Model selection by penalization in mixture of experts models with a non-asymptotic approach

TrungTin Nguyen

53èmes Journées de Statistique
Lyon, France

Outline and our contributions

1. Collection of GLoME models
   - Context and motivating example
   - Boundedness conditions
2. Model selection in GLoME models
   - Asymptotic approach
   - Non-asymptotic approach with oracle inequality
3. Main positive messages and perspectives

Outline

Outline

Context

- **We have**: n random samples $(\mathbf{X}_i, \mathbf{Y}_i)_{i\in[n]} \in (\mathcal{X} \times \mathcal{Y})^n$ with observed values $(\mathbf{x}_i, \mathbf{y}_i)_{i\in[n]}$, $[n] = \{1, \ldots, n\}$, arising from an unknown conditional density s_0.
- **Learning**: Regression analysis + Clustering + Model selection (*e.g.*, number of clusters, complexity in each cluster).
- **Our proposal**: using **mixture of experts (MoE[1])** models due to their flexibility and effectiveness (several universal approximation theorems [2] [3] [4]).

[1] Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. Neural computation.

[2] Nguyen, H. D., Chamroukhi, F., and Forbes, F. (2019). Approximation results regarding the multiple-output Gaussian gated mixture of linear experts model. Neurocomputing.

[3] Nguyen, H. D., **Nguyen, T.**, Chamroukhi, F., and McLachlan, G. J. (2021). Approximations of conditional probability density functions in Lebesgue spaces via mixture of experts models. *Journal of Statistical Distributions and Applications.*

[4] **Nguyen, T.**, Chamroukhi, F., Nguyen, H. D., and McLachlan, G. J. (2022). Approximation of probability density functions via location-scale finite mixtures in Lebesgue spaces. *Communications in Statistics - Theory and Method.*

Context

- **We have**: n random samples $(\mathbf{X}_i, \mathbf{Y}_i)_{i\in[n]} \in (\mathcal{X} \times \mathcal{Y})^n$ with observed values $(\mathbf{x}_i, \mathbf{y}_i)_{i\in[n]}$, $[n] = \{1, \ldots, n\}$, arising from an unknown conditional density s_0.
- **Learning**: Regression analysis + Clustering + Model selection (*e.g.*, number of clusters, complexity in each cluster).
- **Our proposal**: using **mixture of experts (MoE[1])** models due to their flexibility and effectiveness (several universal approximation theorems [2] [3] [4]).

---

[1] Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. Neural computation.

[2] Nguyen, H. D., Chamroukhi, F., and Forbes, F. (2019). Approximation results regarding the multiple-output Gaussian gated mixture of linear experts model. Neurocomputing.

[3] Nguyen, H. D., **Nguyen, T.**, Chamroukhi, F., and McLachlan, G. J. (2021). Approximations of conditional probability density functions in Lebesgue spaces via mixture of experts models. *Journal of Statistical Distributions and Applications.*

