When dependence opens up possibilities

Elisabeth Gassiat

Université Paris-Saclay (LM-Orsay) and CNRS

Colloquium Ceremade, 7 Décembre 2021
Hidden structure

Dimension reduction

Self-supervised methods

WAVING HANDS

Dependence

State space models

Variational autoencoders

The power of stochastic modeling and statistical inference

Identifiability

Inference/Learning

Uncertainty quantification
Outline

1. Clustering
2. Non parametric Hidden Markov Models
3. Multiple testing
4. Fundamental limits
5. Some other dependence stories
1 Clustering

2 Non parametric Hidden Markov Models

3 Multiple testing

4 Fundamental limits

5 Some other dependence stories
Cluster analysis is used in multivariate statistics to uncover latent groups suspected in the data or to discover groups of homogeneous observations.
Clustering

Cluster analysis can be used
- as an exploratory tool to detect structure
- to perform vector quantization and compress the data
- to reveal a latent group structure corresponding to unobserved heterogeneity

Examples
- Bioinformatics. Analysing gene expression data.
- Marketing. Determining market segments.
- Psychology and sociology. Revealing latent structures.
- Disease monitoring
- ...
Clustering

Clustering is an ill-posed problem which aims to find out interesting structures in the data or to derive a useful grouping of the observations.

But what is a cluster?

Heuristic clustering: hierarchical clustering; partitioning clustering.

Different cluster analysis results on "mouse" data set:
Original Data  k-Means Clustering  EM Clustering
Model based clustering : mixture models

Observations $X = (X_k)_{1 \leq k \leq n}$ coming from $J$ populations.

Given that the observation comes from population $j$, it has distribution $F_j$.

Define latent variables $Z = (Z_k)_{1 \leq k \leq n}$ such that: for each $k$,

$$X_k | Z_k = j \sim F_j.$$ 

Then $X_k$ has distribution

$$\sum_{j=1}^{J} \pi_j F_j.$$ 

$\pi_j$ : probability to come from population $j$. 
Mixture models

Example: Gaussian mixtures $\sum_{j=1}^{J} \pi_j \mathcal{N}(\mu_j; \Sigma_j)$.

Possibly unknown: weights $\pi = (\pi_j)_{1 \leq j \leq J}$, means $(\mu_j)_{1 \leq j \leq J}$, variances $(\Sigma_j)_{1 \leq j \leq J}$.
Mixture models: learning / statistical inference

- Inference of the weights
- Inference of the distributions in each population
- Inference on the latent variables

The Bayes classifier minimizes the classification risk:

\[ \hat{Z}_k \text{ is the } \ell \text{ maximizing } \ell \mapsto P(Z_k = \ell | X). \]
Mixture models: learning / statistical inference

- Inference of the weights
- Inference of the distributions in each population
- Inference on the latent variables

The Bayes classifier minimizes the classification risk:

\[ \hat{Z}_k \text{ is the } \ell \text{ maximizing } \ell \mapsto P(Z_k = \ell | X). \]

Proof. The classification risk for any $X$-measurable $\hat{Z}_k$ is

\[ E \left( 1\{\hat{Z}_k \neq Z_k\} \right) \]
Mixture models: learning / statistical inference

- Inference of the weights
- Inference of the distributions in each population
- Inference on the latent variables

The Bayes classifier minimizes the classification risk:

\[ \hat{Z}_k \text{ is the } \ell \text{ maximizing } \ell \mapsto P(Z_k = \ell | X). \]

Proof. The classification risk for any \( X \)-measurable \( \hat{Z}_k \) is

\[
E \left( \mathbf{1}\{\hat{Z}_k \neq Z_k\} \right) = 1 - E \left( \sum_{j=1}^{J} \mathbf{1}\{Z_k = j, \hat{Z}_k = j\} \right)
\]
Mixture models : learning / statistical inference

- Inference of the weights
- Inference of the distributions in each population
- Inference on the latent variables

