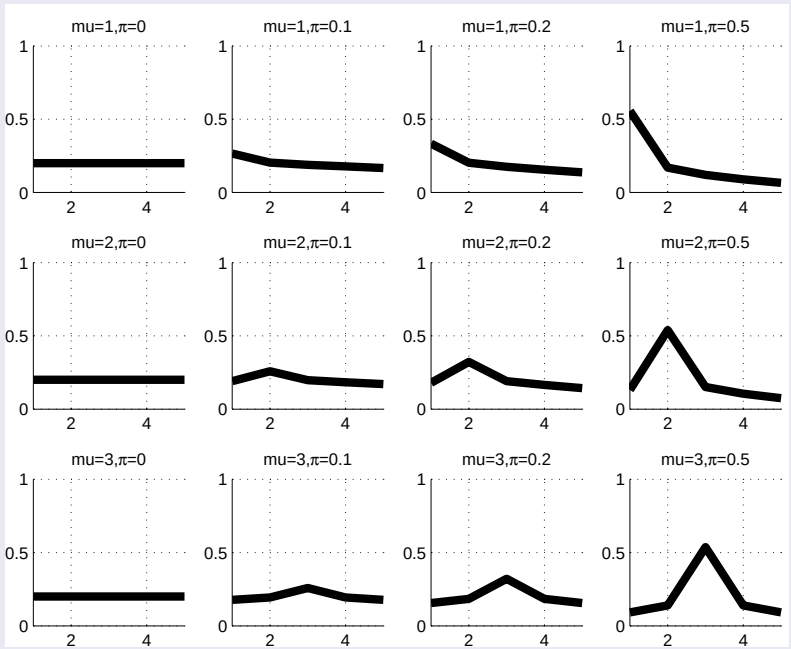
- meaningful parameters:
  - μ: position parameter (mode if $\pi > 0$),
  - π: dispersion parameter (uniform if $\pi = 0$ and Dirac if $\pi = 1$),
- illustration of the distribution:

$\pi = 0$ $\pi = .1$ $\pi = .2$ $\pi = .5$

$\mu = 1$

$\mu = 2$

$\mu = 3$



Extension to clustering of multivariate ordinal data

Mixture of multivariate DSO

- clustering $\longrightarrow$ mixture model,
- multivariate ordinal data $\longrightarrow$ 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

University	PT	EP	SS	EFS
Bordeaux 1	A	A	A	B
Bordeaux 2	A+	A	A+	A
Bordeaux 3	B	A	B	B
Bordeaux 4	B	A	A+	A
Pau	C	B	B	C
Toulouse 1	B	B	B	B
Toulouse 2	B	B	A	B
Toulouse 3	A	A	A+	A
Champollion	A	B	B	B
Lyon 1	A	A+	A	A
Lyon 2	B	A	B	B
Lyon 3	B	A+	B	B
St Etienne	A	B	A	B
Montpellier 1	B	A	A	B
Montpellier 2	A	A	A	B
Montpellier 3	B	B	A	B
Nîmes	C	B	C	C
Perpignan	B	B	B	B
Grenoble 1	B	B	A+	A
Grenoble 2	A	A	B	B
Grenoble 3	C	B	B	C
Savoie	A	A	A	B

AÉRES evaluation: Global analysis

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

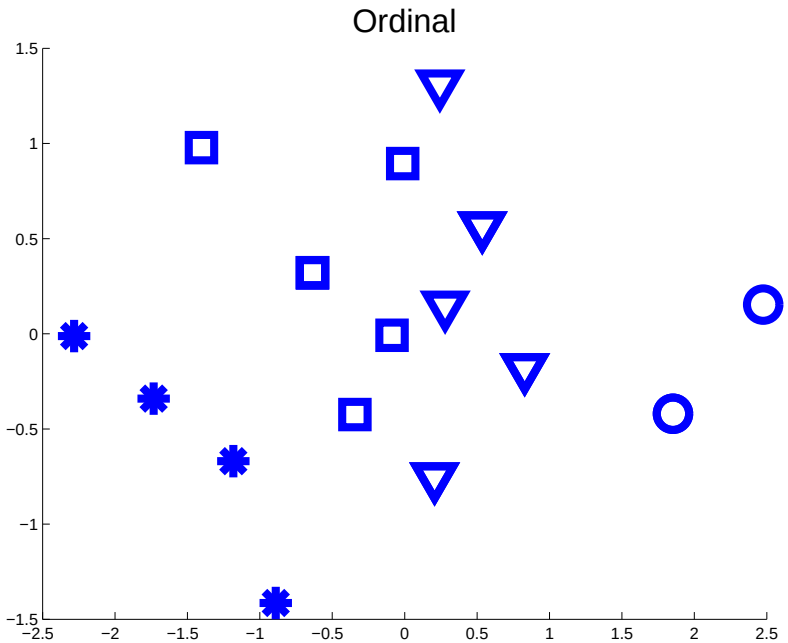
$$\hat{\mu} = (\text{B,A,B,B}) \quad \text{and} \quad \hat{\pi} = (0.37, 0.39, 0.27, 0.59)$$