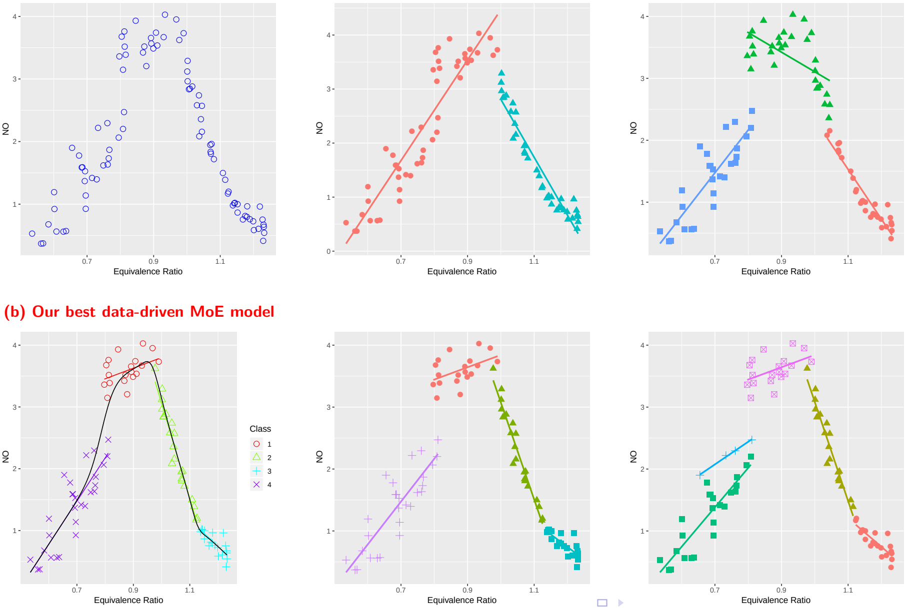
[4] **Nguyen, T.**, Chamroukhi, F., Nguyen, H. D., and McLachlan, G. J. (2022). Approximation of probability density functions via location-scale finite mixtures in Lebesgue spaces. *Communications in Statistics - Theory and Method.*

Motivating example: Ethanol data set 88 observations

(a) Raw Ethanol data set

Collection of MoE models with linear mean functions characterized by 2-5 clusters

(b) Our best data-driven MoE model



Definition: Gaussian-gated Localized MoE (GLoME) models

$$s_{\psi_{K,d}}(\mathbf{x}|\mathbf{y}) = \sum_{k=1}^{K} \underbrace{\frac{\boldsymbol{\pi}_k \mathcal{N}_L(\mathbf{y}; \mathbf{c}_k, \boldsymbol{\Gamma}_k)}{\sum_{j=1}^{K} \boldsymbol{\pi}_j \mathcal{N}_L(\mathbf{y}; \mathbf{c}_j, \boldsymbol{\Gamma}_j)}}_{\text{Gaussian gating network}} \underbrace{\mathcal{N}_D(\mathbf{x}; \boldsymbol{\upsilon}_{k,d}(\mathbf{y}), \boldsymbol{\Sigma}_k)}_{\text{Gaussian expert}}.$$

- $\boldsymbol{\omega} = (\boldsymbol{\pi}, \boldsymbol{c}, \boldsymbol{\Gamma}) \in (\boldsymbol{\Pi}_{K-1} \times \mathbf{C}_K \times \boldsymbol{V}_K') = \boldsymbol{\Omega}_K$, $\boldsymbol{\Pi}_{K-1}$: probability simplex, $K \in \mathbb{N}^\star$: number of mixture components.
- $d \in \mathbb{N}^\star$: mean functions' hyperparameter *e.g.,* degree of polynomial.
- $\psi_{K,d} = (\boldsymbol{\omega}, \boldsymbol{\upsilon}, \boldsymbol{\Sigma}) \in \boldsymbol{\Omega}_K \times \boldsymbol{\Upsilon}_{K,d} \times \mathbf{V}_K$: model parameter.

High-dimensional data using inverse regression frameworks (Gaussian Locally-Linear Mapping (GLLiM[5]) models): $\mathbf{Y} \equiv$ input, $\mathbf{X} \equiv$ output, $\mathcal{X} \subset \mathbb{R}^D$, $\mathcal{Y} \subset \mathbb{R}^L$, with $D \gg L$ and $D, L \in \mathbb{N}^\star$.

[5] Deleforge, A., Forbes, F., and Horaud, R. (2015). High-dimensional regression with gaussian mixtures and partially-latent response variables. Statistics and Computing.