The Bayes classifier minimizes the classification risk:

\[ \hat{Z}_k \text{ is the } \ell \text{ maximizing } \ell \mapsto P(Z_k = \ell|X). \]

**Proof.** The classification risk for any \(X\)-measurable \(\hat{Z}_k\) is

\[
E \left( 1\{\hat{Z}_k \neq Z_k\} \right) = 1 - E \left( \sum_{j=1}^{J} 1\{Z_k = j, \hat{Z}_k = j\} \right) \\
= 1 - E \left( \sum_{j=1}^{J} 1\{\hat{Z}_k = j\} P(Z_k = j|X) \right).
\]
Flexibility – Identifiability

Mixture models are not non parametrically identifiable:

\[ \sum_{j=1}^{J} \pi_j F_j = \left( \pi_1 + \frac{\pi_2}{2} \right) \left( \frac{\pi_1 F_1 + \frac{\pi_2}{2} F_2}{\pi_1 + \frac{\pi_2}{2}} \right) + \frac{\pi_2}{2} F_2 + \sum_{j=3}^{J} \pi_j F_j \]

Learning of population components possible only under additional structural assumptions such as

- Parametric mixtures
- One known component in two populations
- Shape restrictions

May lead to poor results in applications
Flexibility – Identifiability

Copy number variation in DNA

![Graph showing copy number variation in DNA](image)

Disease monitoring

![Graph showing disease monitoring](image)
1. Clustering
2. Non parametric Hidden Markov Models
3. Multiple testing
4. Fundamental limits
5. Some other dependence stories
Hidden Markov Models (HMM)

Observations \((X_k)_{k \geq 1}\) are independent conditionally to \((Z_k)_{k \geq 1}\)

\[
\mathcal{L}((X_k)_{k \geq 1}|(Z_k)_{k \geq 1}) = \bigotimes_{k \geq 1} \mathcal{L}(X_k|Z_k)
\]

Latent (unobserved) variables \((Z_k)_{k \geq 1}\) form a Markov chain
HMMs = Mixture models with Markov regime
EG, A. Cleynen, S. Robin (Stat. and Comput. 2016)

$J$ states. Transition matrix $Q$, initial distribution $\pi$.

The distribution of $X = (X_k)_{1 \leq k \leq n}$ is

$$
\sum_{j_1, \ldots, j_n} \pi(j_1) \prod_{i=1}^{n-1} Q_{j_i,j_{i+1}} F_{j_1} \otimes F_{j_2} \otimes \cdots \otimes F_{j_n}.
$$

The marginal distribution of $X_k$ is

$$
\sum_{j=1}^{J} \pi(j) F_j.
$$

Good news: HMMs are non parametrically identifiable!
Identifiability – The tensor trick

\[ H = (\pi, Q, F_1, \ldots, F_J). \]

The distribution of \((X_1, X_2, X_3)\) may be written as

\[
P^{(3)}_H = \sum_{i=1}^{J} \sum_{j=1}^{J} \sum_{m=1}^{J} \pi(i) Q_{i,j} Q_{j,m} F_i \otimes F_j \otimes F_m
\]
Identifiability – The tensor trick

\( H = (\pi, Q, F_1, \ldots, F_J). \)

The distribution of \((X_1, X_2, X_3)\) may be written as

\[
\mathbb{P}_H^{(3)} = \sum_{i=1}^{J} \sum_{j=1}^{J} \sum_{m=1}^{J} \pi(i) Q_{i,j} Q_{j,m} F_i \otimes F_j \otimes F_m
\]

\[
= \sum_{j=1}^{J} \pi(j) \left( \sum_{i=1}^{J} \frac{\pi(i)}{\pi(j)} F_i \right) \otimes F_j \otimes \left( \sum_{m=1}^{J} Q_{j,m} F_m \right)
\]
Identifiability – The tensor trick

\[ H = (\pi, Q, F_1, \ldots, F_J). \]