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

BIC values

Model	$g = 1$	$g = 2$	$g = 3$	$g = 4$	$g = 5$	$g = 6$
DSO	-111.90	-109.14	-107.80	-104.25	-108.49	-114.28

AÉRES evaluation: Analysis for $g = 4$



- 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

Functional data

We consider functional data

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

with values in $L_2([0, T])^p$

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

Goal

To define model-based clustering for such data.

Existing functional data clustering algorithms (1/2)

Main difficulty

Functional data live in a infinite-dimensional space.

Standard choices

1. infinite $\longrightarrow$ 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.

Existing functional data clustering algorithms (2/2)

Model-based approaches

Modelling of

- basis expansion coefficients:
  James & Sugar 2003, Giacofci et al. 2012...
- principal component scores:
  Bouveyron & Jacques 2011, Jacques & Preda 2012.

New clustering algorithms for functional data

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

- $\boldsymbol{\mu} = \mathbb{E}[\boldsymbol{X}]$ the mean function of $\boldsymbol{X}$,
- covariance operator of $\boldsymbol{X}$ = integral operator C with kernel

$$C(t,s) = \mathbb{E}\left[(\boldsymbol{X}(t) - \mu(t)) \otimes (\boldsymbol{X}(s) - \mu(s))\right].$$

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

Saporta 1981

$$\boldsymbol{X}(t) = \boldsymbol{\mu}(t) + \sum_{j \geq 1} C_j \boldsymbol{f}_j(t),$$

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

$$C_j = \int_0^T \langle \mathbf{X}(t) - \mu(t), \mathbf{f}_j(t) \rangle_{\mathbb{R}^p} dt.$$

New clustering algorithms for functional data

- $\underline{\boldsymbol{X}} = (\boldsymbol{X}_1, ..., \boldsymbol{X}_n)$ be an i.i.d sample of size n of $\boldsymbol{X}$,
- for each $\boldsymbol{X}_i$, $\boldsymbol{Z}_i = (Z_{i1}, \ldots, Z_{iK}) \in \{0,1\}^K$ is such that $Z_{ik} = 1$ if $\boldsymbol{X}_i$ belongs to the cluster k,

Jacques & Preda 2012

We assume that $\boldsymbol{X}$ has the following density approximation,

$$f_{\boldsymbol{X}}^{(q)}(\boldsymbol{x};\boldsymbol{\theta}) = \sum_{k=1}^{K} \pi_k \prod_{j=1}^{q_k} f_{C_j|z_k=1}(c_{jk}(\boldsymbol{x});\lambda_{jk})$$

where

- C_j is assumed to be Gaussian (true if $\boldsymbol{X}$ is a Gaussian process),
- $\boldsymbol{\theta} = (\pi_k, \lambda_{1k}, \ldots, \lambda_{q_k k})_{1 \leq k \leq K}$ have to be *estimated*,
- $c_{jk}(\boldsymbol{x})$ have to be *computed*,
- $\boldsymbol{q} = (q_1, \ldots, q_K)$ have to be *selected*.