Outline

Mild assumption: Boundedness conditions

- Gaussian gating parameters: there exist positive constants $a_{\pi}, A_{c}, a_{\Gamma}, A_{\Gamma}$ s.t.

$$\widetilde{\Omega}_K = \{\omega \in \Omega_K : \forall k \in [K], \|\mathbf{c}_k\|_\infty \leq A_{c},$$
$$a_{\Gamma} \leq m(\Gamma_k) \leq M(\Gamma_k) \leq A_{\Gamma}, a_{\pi} \leq \pi_k\}.$$

- Gaussian mean experts: linear combination of bounded basis functions: $\upsilon = (\upsilon_{k,d})_{k\in[K]} \in \Upsilon_{K,d} = \otimes_{k\in[K]} \Upsilon_{k,d} = \Upsilon_{k,d}^K$, where $\forall k \in [K]$,

$$\Upsilon_{k,d} = \Upsilon_{Bo,d} = \left\{ \mathbf{y} \mapsto \left(\sum_{i=1}^{d} \alpha_i^{(j)} \theta_{\Upsilon,i}(\mathbf{y}) \right)_{j\in[D]} : \|\alpha\|_\infty \leq T_\Upsilon \right\},$$

Collection of bounded basis functions: $\mathbf{y} \mapsto (\theta_{\Upsilon,i}(\mathbf{y}))_{i\in[d_\Upsilon]}$, $d \in \mathbb{N}^\star$, $T_\Upsilon \in \mathbb{R}^+$.

Classical covariance matrix parameterization[6]

Boundedness conditions on Gaussian expert covariance matrices

$$\mathbf{V}_K = \left\{ (\boldsymbol{\Sigma}_k)_{k\in[K]} \equiv \left(B_k \mathbf{P}_k \mathbf{A}_k \mathbf{P}_k^\top\right)_{k\in[K]} : B_- \leq B_k \leq B_+, \mathbf{P}_k \in SO(D), \mathbf{A}_k \in \mathcal{A}(\lambda_-, \lambda_+) \right\}:$$

- $B_k = |\boldsymbol{\Sigma}_k|^{1/D}$: volume, $B_- \in \mathbb{R}^+$, $B_+ \in \mathbb{R}^+$,
- $\mathbf{P}_k$: eigenvectors of $\boldsymbol{\Sigma}_k$, $SO(D)$: special orthogonal group of dimension D,
- $\mathbf{A}_k$: diagonal matrix of normalized eigenvalues of $\boldsymbol{\Sigma}_k$, $\mathcal{A}(\lambda_-, \lambda_+)$: diagonal matrices $\mathbf{A}_k$, such that $|\mathbf{A}_k| = 1$ and $\forall i \in [D], \lambda_- \leq (\mathbf{A}_k)_{i,i} \leq \lambda_+$, where $\lambda_-, \lambda_+ \in \mathbb{R}$.

[6] Celeux, G. and Govaert, G. (1995). Gaussian parsimonious clustering models. Pattern Recognition.

Definition: Collection of GLoME models

$$S_{\mathbf{m}} = \Big\{ (\mathbf{x}, \mathbf{y}) \mapsto s_{\psi_{K,d}}(\mathbf{x}|\mathbf{y}) = s_{\mathbf{m}}(\mathbf{x}|\mathbf{y}) : \mathbf{m} = (K, d),$$
$$\psi_{K,d} = (\boldsymbol{\omega}, \boldsymbol{\upsilon}, \boldsymbol{\Sigma}) \in \widetilde{\boldsymbol{\Omega}}_K \times \boldsymbol{\Upsilon}_{K,d} \times \mathbf{V}_K = \widetilde{\boldsymbol{\Psi}}_{K,d} \Big\}.$$

- $\mathbf{m} \in \mathcal{M} = [K_{\max}] \times [d_{\max}], K_{\max}, d_{\max} \in \mathbb{N}^{\star}$.