The distribution of \((X_1, X_2, X_3)\) may be written as

\[
\mathbb{P}(^3) = H = \sum_{i=1}^{J} \sum_{j=1}^{J} \sum_{m=1}^{J} \pi(i) Q_{i,j} Q_{j,m} F_i \otimes F_j \otimes F_m
\]

\[
= \sum_{j=1}^{J} \pi(j) \left( \sum_{i=1}^{J} \frac{\pi(i) Q_{i,j}}{\pi(j)} F_i \right) \otimes F_j \otimes \left( \sum_{m=1}^{J} Q_{j,m} F_m \right)
\]

Choose functions \((h_k)_{1 \leq k \leq L}\) and \(K\), define \(M(K)\) the \(L \times L\) matrix such that

\[
M(K)_{i,j} = \int h_i(x_1) K(x_2) h_j(x_3) \, d\mathbb{P}(^3).
\]
Identifiability – The tensor trick

\[ M(K)_{i,j} = \int h_i(x_1)K(x_2)h_j(x_3) d\mathbb{P}_H^{(3)}. \]

Let \( O \) the \( L \times L \) matrix such that \( O_{i,j} = \int h_i(x) dF_j \). Then

\[ M(1) = O \text{diag}(\pi) Q^2 O^T, \]

and with \( D^K = \text{diag}[(\int K(x) dF_j)_j] \),

\[
M(K) = O \text{diag}(\pi) Q D^K Q O^T.
\]
Identifiability – The tensor trick

\[ M(K)_{i,j} = \int h_i(x_1)K(x_2)h_j(x_3)d\mathbb{P}^{(3)}_H. \]

Let \( O \) the \( L \times L \) matrix such that \( O_{i,j} = \int h_i(x)dF_j. \) Then

\[ M(1) = O\text{diag}(\pi)Q^2O^T, \]

and with \( D^K = \text{diag}[(\int K(x)dF_j)_j], \)

\[ M(K) = O\text{diag}(\pi)QD^KQO^T. \]

Assume

- \( Q \) is full rank,
- \( F_1, \ldots, F_j \) are not linearly dependent.

Then for some \( h_1, \ldots, h_J, \) with \( L = J, \) \( O \) is invertible, and

\[ (M(1)^{-1}M(K) = (QO^T)^{-1}D^KQO^T. \]
Identifiability

\[ M(\mathbf{1})^{-1}M(K) = (QO^T)^{-1}D^KQO^T. \]

Assumptions:
- \( Q \) full rank
- \( F_1, \ldots, F_j \) are distinct

Then all can be recovered using the tensor trick and the law of enough consecutive observations: \( J, Q, \pi, F_1, \ldots, F_j \).

By using blocks, particular functions, one may vary the tensor-trick argument and get identifiability for other models, ex. seasonal HMMs.

\[ \rightarrow \text{Any estimation method using the distribution of enough consecutive observations should work.} \]

Additional assumption: ergodicity of the hidden Markov chain.
Non parametric inference
Y. De Castro, EG, C. lacour (JMLR 2016)
Y. De Castro, EG, S. Le Corff (IEEE-IT 2017)
PHD E. Vernet ; PHD Luc Lehéricy

- Spectral method; least squares; \((m\)-marginal or full\) maximum likelihood – EM-like algorithms.
  Together with some model selection method (needed).
- Bayesian method (automatic model selection but needs good choice of prior).
- Propagation of errors to the posterior probabilities, that is control of

\[
P_H(Z_k = j | X_1, \ldots, X_n) - \hat{P}_H(Z_k = j | X_1, \ldots, X_n)
\]
Nonparametric estimation in sup-norm of emission densities

K. Abraham, I. Castillo, EG (JMLR, to appear)

When $\mu$ is the counting measure on $\mathbb{Z}$.

