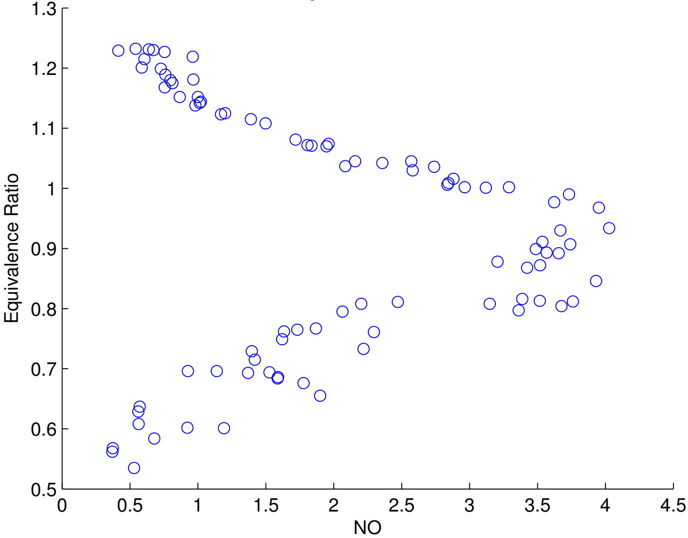
Mixture of Gaussian Regressions with logistic weights and conditional density estimation

E. Le Pennec
(CMAP - École Polytechnique)
and
L. Montuelle (SELECT - Inria Saclay / Université Paris Sud)

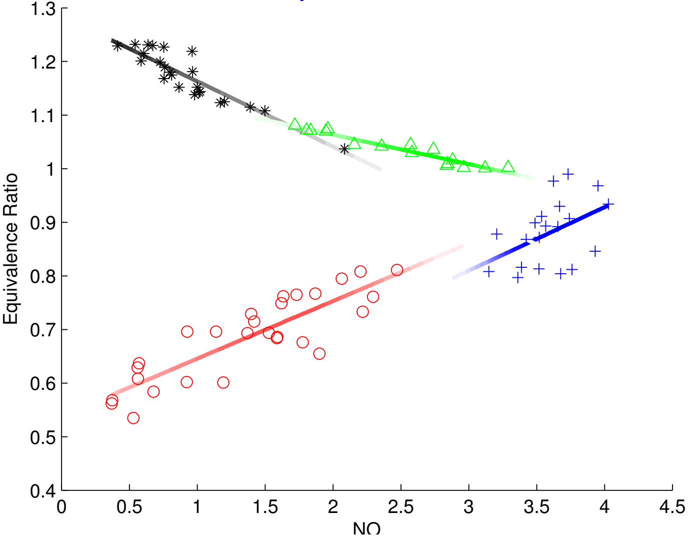
Alea
11 march 2014

NO vs air/ethanol mix



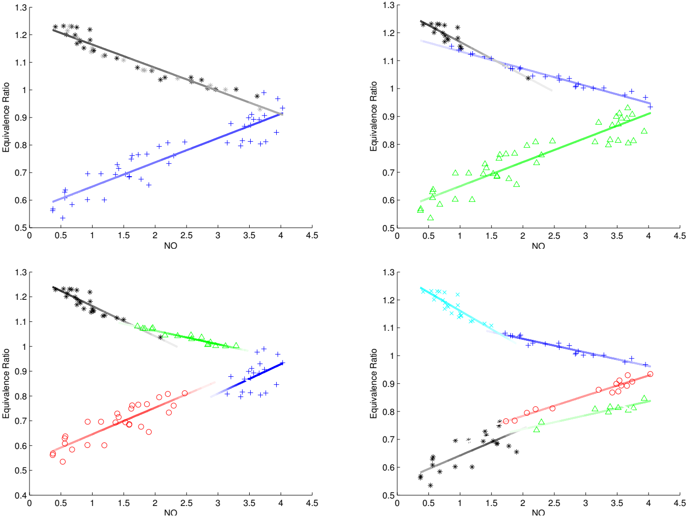
- Regression

NO vs air/ethanol mix



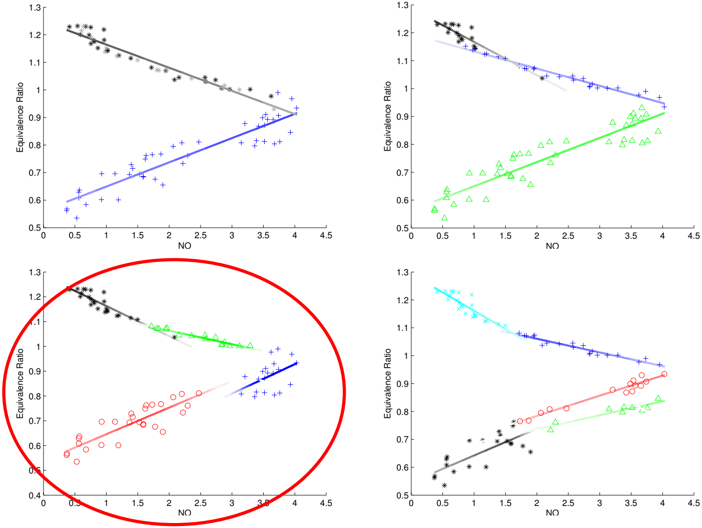
- Regression Mixture

NO vs air/ethanol mix



- Regression Mixture
- Mixtures of Gaussian regressions with logistic weights

NO vs air/ethanol mix



- Regression Mixture
- Mixtures of Gaussian regressions with logistic weights
- Model selection

Statistical modeling

- **Data:** $(X_i, Y_i)_{i \leq n} \in [0;1]^d \times \mathbb{R}^p$
  - $X_i \perp X_j$
  - $Y_i | (X_k)_k \perp Y_j | (X_k)_k$
  - $Y|X$ has a density s_0 w.r.t. Lebesgue measure
- **Regression** = specific modeling of the conditional density $s_0(.|x)$
- **Gaussian regression mixture with logistic weights:**

$$s_{K,\upsilon,\Sigma,w}(y|x) = \sum_{k=1}^{K} \pi_{w,k}(x) \Phi_{\upsilon_k(x),\Sigma_k(x)}(y),$$

with
  - $\pi_{w,k}(x) = \frac{e^{w_k(x)}}{\sum_{k'=1}^{K} e^{w_{k'}(x)}}$, logistic weights
  - $\Phi_{\upsilon_k(x),\Sigma_k(x)}$ density of $\mathcal{N}(\upsilon_k(x), \Sigma_k(x))$

- **Parameters:**
  - K: number of components
  - υ and Σ: K regression functions and covariance matrices functions
  - w: K weights functions defining the mixture proportions

Models

- **Gaussian regression mixture with logistic weights:**

$$s_{K,\upsilon,\Sigma,w}(y|x) = \sum_{k=1}^{K} \pi_{w,k}(x)\Phi_{\upsilon_k(x),\Sigma_k(x)}(y),$$

with
  - $\pi_{w,k}(x) = \frac{e^{w_k(x)}}{\sum_{k'=1}^{K} e^{w_{k'}(x)}}$, logistic weights
  - $\Phi_{\upsilon_k(x),\Sigma_k(x)}$ density of $\mathcal{N}(\upsilon_k(x), \Sigma_k(x))$
- **Parameters:** $\theta = (K, \upsilon, \Sigma, w)$
  - K: number of components
  - υ and Σ: K regression functions and covariance matrices functions
  - w: K weights functions defining the mixture proportions
- **Model** $S_m = \{s_\theta, \theta \in \Theta_m\}$ with $\Theta_m = \{K\} \otimes \Upsilon_K \otimes V_K \otimes W_k$:
  - K: number of components.
  - Υ_K and V_K: sets for the K-tuple of regressions functions and covariance matrices functions.
  - W_K: sets for for the K-tuple of weights functions.
- Typical choice:
  - Υ_K and W_K: tensorial product of polynomial sets of low degree.
  - W_K: constant covariance structures independent of X.