Outline

Definition: Penalized maximum likelihood estimator (PMLE)

$\widehat{s}_{\widehat{\mathbf{m}}}$: an η'-**PMLE** (corresponding **the selected model or best data-driven model** $S_{\widehat{\mathbf{m}}}$ among $(S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$), defined by

$$\sum_{i=1}^{n} -\ln\left(\widehat{s}_{\widehat{\mathbf{m}}}\left(\mathbf{x}_i|\mathbf{y}_i\right)\right) + \mathrm{pen}\left(\widehat{\mathbf{m}}\right) \leq \inf_{\mathbf{m}\in\mathcal{M}} \left(\sum_{i=1}^{n} -\ln\left(\widehat{s}_{\mathbf{m}}\left(\mathbf{x}_i|\mathbf{y}_i\right)\right) + \mathrm{pen}(\mathbf{m})\right) + \eta',$$

- $\widehat{s}_{\mathbf{m}}$: an η-minimizer of the negative log-likelihood (infimum may not be reached) is defined by

$$\sum_{i=1}^{n} -\ln\left(\widehat{s}_{\mathbf{m}}\left(\mathbf{x}_i|\mathbf{y}_i\right)\right) \leq \inf_{s_{\mathbf{m}}\in S_{\mathbf{m}}} \sum_{i=1}^{n} -\ln\left(s_{\mathbf{m}}\left(\mathbf{x}_i|\mathbf{y}_i\right)\right) + \eta,$$

- $\mathrm{pen}(\mathbf{m})$: penalty function $\leftarrow$ trade-off between good data fit and model complexity.

Definition: Loss functions for conditional densities

- Tensorized Kullback-Leibler divergence $\mathrm{KL}^{\otimes \mathrm{n}}$ **(conditional densities and random covariate variables)**:

$$\mathrm{KL}^{\otimes \mathrm{n}}(s,t) = \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{KL}\left(s\left(\cdot|\mathbf{Y}_i\right), t\left(\cdot|\mathbf{Y}_i\right)\right)\right],$$

if $sdy \ll tdy$, $+\infty$ otherwise. Fixed predictors $\Rightarrow$ no $\mathbb{E}_{\mathbf{Y}_{[n]}}[\cdot]$.

- Tensorized Jensen-Kullback-Leibler divergence $\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}$ **(technical difficulties with conditional densities)**: given $\rho \in (0,1)$,

$$\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}(s,t) = \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\rho}\mathrm{KL}\left(s\left(\cdot|\mathbf{Y}_i\right), (1-\rho)\, s\left(\cdot|\mathbf{Y}_i\right) + \rho t\left(\cdot|\mathbf{Y}_i\right)\right)\right].$$

Definition: Loss functions for conditional densities

- Tensorized Kullback-Leibler divergence $\mathrm{KL}^{\otimes \mathrm{n}}$ **(conditional densities and random covariate variables)**:

$$\mathrm{KL}^{\otimes \mathrm{n}}(s,t) = \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^{n}\mathrm{KL}\left(s\left(\cdot|\mathbf{Y}_i\right),t\left(\cdot|\mathbf{Y}_i\right)\right)\right],$$

  if $sdy \ll tdy$, $+\infty$ otherwise. Fixed predictors $\Rightarrow$ no $\mathbb{E}_{\mathbf{Y}_{[n]}}[\cdot]$.

- Tensorized Jensen-Kullback-Leibler divergence $\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}$ **(technical difficulties with conditional densities)**: given $\rho \in (0,1)$,

$$\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}(s,t) = \mathbb{E}_{\mathbf{Y}_{[n]}}\left[\frac{1}{n}\sum_{i=1}^{n}\frac{1}{\rho}\mathrm{KL}\left(s\left(\cdot|\mathbf{Y}_i\right),(1-\rho)\,s\left(\cdot|\mathbf{Y}_i\right)+\rho t\left(\cdot|\mathbf{Y}_i\right)\right)\right].$$

Outline

- Akaike information criterion (AIC) [Akaike, 1974], Bayesian information criterion (BIC) [Schwarz et al., 1978] and BIC-like approximation of integrated classification likelihood (ICL-BIC) [Biernacki et al., 2000] criteria:

$$\text{pen}_{\text{AIC}}(\mathbf{m}) = \dim(S_{\mathbf{m}}), \quad \text{pen}_{\text{BIC}}(\mathbf{m}) = \frac{\ln(n)\dim(S_{\mathbf{m}})}{2}.$$

$$\text{pen}_{\text{ICL-BIC}}(\mathbf{m}) = \text{pen}_{\text{BIC}}(\mathbf{m}) + \text{ENT}(\mathbf{m}) \longleftarrow \text{ estimated mean entropy.}$$