**Theorem**

Let $M_n$ be a sequence tending to infinity, arbitrarily slowly. There exist estimators $\hat{f}_1, \ldots, \hat{f}_J$ and a permutation $\tau$ such that

$$\mathbb{P}_H(\|\hat{f}_j - f_{\tau(j)}\|_\infty \geq M_n n^{-1/2}) \to 0.$$
Nonparametric estimation in sup-norm of emission densities
K. Abraham, I. Castillo, EG (JMLR, to appear)

When $\mu$ is Lebesgue measure on $\mathbb{R}$.

**Theorem**

Let $L_n$ be a sequence tending to infinity sufficiently slowly. Suppose $f_1, \ldots, f_J$ belong to some Holder space with regularity $s$. Then there exist estimators $\hat{f}_1, \ldots, \hat{f}_J$ and a permutation $\tau$ such that

$$\mathbb{P}_H \left( \| \hat{f}_j - f_{\tau(j)} \|_{\infty} \geq L_n \left( \frac{n}{\log(n)} \right)^{-s/(1+2s)} \right) \rightarrow 0.$$ 

- Uniformity.
- Lower bound / minimax risk.
- Adaptivity.
Seasonal HMMs – with trends

Goal: Multivariate (temperature, precipitations, wind, radiation) stochastic generator.

Latent variable: type of weather.

Mimic the behavior of each variable correctly.
Mimic the global dependence between variables correctly.

Seasonality: in the transition matrix and/or emission distributions.
Trend: in the temperature mean.

Results: theorems; applications
Precipitation-Temperature-Wind

- Transition probabilities: $Q_{ij}(t) \propto \exp(P_{ij}(t))$, with $P_{ij}$ a trigonometric polynomial with period 365 and degree $d$.

- Emission distribution in latent state $k$:

$$X_t \mid \{Z_t = k\} \sim \left( \sum_{m=1}^{M} p_{km} \mu_{km}^{\text{Precip}}(t) \otimes \mu_{km}^{\text{Temp}}(t) \right) \otimes \mu_{k}^{\text{Vent}}$$

- $\mu_{km}^{\text{Precip}}(t) = \begin{cases} \delta_0, & 1 \leq m \leq M_1 \\ \mathcal{E} \left( \frac{\lambda_{km}}{1+\sigma_k^P(t)} \right), & M_1 < m \leq M \end{cases}$

- $\mu_{km}^{\text{Temp}}(t) = \mathcal{N} \left( T_k(t) + S_k(t) + \mu_{km}, \sigma_{km}^2 \right)$

- $\mu_k^{\text{Vent}} = \mathcal{W}(a_k, b_k)$ (loi de Weibull) ou $\mu_k^{\text{Vent}}(t) = \mathcal{W}(a_k, b_k(t))$
Relative frequencies of latent states

\[ P(X_t = k), \ 1 \leq t \leq 365 \]
Trends and seasonalities

Tendances température

Saison température

Saison précip

E. Gassiat (UPS and CNRS)
Emission densities

**Densité vent**

Vitesse (m/s)

**Densité précip**

Precip (mm)

**Densité bruit température**

Température (°C)

**Poids de δ₀**

Etat
1. Clustering

2. Non parametric Hidden Markov Models

3. Multiple testing

4. Fundamental limits

5. Some other dependence stories
Multiple testing

K. Abraham, I. Castillo, EG (JMLR, to appear)

For $i = 1, \ldots, N$, test $H_{0,i} : \theta_i = 0$ against $H_{1,i} : \theta_i = 1$

with observations $X = (X_n)_{n \leq N}$ such that, for $\theta = (\theta_n)_{n \leq N}$,

$$X_n \mid \theta \sim f_{\theta_n}, \quad 1 \leq n \leq N$$

Hidden Markov model prior : $\theta = (\theta_n)_{n \leq N} \sim \text{Markov}(\pi, Q)$.