New clustering algorithms for functional data

Bouveyron & Jacques 2011

We assume that $\boldsymbol{X}$ has the following density approximation,

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

where

- C_j is assumed to be Gaussian (true if $\boldsymbol{X}$ is a Gaussian process),
- $\boldsymbol{\theta} = (\pi_k, \lambda_{1k}, \ldots, \lambda_{q_k k})_{1\leq k\leq K}$ have to be *estimated*,
- $c_{jk}(\boldsymbol{x})$ 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

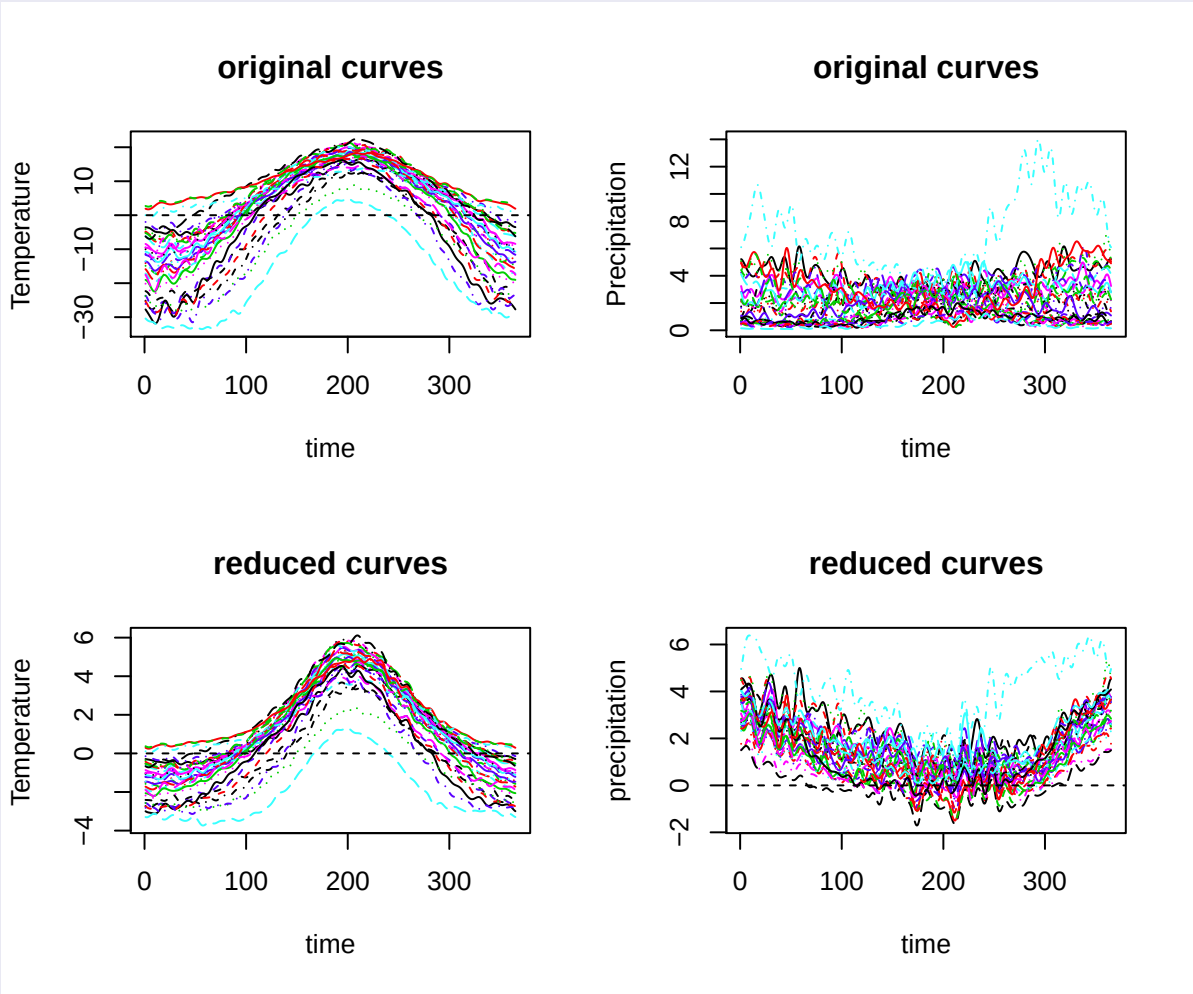
Estimation procedure

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.