- AIC (based on asymptotic theory), BIC, ICL-BIC (based on Bayesian approach):
  - May be wrong in a non-asymptotic context: $\dim(S_{\mathbf{m}})$ and $\text{card}(\mathcal{M})$ depend on and can be much larger than n.
  - No finite sample guarantees.
- Obtain an upper bound on $\mathbb{E}\left[\text{KL}^{\otimes n}(s_0, \widehat{s}_{\mathbf{m}})\right]$:
  - ✔ Finite sample guarantee.
  - ✖ Strong regularity assumptions of [White, 1982].

- Akaike information criterion (AIC) [Akaike, 1974], Bayesian information criterion (BIC) [Schwarz et al., 1978] and BIC-like approximation of integrated classification likelihood (ICL-BIC) [Biernacki et al., 2000] criteria:

$$\mathrm{pen}_{\mathrm{AIC}}(\mathbf{m}) = \dim(S_{\mathbf{m}}), \quad \mathrm{pen}_{\mathrm{BIC}}(\mathbf{m}) = \frac{\ln(n)\dim(S_{\mathbf{m}})}{2}.$$

$$\mathrm{pen}_{\mathrm{ICL\text{-}BIC}}(\mathbf{m}) = \mathrm{pen}_{\mathrm{BIC}}(\mathbf{m}) + \mathrm{ENT}(\mathbf{m}) \longleftarrow \text{ estimated mean entropy.}$$

- AIC (based on asymptotic theory), BIC, ICL-BIC (based on Bayesian approach):
  - May be wrong in a non-asymptotic context: $\dim(S_{\mathbf{m}})$ and $\mathrm{card}(\mathcal{M})$ depend on and can be much larger than n.
  - No finite sample guarantees.
- Obtain an upper bound on $\mathbb{E}\left[\mathrm{KL}^{\otimes n}(s_0, \widehat{s}_{\mathbf{m}})\right]$:
  - ✔ Finite sample guarantee.
  - ✖ Strong regularity assumptions of [White, 1982].

Some asymptotic approaches for model selection in MoE models

- Akaike information criterion (AIC) [Akaike, 1974], Bayesian information criterion (BIC) [Schwarz et al., 1978] and BIC-like approximation of integrated classification likelihood (ICL-BIC) [Biernacki et al., 2000] criteria:

$$\mathrm{pen}_{\mathrm{AIC}}(\mathbf{m}) = \dim(S_{\mathbf{m}}), \quad \mathrm{pen}_{\mathrm{BIC}}(\mathbf{m}) = \frac{\ln(n)\dim(S_{\mathbf{m}})}{2}.$$

$$\mathrm{pen}_{\mathrm{ICL\text{-}BIC}}(\mathbf{m}) = \mathrm{pen}_{\mathrm{BIC}}(\mathbf{m}) + \mathrm{ENT}(\mathbf{m}) \longleftarrow \text{ estimated mean entropy.}$$

- AIC (based on asymptotic theory), BIC, ICL-BIC (based on Bayesian approach):
  - May be wrong in a non-asymptotic context: $\dim(S_{\mathbf{m}})$ and $\mathrm{card}(\mathcal{M})$ depend on and can be much larger than n.
  - No finite sample guarantees.

- Obtain an upper bound on $\mathbb{E}\left[\mathrm{KL}^{\otimes n}\left(s_0, \widehat{s}_{\mathbf{m}}\right)\right]$:
  - ✔ Finite sample guarantee.
  - ✖ Strong regularity assumptions of [White, 1982].