- Multiple testing procedure by thresholding posterior probabilities

$$\varphi_{\lambda,H}(X) = \left(1 \{ P_H(\theta_i = 0 \mid X) < \lambda \} \right)_{1 \leq i \leq N}$$

- Empirical Bayes : estimate the parameters.
Multiple testing
K. Abraham, I. Castillo, EG (JMLR, to appear)

Uncertainty quantification

- The HMM prior allows unknown $\pi$, $Q$, and $f_0$ and $f_1$ without modeling assumption.
- Precise control of the false discovery rate
- Maximal true discovery proportion
- Less conservative than BH- procedures built under independence
1. Clustering

2. Non parametric Hidden Markov Models

3. Multiple testing

4. Fundamental limits

5. Some other dependence stories
Two-state HMM with multinomial emissions

$n$ observations from $X = (X_1, X_2, \ldots) \in \{1, \ldots, K\}^\mathbb{N}$, with binary latent variables $Z = (Z_1, Z_2, \ldots) \in \{0, 1\}^\mathbb{N}$,

$$P(X_n = k \mid Z) = f_{Z_i}(k),$$

$$Z = (Z_n)_{n \in \mathbb{N}} \sim \text{Stat. Markov}(Q), \quad Q = \begin{pmatrix} 1 - p & p \\ q & 1 - q \end{pmatrix}.$$

We denote $\theta = (p, q, f_0, f_1)$,

$$P_\theta(Z_1 = 0) = \frac{q}{p + q}, \quad P_\theta(Z_1 = 1) = \frac{p}{p + q}.$$

Identifiability and inference with parametric rates as soon as :

- $0 < p < 1$ and $0 < q < 1$ (ergodic Markov chain)
- $p + q \neq 1$ (det $Q \neq 0$ : non independent latent variables)
- $\|f_0 - f_1\| > 0$ (distinct populations)
Fundamental limits of the learning domain

We might be interested in sparsity. Sparsity occurs when \( \#\{i : Z_i = 1\} = o(n) \), that is \( p = o(1) \) (and \( q \gg p \)).

We might be interested in how far from each other have populations to be so that they can be separated: that is in the case \( \|f_0 - f_1\| = o(1) \).

We might be interested in how much have the latent variables to be dependent so that two different populations can be identified.
Fundamental limits of the learning domain

We might be interested in sparsity. Sparsity occurs when \( \# \{ i : Z_i = 1 \} = o(n) \), that is \( p = o(1) \) (and \( q \gg p \)).

We might be interested in how far from each other have populations to be so that they can be separated: that is in the case \( \| f_0 - f_1 \| = o(1) \).

We might be interested in how much have the latent variables to be dependent so that two different populations can be identified.

With finitely many data, how is learning impacted by being near the limit cases? That is: how do the constants, in the upper bound on the risk, depend on being near the limit cases?

\[
\Theta = \Theta(\delta, \epsilon, \zeta) = \{ \theta : p, q \geq \delta, |1 - p - q| \geq \epsilon, \| f_0 - f_1 \| \geq \zeta \}. 
\]
Fundamental limits of the learning domain: upper bounds


Fix \( L > 0 \), and define

\[
\Theta_L = \Theta_L(\delta, \epsilon, \zeta) = \Theta \cap \{1 - |1 - p - q| \geq L\}.
\]

**Theorem**

There exist an estimator \( \hat{\theta} = (\hat{p}, \hat{q}, \hat{f}_0, \hat{f}_1) \) and a constant \( C = C(K, L) > 0 \) such that for all \( 1 \leq x^2 \leq n\delta^2\epsilon^4\zeta^6 \)

\[
\sup_{\theta \in \Theta_L} \mathbb{P}_\theta \left( \max(|\hat{p} - p|, |\hat{q} - q|) > \frac{Cx}{\sqrt{n\delta^2\epsilon^4\zeta^6}} \max(\delta, \epsilon\zeta) \right) \leq e^{-x^2},
\]

\[
\sup_{\theta \in \Theta_L} \mathbb{P}_\theta \left( \max(||\hat{f}_0 - f_0||, ||\hat{f}_1 - f_1||) > \frac{Cx}{\sqrt{n\delta^2\epsilon^4\zeta^4}} \right) \leq e^{-x^2}.
\]
Fundamental limits of the learning domain: lower bounds