Maximum likelihood and penalization

- **Model** $S_m = \{s_\theta, \theta \in \Theta_m\}$ with $\Theta_m = \{K\} \otimes \Upsilon_K \otimes V_K \otimes W_K$:
  - K: number of components.
  - Υ_K and V_K: sets for the K-tuple of regressions functions and covariance matrices functions.
  - W_K: sets for for the K-tuple of weights functions.
- **Maximum likelihood estimation** within each model:

$$\widehat{s}_m = \operatorname*{argmax}_{\theta \in \Theta_m} - \sum_{i=1}^{n} \ln s_\theta(Y_i|X_i)$$

- **Model selection** by a penalization proportional to the dimension:

$$\widehat{m} = \operatorname*{argmin}_{m \in \mathcal{M}} \sum_{k=1}^{K} - \ln \widehat{s}_m(Y_i|X_i) + \kappa \dim \Theta_m$$

- Usual complexity/fidelity tradeoff.

Contributions

Characterization of the theoretical performances

- **Penalty choice:** $\mathrm{pen}(m) = \kappa(C + \ln n) \dim(S_m)$.
- **Oracle inequality:**

$$\mathbb{E}\left[JKL_{\rho}^{\otimes n}(s_0, \widehat{s}_{\widehat{m}})\right] \leq C_1 \inf_{m \in \mathcal{M}} \left(\inf_{s_m \in S_m} KL^{\otimes n}(s_0, s_m) + \frac{\mathrm{pen}(m)}{n} \right) + \frac{C_2}{n}$$

Numerical implementation of the penalized maximum likelihood

- **EM type minimization scheme** with a focus on initialization issues.
- Practical scheme for the **penalty calibration** with the slope heuristic approach.

Conditional density and selection

- **General framework:** observation of (X_i, Y_i) with X_i independent and Y_i cond. independent of law of density $s_0(y|X_i)$.
- **Goal:** estimation of $s_0(y|x)$.
- **Penalized model selection principle:**
  - choice of a collection of cond. dens. models $S_m = \{s_m(y|x)\}$ with $m \in \mathcal{S}$,
  - Maximum likelihood estimation of a cond. density $\hat{s}_m$ for each model S_m:

$$\hat{s}_m = \operatorname*{argmin}_{s_m \in S_m} - \sum_{i=1}^{n} \ln s_m(Y_i|X_i)$$

  - Selection of a model $\widehat{m}$ by

$$\widehat{m} = \operatorname*{argmin}_{m \in \mathcal{S}} - \sum_{i=1}^{n} \ln \widehat{s}_m(Y_i|X_i) + \mathrm{pen}(m).$$

    with $\mathrm{pen}(m)$ *well chosen*.
- **Typical oracle inequality result:**

$$\mathbb{E}\left[d^2(s_0, \widehat{s}_{\widehat{m}})\right] \leq C \inf_{m \in \mathcal{S}} \left(\inf_{s_m \in S_m} KL(s_0, s_m) + \frac{\mathrm{pen}(m)}{n} \right) + \frac{C'}{n}.$$

- **Short bib.**: Rosenblatt, Fan et al., de Gooijer and Zerom, Efromovitch, Brunel, Comte, Lacour... / Plugin, direct estimation, L^2, minimax, censure...

Ideal oracle inequality

- **Oracle inequality:**

$$\mathbb{E}\left[KL^{\otimes_n}(s_0, \widehat{s}_{\widehat{m}})\right] \leq C_1 \inf_{S_m \in \mathcal{S}} \left(\underbrace{\inf_{s_m \in S_m} KL^{\otimes_n}(s_0, s_m)}_{\text{Bias term}} + \underbrace{\frac{\Delta(m)}{n}}_{\text{Variance term}} \right)$$

  as soon as $\mathrm{pen}(m)$ is *large enough*
- **Divergence adapted to the conditional density setting:**
  - Divergence on the product density conditioned on the design (Kolaczyk, Bigot).
  - *Tensorization* principle and expectation on the design: design:

$$KL \to KL^{\otimes_n}(s, s') = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} KL\left(s(\cdot|X_i), s'(\cdot|X_i)\right)\right]$$

- Much **more information** using the second approach because losses used are *larger*.
- Ability to handle **independent but non i.i.d. case** and **integrated loss**.
- Classical density estimation theorem if $s(\cdot|X_i) = s(\cdot)$.

Notations

- Let for any function $g(x,y)$,
  - $P_n^{\otimes_n}(g)$: its **empirical process** $P_n^{\otimes_n}(g) = \frac{1}{n}\sum_{i=1}^{n} g(X_i, Y_i)$.
  - $P^{\otimes_n}(g)$: its **expectation** $P^{\otimes_n}(g) = \mathbb{E}\left[P_n^{\otimes_n}(g)\right] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} g(X_i, Y_i)\right]$.
  - $\nu_n^{\otimes_n}(g) = P_n^{\otimes_n}(g) - P^{\otimes_n}(g)$: its **recentered** process.
- **Maximum likelihood estimate:**

$$\widehat{s}_m = \operatorname*{argmin}_{s_m \in S_m} \sum_{i=1}^{n} -\ln s_m(Y_i|X_i) = \operatorname*{argmin}_{s_m \in S_m} P_n^{\otimes_n}(-\ln s_m)$$
$$= \operatorname*{argmin}_{s_m \in S_m} P_n^{\otimes_n}(-\ln \frac{s_m}{s_0})$$

- **Best *projection*:**

$$\widetilde{s}_m = \operatorname*{argmin}_{s_m \in S_m} KL^{\otimes_n}(s_0, s_m) = \operatorname*{argmin}_{s_m \in S_m} P^{\otimes_n}(-\ln \frac{s_m}{s_0})$$
$$= \operatorname*{argmin}_{s_m \in S_m} P^{\otimes_n}(-\ln s_m)$$

Ideal penalty

- By definition:

$$KL^{\otimes n}(s_0, \widehat{s}_m) = P_n^{\otimes n}\left(-\ln \frac{\widehat{s}_m}{s_0}\right) \underbrace{-\nu_n^{\otimes n}\left(-\ln \frac{\widehat{s}_m}{s_0}\right)}_{\mathrm{pen}_{\mathrm{id}}(m)/n}$$

- With the ***ideal* penalty** $\mathrm{pen}_{\mathrm{id}}(m)$:

$$\begin{aligned} KL^{\otimes n}(s_0, \widehat{s}_{\widehat{m}}) &= P_n^{\otimes n}\left(-\ln \frac{\widehat{s}_{\widehat{m}}}{s_0}\right) + \frac{\mathrm{pen}_{\mathrm{id}}(\widehat{m})}{n} \\ &\leq \inf_m P_n^{\otimes n}\left(-\ln \frac{\widehat{s}_m}{s_0}\right) + \frac{\mathrm{pen}_{\mathrm{id}}(m)}{n} \leq \inf_m KL^{\otimes n}(s_0, \widehat{s}_m) \\ &\leq \inf_m \left(KL^{\otimes n}(s_0, \widetilde{s}_m) + \left(KL^{\otimes n}(s_0, \widehat{s}_m) - KL^{\otimes n}(s_0, \widetilde{s}_m)\right)\right) \end{aligned}$$

- **Ideal penalty oracle inequality:**

$$\mathbb{E}\left[KL^{\otimes n}(s_0, \widehat{s}_{\widehat{m}})\right] \leq \inf_{S_m \in \mathcal{S}} \left(\underbrace{KL^{\otimes n}(s_0, \widetilde{s}_m)}_{\text{Bias term}} + \underbrace{\mathbb{E}\left[KL^{\otimes n}(s_0, \widehat{s}_m) - KL^{\otimes n}(s_0, \widetilde{s}_m)\right]}_{\text{Variance term}}\right)$$

Non ideal penalization

- By construction

$$
\begin{aligned}
KL^{\otimes n}(s_0, \widehat{s}_{\widehat{m}}) &= P_n^{\otimes n}\left(-\ln \frac{\widehat{s}_{\widehat{m}}}{s_0}\right) - \nu_n^{\otimes n}\left(-\ln \frac{\widehat{s}_{\widehat{m}}}{s_0}\right) \\
&= P_n^{\otimes n}\left(-\ln \frac{\widehat{s}_{\widehat{m}}}{s_0}\right) + \frac{\mathrm{pen}(\widehat{m})}{n} \\
&\qquad - \nu_n^{\otimes n}\left(-\ln \frac{\widehat{s}_{\widehat{m}}}{s_0}\right) - \frac{\mathrm{pen}(\widehat{m})}{n} \\
&\leq \min_{S_m \in \mathcal{S}}\left(P_n^{\otimes n}\left(-\ln \frac{\widehat{s}_m}{s_0}\right) + \frac{\mathrm{pen}(m)}{n}\right) \\
&\qquad - \nu_n^{\otimes n}\left(-\ln \frac{\widehat{s}_{\widehat{m}}}{s_0}\right) - \frac{\mathrm{pen}(\widehat{m})}{n}
\end{aligned}
$$