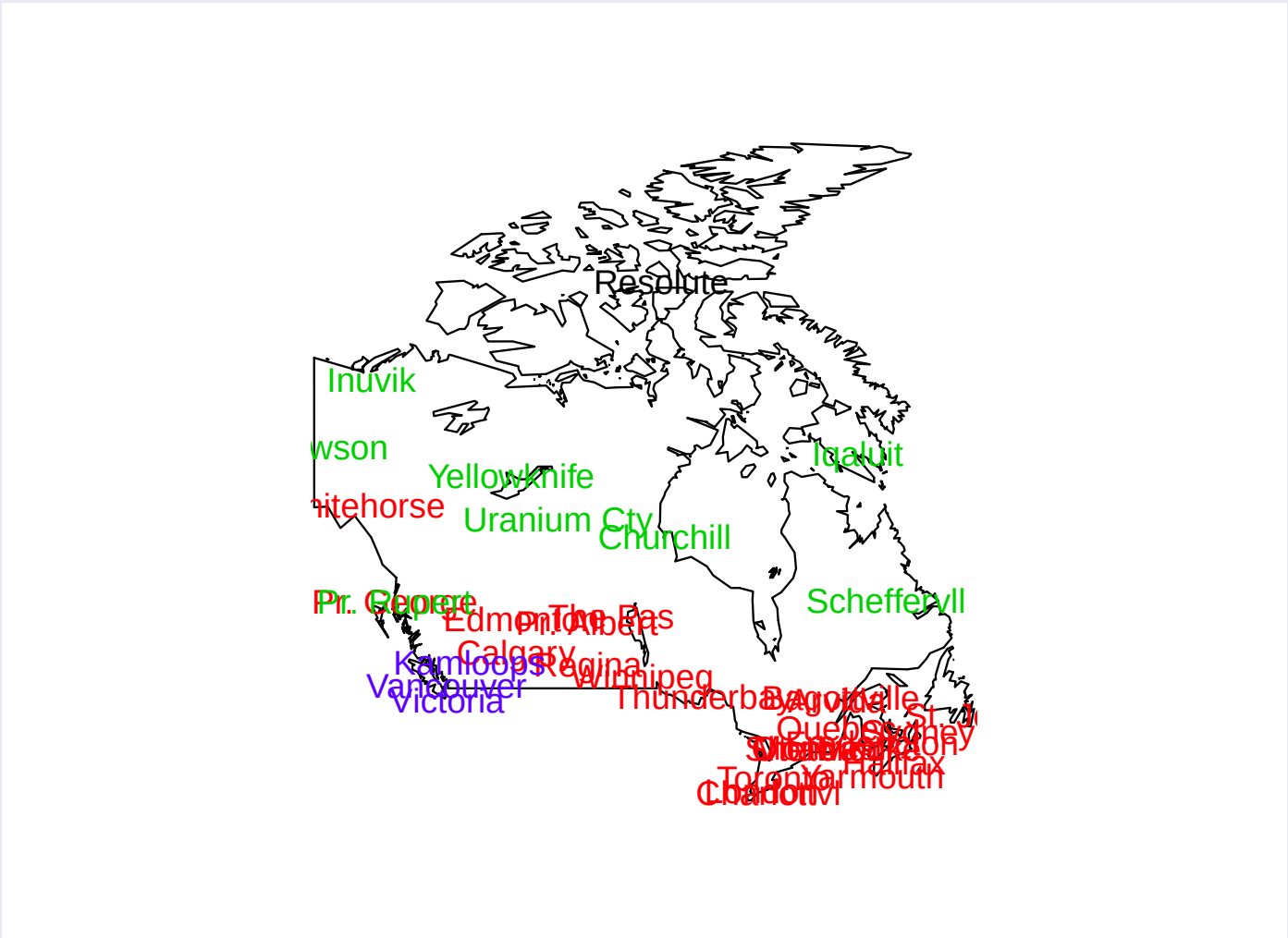
Application: Canadian weather data

Temperature and precipitation curves are first normalized



Clustering Canadian Weather data

Clustering visualization (Jacques & Preda 2012)



Plan

And now ?

Further exploitation of the proposed models

- ranking and ordinal data:
  - new model with search algo. stopped before the end (ordinal), π 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

- **R** package for each kind of data.

And now ?

Heterogeneous data

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

- simple solution:
  - conditional independence
    ($\rightarrow$ 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),
  - ...

Acknowledgements

Université Lille1
Sciences et Technologies
Laboratoire Paul Painlevé
POLYTECH LILLE
Inria INVENTEURS DU MONDE NUMÉRIQUE
cnrs