**Theorem**

Assume $\delta \leq 1/6$, $\epsilon \leq 1/3$, $\zeta \leq \frac{\sqrt{2[K/2]}}{4K}$, $L \leq 1/3$ and $n\delta^2\epsilon^4\zeta^6 \geq 1$. Then for some $c = c(K) > 0$,

$$\inf_{\tilde{\theta}} \sup_{\theta \in \Theta_L} \mathbb{P}_{\theta} \left( \max(|\tilde{\theta} - \theta|, |\tilde{q} - q|) > \frac{c}{\sqrt{n\delta^2\epsilon^4\zeta^6}} \max(\delta, \epsilon\zeta) \right) \geq 1/4,$$

$$\inf_{\tilde{\theta}} \sup_{\theta \in \Theta_L} \mathbb{P}_{\theta} \left( \max(||\tilde{f}_0 - f_0||, ||\tilde{f}_1 - f_1||) > \frac{c}{\sqrt{n\delta^2\epsilon^4\zeta^4}} \right) \geq 1/4,$$

where the infima are over all estimators $\tilde{\theta} = (\tilde{\theta}, \tilde{q}, \tilde{f}_0, \tilde{f}_1)$. 
HMMs: Conclusions/Questions

- HMM modelling allows flexible emission distributions: fully nonparametric identifiable under minimal assumptions.
- Usual adaptive minimax rates for learning all parameters of the model. Various possible inference methodologies.
- Control of the propagation of errors to posterior probabilities.
- Learning rates are deteriorated when the latent sequence becomes sparse / nearly independent / the emission multinomials become close.
- Algorithms/theory: detection of problematic regions?
- Robustness to the hidden Markov modelling?
- Non parametric clustering for more general dependence structures?
1. Clustering
2. Non parametric Hidden Markov Models
3. Multiple testing
4. Fundamental limits
5. Some other dependence stories
Nonparametric regression on hidden phi-mixing variables

PHD T. Dumont, CIFRE IdServices

Observations:

\[ X_k = f_*(Z_k) + \epsilon_k. \]

\( f_* \) is an unknown function defined on a subset \( \mathbb{X} \) of \( \mathbb{R}^m \) and taking values in \( \mathbb{R}^\ell \).

\( (\epsilon_k)_{k \geq 1} \) is an i.i.d. sequence of standard Gaussian random vectors. The \( Z_k \)'s are not observed and their distribution is unknown.

Some identifiability can be obtained when the \( (Z_k)'s \) are not independent under various (non parametric) assumptions.
Regression on hidden variables: Estimation

**Figure** – Estimates with no isometry for $f_1$ [top] and the isometry $x \mapsto 1 - x$ for $f_1$ [bottom].
Regression on hidden variables : Estimation

Figure – True image $f_\star([-1, 1])$ (red) and estimations after 100 iterations of the algorithm over 100 Monte Carlo runs (grey).
Deconvolution with unknown noise distribution
EG, S. Le Corff, L. Lehéricy, Annals of Stat., to appear

The observation $Y$ is given by

$$Y = X + \varepsilon,$$

$X$ is the signal and $\varepsilon$ is the noise, they are independent random variables with both unknown distribution.

No assumption on the noise and weak structure assumptions on the signal allow identifiability:

- Multidimensional observations: $Y, X, \varepsilon$ are in $\mathbb{R}^d$, $d \geq 2$
- No distributional assumption on the noise, except that it has independent components
- The distribution of the signal has light tails
- Some dependency assumption on the components of the signal
Deconvolution with unknown noise distribution

- State space models/General HMMs (EG, S. Le Corff, L. Lehéricy, JMLR 2020).
- Denoising/Self-supervised algorithms (J. Ollion, C. Ollion, EG, L. Lehéricy, S. Le Corff, 2021)
- Manifold learning PHD J. Capitao-Miniconi (on-going)

But this is another full story...
Thank you for your attention!