Non ideal penalization

- Using $\widetilde{s}_m = \mathrm{argmin}_{s_m \in S_m} KL^{\otimes_n}(s_0, s_m)$:

$$KL^{\otimes_n}(s, \widehat{s}_{\widehat{m}}) \leq \min_{S_m \in \mathcal{S}} \left(P_n^{\otimes_n}\left(-\ln \frac{\widehat{s}_m}{s_0}\right) + \frac{\mathrm{pen}(m)}{n} \right) - \nu_n^{\otimes_n}\left(-\ln \frac{\widehat{s}_{\widehat{m}}}{s_0}\right) - \frac{\mathrm{pen}(\widehat{m})}{n}$$

$$KL^{\otimes_n}(s_0, \widehat{s}_{\widehat{m}}) \leq \min_{S_m \in \mathcal{S}} \left(P_n^{\otimes_n}\left(-\ln \frac{\widetilde{s}_m}{s_0}\right) + \frac{\mathrm{pen}(m)}{n} \right) - \nu_n^{\otimes_n}\left(-\ln \frac{\widehat{s}_{\widehat{m}}}{s_0}\right) - \frac{\mathrm{pen}(\widehat{m})}{n}$$

Non ideal penalization

- **Summary:**

$$KL^{\otimes_n}(s_0, \widehat{s}_{\widehat{m}}) \leq \min_{S_m \in \mathcal{S}} \left(P_n^{\otimes_n} \left(-\ln \frac{\widetilde{s}_m}{s_0} \right) + \frac{\mathrm{pen}(m)}{n} \right) - \nu_n^{\otimes_n} \left(-\ln \frac{\widehat{s}_{\widehat{m}}}{s_0} \right) - \frac{\mathrm{pen}(\widehat{m})}{n}$$

- **Oracle inequality** up to something:

$$\mathbb{E}\left[KL^{\otimes_n}(s_0, \widehat{s}_{\widehat{m}}) \right] \leq \min_{S_m \in \mathcal{S}} \left(KL^{\otimes_n}(s_0, \widetilde{s}_m) + \frac{\mathrm{pen}(m)}{n} \right) + \mathbb{E}\left[-\nu_n^{\otimes_n} \left(-\ln \frac{\widehat{s}_{\widehat{m}}}{s_0} \right) - \frac{\mathrm{pen}(\widehat{m})}{n} \right]$$

- If $\mathbb{E}\left[-\nu_n^{\otimes_n} \left(-\ln \frac{\widehat{s}_{\widehat{m}}}{s_0} \right) - \frac{\mathrm{pen}(\widehat{m})}{n} \right] \leq 0$ then **exact** oracle inequality!
- If $\mathbb{E}\left[-\nu_n^{\otimes_n} \left(-\ln \frac{\widehat{s}_{\widehat{m}}}{s_0} \right) - \frac{\mathrm{pen}(\widehat{m})}{n} - \epsilon KL^{\otimes_n}(s_0, \widehat{s}_{\widehat{m}}) \right] \leq 0$ then **inexact** oracle inequality.

Kullback-Leibler and extension

- **Issue** in the previous approach: control of

$$\nu_n^{\otimes_n}\left(-\ln\frac{\widehat{s}_{\widehat{m}}}{s_0}\right)$$

hard due to the **unboundedness** of $-\ln\frac{\widehat{s}_{\widehat{m}}}{s_0}$

- **Trick:** replace this quantity by the bounded one

$$-\frac{1}{\rho}\ln\frac{\rho\widehat{s}_{\widehat{m}}+(1-\rho)s_0}{s_0}\leq-\frac{1}{\rho}\ln(1-\rho)$$

- By convexity,

$$-\frac{1}{\rho}\ln\frac{\rho\widehat{s}_{\widehat{m}}+(1-\rho)s_0}{s_0}\leq-\ln\frac{\widehat{s}_{\widehat{m}}}{s_0}$$

- **Jensen-Kullback-Leibler divergence:**

$$JKL_\rho^{\otimes_n}(s_0,\widehat{s}_m)=P^{\otimes_n}\left(-\frac{1}{\rho}\ln\frac{\rho\widehat{s}_{\widehat{m}}+(1-\rho)s_0}{s_0}\right)=\frac{1}{\rho}KL^{\otimes_n}(s_0,\rho\widehat{s}_m+(1-\rho)s_0)$$
$$\leq P^{\otimes_n}\left(-\ln\frac{\widehat{s}_{\widehat{m}}}{s_0}\right)=KL^{\otimes_n}(s_0,\widehat{s}_m)$$

$JKL^{\otimes n}$ and non ideal penalization

- By construction

$$
\begin{aligned}
JKL_\rho^{\otimes n}(s_0, \widehat{s}_{\widehat{m}}) &= P_n^{\otimes n}\left(-\frac{1}{\rho}\ln\frac{\rho\widehat{s}_{\widehat{m}} + (1-\rho)s_0}{s_0}\right) \\
&\qquad - \nu_n^{\otimes n}\left(-\frac{1}{\rho}\ln\frac{\rho\widehat{s}_{\widehat{m}} + (1-\rho)s_0}{s_0}\right) \\
&\leq P_n^{\otimes n}\left(-\ln\frac{\widehat{s}_{\widehat{m}}}{s_0}\right) - \nu_n^{\otimes n}\left(-\frac{1}{\rho}\ln\frac{\rho\widehat{s}_{\widehat{m}} + (1-\rho)s_0}{s_0}\right) \\
&\leq P_n^{\otimes n}\left(-\ln\frac{\widehat{s}_{\widehat{m}}}{s_0}\right) + \frac{\mathrm{pen}(\widehat{m})}{n} \\
&\qquad - \nu_n^{\otimes n}\left(-\frac{1}{\rho}\ln\frac{\rho\widehat{s}_{\widehat{m}} + (1-\rho)s_0}{s_0}\right) - \frac{\mathrm{pen}(\widehat{m})}{n} \\
&\leq \min_{S_m\in\mathcal{S}}\left(P_n^{\otimes n}\left(-\ln\frac{\widehat{s}_m}{s_0}\right) + \frac{\mathrm{pen}(m)}{n}\right) \\
&\qquad - \nu_n^{\otimes n}\left(-\frac{1}{\rho}\ln\frac{\rho\widehat{s}_{\widehat{m}} + (1-\rho)s_0}{s_0}\right) - \frac{\mathrm{pen}(\widehat{m})}{n}
\end{aligned}
$$

$JKL^{\otimes_n}$ and non ideal penalization

- Using $\widetilde{s}_m = \mathrm{argmin}_{s_m \in S_m} KL^{\otimes_n}(s_0, s_m)$:

$$
\begin{aligned}
JKL_\rho^{\otimes_n}(s_0, \widehat{s}_{\widehat{m}}) \leq \min_{S_m \in \mathcal{S}} & \left(P_n^{\otimes_n}\left(-\ln \frac{\widehat{s}_m}{s_0}\right) + \frac{\mathrm{pen}(m)}{n} \right) \\
& - \nu_n^{\otimes_n}\left(-\frac{1}{\rho} \ln \frac{\rho \widehat{s}_{\widehat{m}} + (1-\rho) s_0}{s_0}\right) - \frac{\mathrm{pen}(\widehat{m})}{n} \\
JKL_\rho^{\otimes_n}(s_0, \widehat{s}_{\widehat{m}}) \leq \min_{S_m \in \mathcal{S}} & \left(P_n^{\otimes_n}\left(-\ln \frac{\widetilde{s}_m}{s_0}\right) + \frac{\mathrm{pen}(m)}{n} \right) \\
& - \nu_n^{\otimes_n}\left(-\frac{1}{\rho} \ln \frac{\rho \widehat{s}_{\widehat{m}} + (1-\rho) s_0}{s_0}\right) - \frac{\mathrm{pen}(\widehat{m})}{n}
\end{aligned}
$$

$JKL^{\otimes_n}$ and non ideal penalization

- **Summary:**

$$JKL_\rho^{\otimes_n}(s_0,\widehat{s}_{\widehat{m}}) \leq \min_{S_m\in\mathcal{S}} \left(P_n^{\otimes_n}\left(-\ln\frac{\widetilde{s}_m}{s_0}\right) + \frac{\mathrm{pen}(m)}{n}\right) - \nu_n^{\otimes_n}\left(-\frac{1}{\rho}\ln\frac{\rho\widehat{s}_{\widehat{m}}+(1-\rho)s_0}{s_0}\right) - \frac{\mathrm{pen}(\widehat{m})}{n}$$

- **Oracle inequality up to something:**

$$\mathbb{E}\left[JKL_\rho^{\otimes_n}(s_0,\widehat{s}_{\widehat{m}})\right] \leq \min_{S_m\in\mathcal{S}} \left(KL^{\otimes_n}(s_0,\widetilde{s}_m) + \frac{\mathrm{pen}(m)}{n}\right) + \mathbb{E}\left[-\nu_n^{\otimes_n}\left(-\frac{1}{\rho}\ln\frac{\rho\widehat{s}_{\widehat{m}}+(1-\rho)s_0}{s_0}\right) - \frac{\mathrm{pen}(\widehat{m})}{n}\right]$$

- Under some assumptions on the model collection, **it exists a penalty such that**
$$\mathbb{E}\left[-\nu_n^{\otimes_n}\left(-\ln\frac{\widehat{s}_{\widehat{m}}}{s_0}\right) - \frac{\mathrm{pen}(\widehat{m})}{n} - \epsilon JKL_\rho^{\otimes_n}(s_0,\widehat{s}_{\widehat{m}})\right] \leq 0$$
- For such a penalty, one has an **inexact** oracle inequality.

Theorem

Assumption (H): For every model S_m in the collection $\mathcal{S}$, there is a non-decreasing function $\phi_m(\delta)$ such that $\delta \mapsto \frac{1}{\delta}\phi_m(\delta)$ is non-increasing on $(0, +\infty)$ and for every $\sigma \in \mathbb{R}^+$ and every $s_m \in S_m$

$$\int_0^\sigma \sqrt{H_{[\cdot],d^{\otimes_n}}(\epsilon, S_m(s_m, \sigma))}\, d\epsilon \leq \phi_m(\sigma).$$

Assumption (K): There is a family $(x_m)_{m\in\mathcal{M}}$ of non-negative number such that

$$\sum_{m\in\mathcal{M}} e^{-x_m} \leq \Sigma < +\infty$$

Theorem

Assume we observe (X_i, Y_i) with unknown conditional s_0. Let $\mathcal{S} = (S_m)_{m\in\mathcal{M}}$ a at most countable collection of conditional density sets. Assume Assumptions (H), (K) and (S) hold.
Let $\widehat{s}_m$ be a δ -log-likelihood minimizer in S_m:

$$\sum_{i=1}^n -\ln(\widehat{s}_m(Y_i|X_i)) \leq \inf_{s_m\in S_m}\left(\sum_{i=1}^n -\ln(s_m(Y_i|X_i))\right) + \delta$$

Then for any $\rho \in (0,1)$ and any $C_1 > 1$, there is a constant κ_0 depending only on ρ and C_1 such that, as soon as for every index $m \in \mathcal{M}$ $\mathrm{pen}(m) \geq \kappa(\mathfrak{D}_m + x_m)$ with $\kappa > \kappa_0$
where $\mathfrak{D}_m = n\sigma_m^2$ with σ_m the unique root of $\frac{1}{\sigma}\phi_m(\sigma) = \sqrt{n}\sigma$,
the penalized likelihood estimate $\widehat{s}_{\widehat{m}}$ with $\widehat{m}$ defined by

$$\widehat{m} = \operatorname*{argmin}_{m\in\mathcal{M}} \sum_{i=1}^n -\ln(\widehat{s}_m(Y_i|X_i)) + \mathrm{pen}(m)$$

satisfies
$$\mathbb{E}\left[JKL_\rho^{\otimes_n}(s_0, \widehat{s}_{\widehat{m}})\right] \leq C_1\left(\inf_{S_m\in\mathcal{S}}\left(\inf_{s_m\in S_m} KL^{\otimes_n}(s_0, s_m) + \frac{\mathrm{pen}(m)}{n}\right) + \frac{\kappa_0\Sigma + \delta}{n}\right).$$

Simplified Theorem...

- **Oracle inequality:**

$$\mathbb{E}\left[JKL_{\rho}^{\otimes_n}(s_0, \widehat{s}_{\widehat{m}})\right] \leq C_1 \left(\inf_{S_m \in \mathcal{S}} \left(\inf_{s_m \in S_m} KL^{\otimes_n}(s_0, s_m) + \frac{\mathrm{pen}\, m}{n} \right) + \frac{\kappa_0 \Sigma + \delta}{n} \right)$$

as soon as

$$\mathrm{pen}(m) \geq \kappa \left(\mathfrak{D}_m + x_m\right) \quad \text{with } \kappa > \kappa_0,$$

where $\mathfrak{D}_m$ measure the **complexity of the model** S_m (entropy term) and x_m the **coding cost** within the collection.

- $\mathfrak{D}_m$ linked to the **bracketing entropy** of S_m with respect to the tensorized Hellinger distance $d^{2\otimes_n}$.
- Often $\mathfrak{D}_m \propto (\log n) \dim(S_m)$...

Penalty and complexities

- Control required on

$$-\nu_n^{\otimes_n}\left(-\ln\frac{\widehat{s}_{\widehat{m}}}{s_0}\right)-\frac{\mathrm{pen}(\widehat{m})}{n}-\epsilon JKL_\rho^{\otimes_n}(s_0,\widehat{s}_{\widehat{m}})$$

  through a **supremum**!
- **Control in expectation requires** a $\mathrm{pen}(m)$ taking into account
  - the **intrinsic complexity of the model**,
  - the **complexity of the collection**.
- Here:
  - **Model complexity:** entropy complexity $\mathfrak{D}_m$ defined from the *bracketing entropy* $H_{[\cdot],d^{\otimes_n}}(\epsilon, S_m)$ of S_m with respect to the tensorized Hellinger distance $d^{2\otimes_n}$.
  - **Collection (coding):** Kraft type inequality $\sum_{m\in\mathcal{S}} e^{-x_m} \leq \Sigma < +\infty$
- **Classical constraint on the penalty**

$$\mathrm{pen}(m) \geq \kappa\left(\mathfrak{D}_m + x_m\right) \quad \text{with } \kappa > \kappa_0.$$

- Often $\mathfrak{D}_m \propto (\ln(n))\dim(S_m)$ and thus **classical penalization by dimension** setting...

Brackets and complexity

- **Bracketing entropy:** $H_{[\cdot],d^{\otimes_n}}(\epsilon, S) =$ logarithm of the minimum number of brackets $[t_i^-, t_i^+]$ such that
  - $\forall i, d^{\otimes_n}(t_i^-, t_i^+) \leq \epsilon$
  - $\forall s \in S, \exists i, t_i^- \leq s \leq t_i^+$

  where $d^{\otimes_n} = \sqrt{d^{2\otimes_n}} = \sqrt{\mathbb{E}\left[\frac{1}{n}\sum d^2(s(\cdot|X_i), s'(\cdot|X_i))\right]}$ is the tensorized Hellinger distance.
- **Assumption** (H)**:** for all model S_m, there is a non decreasing $\phi_m(\delta)$ such that $\delta \mapsto \frac{1}{\delta}\phi_m(\delta)$ is non increasing $(0, +\infty)$ and such that for all $\sigma \in \mathbb{R}^+$ and all $s_m \in S_m$

$$\int_0^\sigma \sqrt{H_{[\cdot],d^{\otimes_n}}(\epsilon, S_m)}\, d\epsilon \leq \phi_m(\sigma),$$

- **Complexity** $\mathfrak{D}_m$ def. as $n\sigma_m^2$ with σ_m unique root of $\phi_m(\sigma) = \sqrt{n}\sigma^2$.
- *Key*: **Dudley type integral** and optimization of a deviation bound.
- Typically, $H_{[\cdot],d^{\otimes_n}}(\epsilon, S_m)) \sim \dim S_m(C + \log 1/\epsilon)$ which implies $\mathfrak{D}_m \propto (\ln n) \dim(S_m)$...

Gaussian regression mixtures

- **Model** $S_m = \{s_\theta, \theta \in \Theta_m\}$ with $\Theta_m = \{K\} \otimes \Upsilon_K \otimes V_K \otimes W_k$:
  - Υ_K and V_K: sets for the K-tuple of regressions functions and covariance matrices functions.
  - W_K: sets for for the K-tuple of weights functions.
- **Structural assumptions:**
  - V_K is a set of covariance matrices independent of the covariate,
  - Υ_K and W_K are such that

$$H_{\max_{k=1}^K \sup_x \|\cdot_k(x)\|}(\delta, W_K) \leq \dim(W_K)\left(C_W + \ln\frac{1}{\delta}\right)$$

$$H_{\max_{k=1}^K \sup_x \|\cdot_k(x)\|_2}(\delta, \Upsilon_K) \leq \dim(\Upsilon_K)\left(C_\Upsilon + \ln\frac{1}{\delta}\right)$$

- Satisfied for instance if Υ_K and W_K are K-tuples of **polynomials with bounded coefficients** and x is bounded.
- **Th:** Under this assumption, if $\mathrm{pen}(m) = \kappa(C + \ln n)\dim(S_m)$ then

$$\mathbb{E}\left[JKL_\rho^{\otimes n}(s_0, \widehat{s}_{\widehat{m}})\right] \leq C_1 \inf_{m\in\mathcal{M}}\left(\inf_{s_m\in S_m} KL^{\otimes n}(s_0, s_m) + \frac{\mathrm{pen}(m)}{n}\right) + \frac{C_2}{n}$$

- **Key:** upper bound of the bracketing entropy $H_{[\cdot],d^{\otimes n}}(\epsilon, S_m)$.

Bracketing entropy decomposition

- **Model:**

$$S_m = \{\sum_{k=1}^{K} \pi_{w,k}(x)\Phi_{\upsilon_k(x),\Sigma_k(x)}(y), (K, \upsilon, \Sigma, w) \in \Theta_m\}$$

with $\Theta_m = \{K\} \otimes \Upsilon_K \otimes V_K \otimes W_k$

- **Weight and regression models:**

$$\mathcal{W}_K = \left\{ (\pi_{w,k}(x))_{k=1}^{K}, w \in W_K \right\}$$

$$\mathcal{R}_K = \left\{ \left(\Phi_{\upsilon_k(x),\Sigma_k(x)}(y) \right)_{k=1}^{K}, (\upsilon, \Sigma) \in \Upsilon_K \times V_K \right\}$$

- **Splitting properties:**

$$H_{[.],d^{\otimes n}}(\delta, S_m) \leq H_{[.],\sup_x \max_k d}\left(\frac{\delta}{5}, \mathcal{R}_K\right) + H_{[.],\sup_x d}\left(\frac{\delta}{5}, \mathcal{W}_K\right)$$

Bracketing entropy decomposition

- **Gaussian K-tuple bracketing entropy:**

$$H_{[.],\sup_x \max_k d}\left(\frac{\delta}{5}, \mathcal{R}_K\right) \le H_{\max_{k=1}^K \sup_x \|\cdot_k(x)\|_2}(\epsilon_1\delta, \Upsilon_K) + H_d(\epsilon_2\delta, V_K)$$

$$\le (\dim(\Upsilon_K) + \dim(V_K))\left(C_{\mathcal{R}} + \ln\frac{1}{\delta}\right)$$

- **Logistic weight K-tuple bracketing entropy:**

$$H_{[.],\sup_x d}\left(\frac{\delta}{5}, \mathcal{W}_K\right) \le H_{\max_k \sup_x \|\cdot_k\|}\left(\frac{\epsilon_3\delta}{\sqrt{K}}, W_K\right) \le \dim(W_K)\left(C_{\mathcal{W}} + \ln\frac{1}{\delta}\right)$$

- **Bracketing entropy bound** if $K \le K_{\max}$:

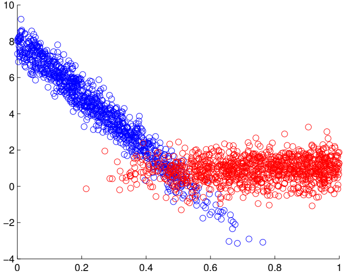
$$H_{[.],d^{\otimes n}}(\delta, S_m) \le H_{[.],\sup_x \max_k d}\left(\frac{\delta}{5}, \mathcal{R}_K\right) + H_{[.],\sup_x d}\left(\frac{\delta}{5}, \mathcal{W}_K\right)$$

$$\le (\dim(\Upsilon_K) + \dim(V_K) + \dim(W_K))\left(C + \ln\frac{1}{\delta}\right)$$

$$\le \dim(S_m)\left(C + \ln\frac{1}{\delta}\right)$$

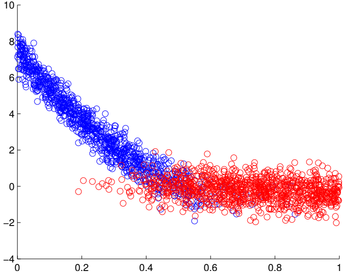
Numerical experiments

❶ $s_0 \in (S_m)_{m \in \mathcal{M}}$



Well specified

❷ $s_0 \notin (S_m)_{m \in \mathcal{M}}$



Misspecified

2 000 points

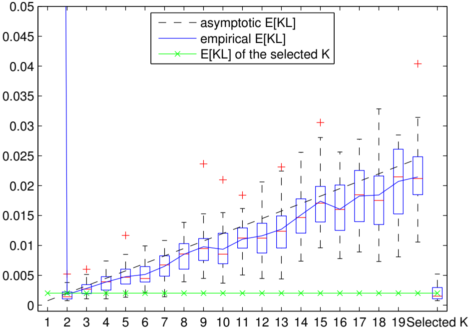
- Models S_m used:
  - Affine models for the weights and the regressions:

$$\Upsilon_K = W_K = \{(a_k x + b_k)_{k=1}^{K}, (a, b) \in \mathbb{R}^{K \times 2}\}$$

  - Free variance: $V_K = \mathbb{R}_+^K$
  - Only choice is the number of components K

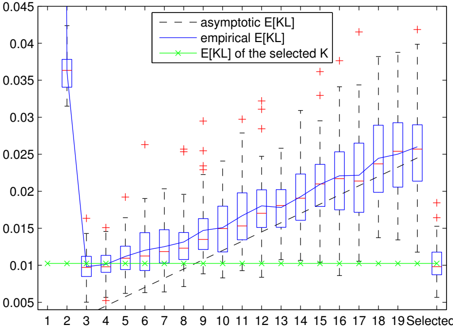
KL risk
2 000 points

❶ $s_0 \in (S_m)_{m \in \mathcal{M}}$



Well specified

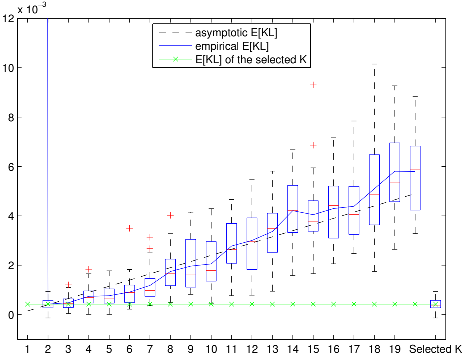
❷ $s_0 \notin (S_m)_{m \in \mathcal{M}}$



Misspecified

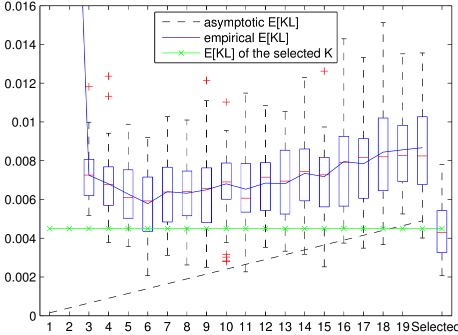
KL risk
10 000 points

❶ $s_0 \in (S_m)_{m \in \mathcal{M}}$



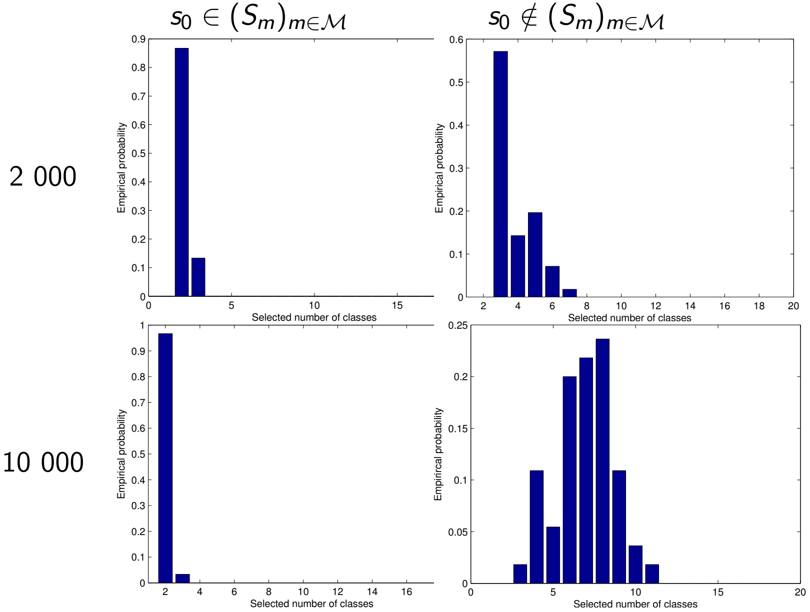
Well specified

❷ $s_0 \notin (S_m)_{m \in \mathcal{M}}$



Misspecified

Histograms of the selected K



Numerical optimization

- **Model** $S_m = \{s_\theta, \theta \in \Theta_m\}$ with $\Theta_m = \{K\} \otimes \Upsilon_K \otimes V_K \otimes W_k$:
  - K: number of components.
  - Υ_K and V_K: sets for the K-tuple of regressions functions and covariance matrices functions.
  - W_K: sets for the K-tuple of weights functions.
- **Maximum likelihood estimation:**

$$\widehat{s}_m = \underset{\theta \in \Theta_m}{\operatorname{argmin}} - \sum_{i=1}^{N} \ln s_\theta(Y_i|X_i)$$

- **Penalized model selection:**

$$\widehat{m} = \underset{m}{\operatorname{argmin}} - \sum_{i=1}^{N} \ln \widehat{s}_m(Y_i|X_i) + \kappa \dim \Theta_m$$

- Model selection computed by **exhaustive exploration.**
- Focus on maximum likelihood estimation!

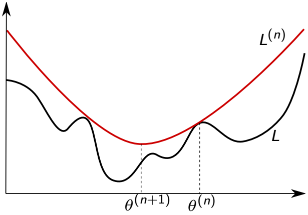
Maximum likelihood estimation

- **Model** $S_m = \{s_\theta, \theta \in \Theta_m\}$ with $\Theta_m = \{K\} \otimes \Upsilon_K \otimes V_K \otimes W_k$:
  - K: number of components.
  - Υ_K and V_K: sets for the K-tuple of regressions functions and covariance matrices functions.
  - W_K: sets for the K-tuple of weights functions.
- **Maximum likelihood estimation:**

$$\widehat{s}_m = \underset{\theta \in \Theta_m}{\operatorname{argmin}} \underbrace{- \sum_{i=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_{w,k}(X_i) \Phi_{\upsilon_k(X_i), \Sigma_k(X_i)}(Y_i) \right)}_{L(\theta)}$$

- **Non convex** minimization problem!
- **Majorization/Minimization** approach

MM approach



- **Iterative** approach to minimize $L(\theta)$ by minimizing a **sequence of (convex) proxies** of L.
- **Majorization/Minimization:**
  - **Current estimate** of the minimizer: $\theta^{(n)}$
  - Construction of a **Majorization** $L^{(n)}$ of L such that $L^{(n)}(\theta^{(n)}) = L(\theta^{(n)})$ with $L^{(n)}$ easy to minimize (convex for example).
  - Computation of a **Minimizer**

$$\theta^{(n+1)} = \operatorname{argmin} L^{(n)}(\theta)$$

- By construction, $L(\theta^{(n+1)}) \leq L(\theta^{(n)})$!
- Very **generic methodology**...
- Minimization can be replaced by a diminution...

Maximum Likelihood and EM

- Back to **our** maximum likelihood:

$$L(\theta) = L(K, \upsilon, \Sigma, w) = -\sum_{i=1}^{N} \ln \left(\sum_{k=1}^{K} \pi_{w,k}(X_i) \Phi_{\upsilon_k(X_i),\Sigma_k(X_i)}(Y_i) \right)$$

- **EM:** specific case of MM for this type of mixture.
- **(Conditional) Expectation:** at step n, we let

$$P_k^{i,(n)} = P\left(k_i = k \middle| X_i, Y_i, K, \upsilon^{(n)}, \Sigma^{(n)}, w^{(n)}\right)$$

and $\quad L^{(n)}(K, \upsilon, \Sigma, w) = -\sum_{i=1}^{N} \sum_{k=1}^{K} P_k^{i,(n)} \ln \left(\pi_{w,k}(X_i) \Phi_{\upsilon_k(X_i),\Sigma_k(X_i)}(Y_i)\right).$

- **Maj. prop.:** $L \leq L^{(n)} + \mathrm{Cst}^{(n)}$ with equ. at $\theta = (K, \upsilon^{(n)}, \Sigma^{(n)}, w^{(n)})$
- **Separability** in $(\upsilon^{(n)}, \Sigma^{(n)})$ and $w^{(n)}$:

$$L^{(n)}(K, \upsilon, \Sigma, w) = \left(-\sum_{i=1}^{N} \sum_{k=1}^{K} P_k^{i,(n)} \ln \Phi_{\upsilon_k(X_i),\Sigma_k(X_i)}(Y_i) \right)$$

$$+ \left(-\sum_{i=1}^{N} \sum_{k=1}^{K} P_k^{i,(n)} \ln \pi_{w,k}(X_i) \right)$$

Minimization of $L^{(n)}$

- **Separability** in $(\upsilon^{(n)}, \Sigma^{(n)})$ and $w^{(n)}$:

$$L^{(n)}(K, \upsilon, \Sigma, w) = \left(-\sum_{i=1}^{N} \sum_{k=1}^{K} P_k^{i,(n)} \ln \Phi_{\upsilon_k(X_i),\Sigma_k(X_i)}(Y_i) \right) + \left(-\sum_{i=1}^{N} \sum_{k=1}^{K} P_k^{i,(n)} \ln \pi_{w,k}(X_i) \right)$$

- For the **regression parameters** $(\upsilon^{(n)}, \Sigma^{(n)})$:
  - K weighted linear regressions: $-\sum_{i=1}^{N} P_k^{i,(n)} \ln \Phi_{\upsilon_k(X_i),\Sigma_k(X_i)}(Y_i)$
  - Explicit formulas!
- For the **weight parameters** $w^{(n)}$:
  - Single K modality logistic regression: $-\sum_{i=1}^{N} \sum_{k=1}^{K} P_k^{i,(n)} \ln \pi_{w,k}(X_i)$
  - Iterative minimization scheme (Newton = Iterative Reweighted Least Square)

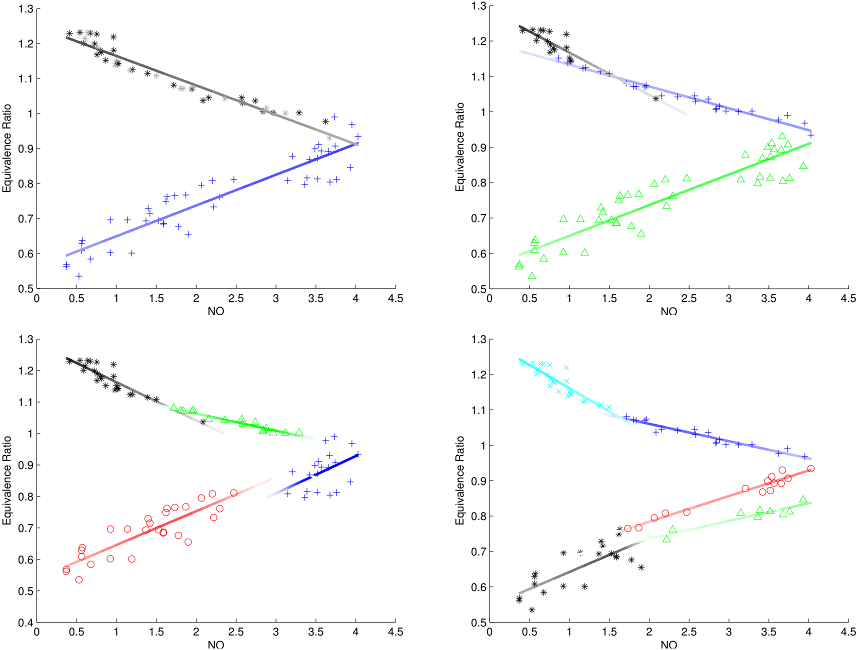
Initialization

- **Very important issue!**
- For the **weights:** initialization to uniform weights seems sufficient.
- For the **means:**
  - Comparison between **several** strategies
    - Naive purely random initialization
    - Small-EM: Random initialization followed by a few minimization steps and selection
    - Advanced Small-EM: Initialization based on a first $2D$ clustering followed by a few minimization steps and selection
    - Advanced Small-EM 2 : Initialization based on a random drawing of lines between points clustering followed by a few minimization steps and selection
  - **Criterion:** lowest likelihood for a given amount of time!
  - Similar results in term on expectation but **different behaviors in term of dispersion**:
    - Too simple strategies fail sometimes to provide a satisfactory answer while too complex ones may not explore sufficient local maxima.
    - **Winner:** Advanced Small-EM 2 with 3 minimizations steps and 50 candidates.

Newton-EM Algorithm

- **Initialization with Advanced Small-EM 2:**
  - Initialization based on a random drawing of lines between points clustering
  - 3 minimizations steps
  - selection among 50 candidates
- **Iterate** until convergence:
  - **Newton steps** over weights $w^{(n)}$ if the likelihood increases (up to 5 times)
  - K **linear regressions** to update mean and variance $(\upsilon^{(n)}, \Sigma^{(n)})$ in each class
- **Note:** Initialization issues in high dimension with this scheme!

Numerical results



Penalization strategy

- **Penalized model selection:**

$$\widehat{m} = \operatorname*{argmin}_{m \in \mathcal{M}} \sum_{k=1}^{K} - \ln \widehat{s}_m(Y_i|X_i) + \mathrm{pen}(m)$$

- **Theoretical analysis:**

$$\mathrm{pen}(m) = \kappa(C + \ln n) \dim \Theta_m$$

- κ and C are only **loosely upper bounded!**
- **In practice,** use $\mathrm{pen}(m) = \kappa \dim \Theta_m$ with κ chosen appropriately.
- **Classical choice:**
  - BIC: $\kappa = \log n/2$
  - AIC: $\kappa = 1$
- Here: **Jump/slope heuristic** = data driven choice of κ

Ideal penalty

- By definition:

$$KL^{\otimes n}(s_0, \widehat{s}_m) = P_n^{\otimes n}\left(-\ln \frac{\widehat{s}_m}{s_0}\right) \underbrace{-\nu_n^{\otimes n}\left(-\ln \frac{\widehat{s}_m}{s_0}\right)}_{\mathrm{pen}_{\mathrm{id}}(m)/n}$$

- With the ***ideal* penalty** $\mathrm{pen}_{\mathrm{id}}(m)$:

$$\begin{aligned} KL^{\otimes n}(s_0, \widehat{s}_{\widehat{m}}) &= P_n^{\otimes n}\left(-\ln \frac{\widehat{s}_{\widehat{m}}}{s_0}\right) + \frac{\mathrm{pen}_{\mathrm{id}}(\widehat{m})}{n} \\ &\leq \inf_m P_n^{\otimes n}\left(-\ln \frac{\widehat{s}_m}{s_0}\right) + \frac{\mathrm{pen}_{\mathrm{id}}(m)}{n} \leq \inf_m KL^{\otimes n}(s_0, \widehat{s}_m) \\ &\leq \inf_m \left(KL^{\otimes n}(s_0, \widetilde{s}_m) + \left(KL^{\otimes n}(s_0, \widehat{s}_m) - KL^{\otimes n}(s_0, \widetilde{s}_m)\right)\right) \end{aligned}$$

- **Ideal penalty oracle inequality:**

$$\mathbb{E}\left[KL^{\otimes n}(s_0, \widehat{s}_{\widehat{m}})\right] \leq \inf_{S_m \in \mathcal{S}} \left(\underbrace{KL^{\otimes n}(s_0, \widetilde{s}_m)}_{\text{Bias term}} + \underbrace{\mathbb{E}\left[KL^{\otimes n}(s_0, \widehat{s}_m) - KL^{\otimes n}(s_0, \widetilde{s}_m)\right]}_{\text{Variance term}} \right)$$

Jump/Slope heuristic

- **Ideal penalty decomposition:**

$$\frac{\mathrm{pen}_{\mathrm{id}}(m)}{n} = -\nu_n^{\otimes n}\left(-\ln \frac{\widehat{s}_{\widehat{m}}}{s_0}\right)$$
$$= \nu_n^{\otimes n}\left(-\ln \frac{\widetilde{s}_m}{\widehat{s}_m}\right) - \nu_n^{\otimes n}\left(-\ln \widetilde{s}_m\right) + \underbrace{\nu_n^{\otimes n}(-\log s_0)}_{\text{independent of } m} .$$

- **Jump/Slope heuristic:**
  - Concentration: $\nu_n^{\otimes n}\left(-\ln \widetilde{s}_m\right) \ll \nu_n^{\otimes n}\left(-\ln \frac{\widetilde{s}_m}{\widehat{s}_m}\right)$
  - Symmetry: $P_n^{\otimes n}\left(-\ln(\widetilde{s}_m / \widehat{s}_m)\right) \sim P^{\otimes n}\left(-\ln(\widehat{s}_m / \widetilde{s}_m)\right)$
- Resulting **approximation**:

$$\frac{\mathrm{pen}_{\mathrm{id}}(m)}{n} \sim 2P_n^{\otimes n}\left(-\ln \frac{\widetilde{s}_m}{\widehat{s}_m}\right) \underbrace{-P^{\otimes n}(-\log s_0)}_{\text{independent of } m} .$$

- $P_n^{\otimes n}\left(-\ln \frac{\widetilde{s}_m}{\widehat{s}_m}\right)$ has still **to be estimated**!

Minimal penalty

- If $\mathrm{pen}(m) = \kappa P_n^{\otimes_n}\left(-\ln \frac{\widetilde{s}_m}{\widehat{s}_m}\right)$ then

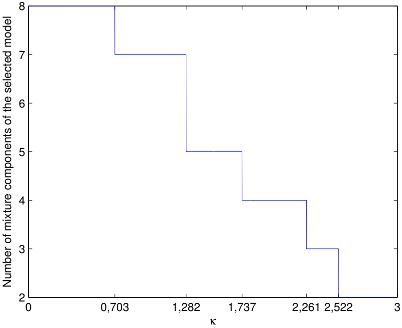
$$P_n^{\otimes_n}(-\ln \widehat{s}_m) + \mathrm{pen}(m) = (1-\kappa)P_n^{\otimes_n}(-\ln \widehat{s}_m) + \kappa P_n^{\otimes_n}(-\ln \widetilde{s}_m)$$

- **No tradeoff is $\kappa < 1$!**
- **Minimal penalty:** $\mathrm{pen}_{\min}(m) = P_n^{\otimes_n}\left(-\ln(\widetilde{s}_m/\widehat{s}_m)\right)$
- Jump/Slope heuristic strongest assumption: **parametric approximation** of $\mathrm{pen}_{\min}$

$$\mathrm{pen}_{\min}(m) = \mathrm{pen}(\kappa, m)$$

  where pen *shape* is given by the theoretical study!
- **Simplest case:** $\mathrm{pen}(\kappa, m) = \kappa \dim S_m$.

Jump heuristic

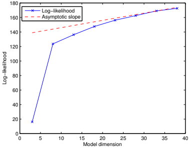


- Minimal penalty for which there is a **tradeoff**:

$$\mathrm{pen}_{\mathrm{min}}(m) = \mathrm{pen}(\kappa, m)$$

- Compute the models selected for several κ and detect **a jump in the model dimensions**.
- Not always a clear single jump...

Slope heuristic



- **Observation:**

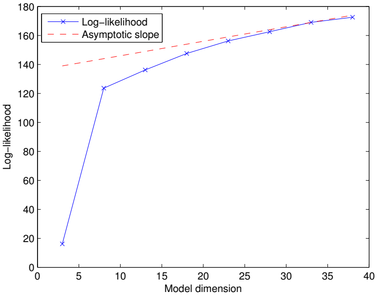
$$\mathrm{pen}_{\mathrm{id}}(m) = P_n^{\otimes n}\left(-\ln \frac{\widetilde{s}_m}{\widehat{s}_m}\right) = P_n^{\otimes n}\left(\ln \widehat{s}_m\right) + P_n^{\otimes n}\left(-\ln \widetilde{s}_m\right)$$

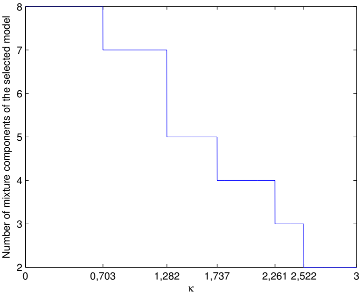
- If the model are more and more complex, one may expect that the ***projection* bias converges to a constant**: $P_n^{\otimes n}\left(-\ln \widetilde{s}_m\right) \sim C$
- This implies $\mathrm{pen}_{\mathrm{id}}(m) \sim P_n^{\otimes n}\left(\ln \widehat{s}_m\right) + C$
- If $\mathrm{pen}_{\mathrm{id}}(indm) = \mathrm{pen}(\kappa, m)$ then κ can be estimated by a **regression** as

$$\mathrm{pen}(\kappa, m) - C \sim \underbrace{P_n^{\otimes n}\left(\ln \widehat{s}_m\right)}_{\text{data driven}}$$

- If $\mathrm{pen}(\kappa, m) = \kappa \dim S_m$, κ measures the **slope** of $P_n^{\otimes n}\left(\ln \widehat{s}_m\right)$ with respect to $\dim S_m$.

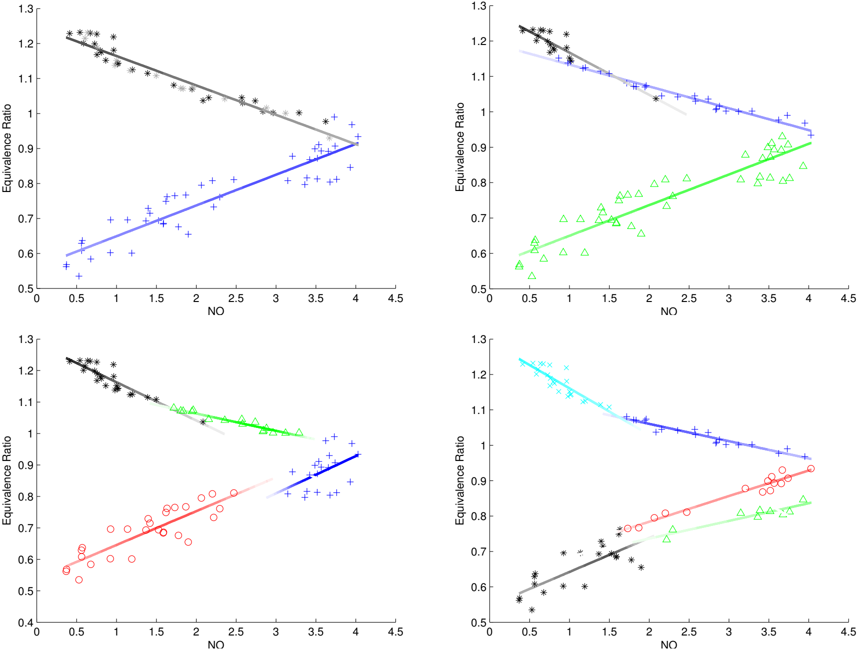
Slope heuristic



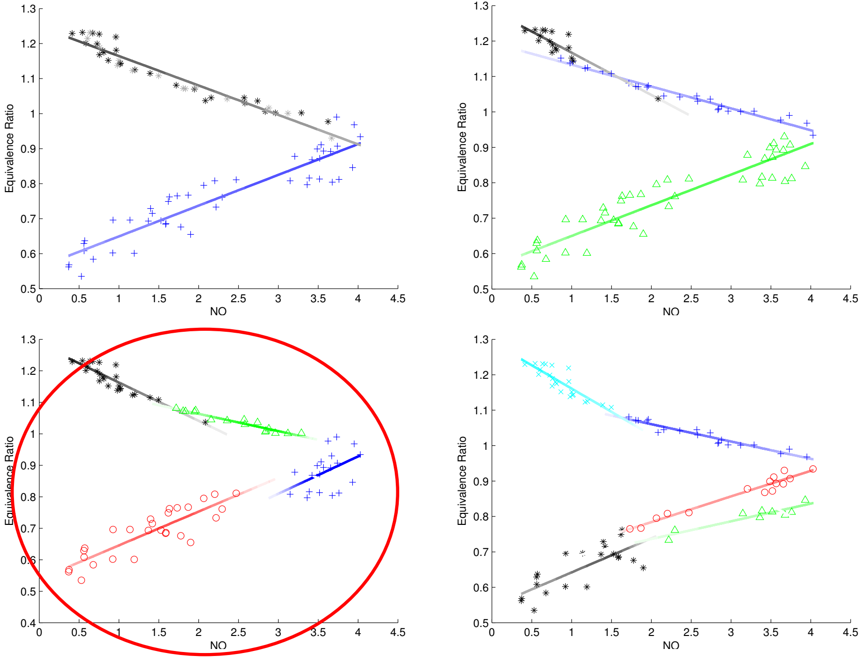


- **Slope heuristic** with $\mathrm{pen}(\kappa, m) = \kappa \dim S_m$: $\kappa \sim 1$
- **Resulting penalties:**
  - Slope heuristic: $\mathrm{pen}(m) = 2 \dim(S_m)$
  - BIC: $\mathrm{pen}(m) = 2{,}23 \dim(S_m)$
  - AIC: $\mathrm{pen}(m) = 2 \dim(S_m)$
- **Selected number of clusters:**
  - Slope heuristic: 4
  - BIC: 4
  - AIC: 7!

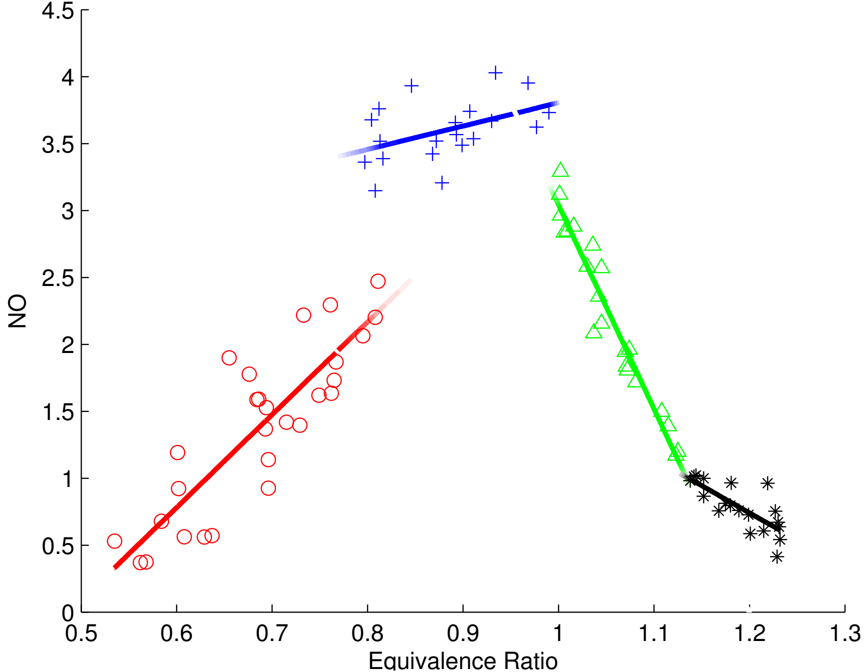
Numerical results



Numerical results



Validation?


4.5
4
3.5
3
2.5
2
1.5
1
0.5
0
NO
0.5
0.6
0.7
0.8
0.9
1
1.1
1.2
1.3
Equivalence Ratio

Conclusion

- **Framework:**
  - Mixture of regressions.
  - Proposed tool: Mixture of Gaussian regressions with logistic mixing weights.
  - Penalized maximum likelihood conditional density estimation.
- **Contributions:**
  - Theoretical guarantee for the conditional density estimation problem.
  - Efficient minimization algorithm.
  - Numerical penalty calibration.
- **Perspectives:**
  - Proof for penalty calibration by slope heuristic.
  - Enhanced Spatialized Gaussian Mixture Model with piecewise logistic weights (S. Cohen).