Outline

Non-asymptotic upper bound of a single model

- **Initial target**:

$$\mathbb{E}\left[\mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_{\mathbf{m}}\right)\right] \leq \left(\inf_{\psi_{\mathbf{m}} \in \mathbf{\Psi}_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, s_{\psi_{\mathbf{m}}}\right) + \frac{1}{2n} \dim\left(S_{\mathbf{m}}\right)\right) + C_2 \frac{1}{n}.$$

- **Our contribution**:

$$\mathbb{E}\left[\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_{\mathbf{m}}\right)\right] \leq C_1 \left(\inf_{\psi_{\mathbf{m}} \in \mathbf{\Psi}_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, s_{\psi_{\mathbf{m}}}\right) + \frac{\kappa}{n} \mathfrak{D}_m\right) + C_2 \frac{1}{n}.$$

1. Different divergences: $\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_m\right) \leq \mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_m\right)$.
2. $C_1 > 1$.

Non-asymptotic upper bound of a single model

- **Initial target**:

$$\mathbb{E}\left[\mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_{\mathbf{m}}\right)\right] \leq \left(\inf_{\psi_{\mathbf{m}} \in \boldsymbol{\Psi}_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, s_{\psi_{\mathbf{m}}}\right) + \frac{1}{2n} \dim\left(S_{\mathbf{m}}\right)\right) + C_2 \frac{1}{n}.$$

- **Our contribution**:

$$\mathbb{E}\left[\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_{\mathbf{m}}\right)\right] \leq C_1 \left(\inf_{\psi_{\mathbf{m}} \in \boldsymbol{\Psi}_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, s_{\psi_{\mathbf{m}}}\right) + \frac{\kappa}{n} \mathfrak{D}_m\right) + C_2 \frac{1}{n}.$$

1. Different divergences: $\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_m\right) \leq \mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_m\right)$.
2. $C_1 > 1$.

Contribution: non-asymptotic oracle inequality[7]

Theorem: We are given: $(S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$, $\rho \in (0,1)$, $C_1 > 1$, $\Xi = \sum_{\mathbf{m}\in\mathcal{M}} e^{-z_m} < \infty, z_m \in \mathbb{R}^+, \forall m \in \mathcal{M}$. There exist constants C and $\kappa(\rho, C_1) > 0$ such that whenever for all $m \in \mathcal{M}$,

$$\mathrm{pen}(\mathbf{m}) \geq \kappa(\rho, C_1)\left[(C + \ln n)\dim(S_{\mathbf{m}}) + z_m\right],$$

the η'-PMLE $\widehat{s}_{\widehat{\mathbf{m}}}$ satisfies

$$\mathbb{E}\left[\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_{\widehat{\mathbf{m}}}\right)\right] \leq C_1 \inf_{\mathbf{m}\in\mathcal{M}}\left(\inf_{s_{\mathbf{m}}\in S_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, s_{\mathbf{m}}\right) + \frac{\mathrm{pen}(\mathbf{m})}{n}\right) + \frac{\kappa(\rho, C_1) C_1 \Xi}{n} + \frac{\eta + \eta'}{n}.$$

[7] **Nguyen, T.**, Nguyen, H.D., Chamroukhi, F., and Forbes, F. (2022). A non-asymptotic approach for model selection via penalization in high-dimensional mixture of experts. arXiv 2104.02640. Under revision, Electronic Journal of Statistics.

Outline

- **Our contributions**: establishing **non-asymptotic risk bounds** that take the form of **weak oracle inequalities**, provided that **lower bounds on the penalties** hold true.
- Partially answering important questions:
  1. **Which value of** K should be chosen, given the sample size n.
  2. Whether it is better to use a **few complex experts** or **combine many simple experts**, given the total number of parameters.
- Minimax lower bounds: only known for mixture models[8].

[8] Maugis-Rabusseau, C. and Michel, B. (2013). Adaptive density estimation for clustering with Gaussian mixtures. ESAIM: Probability and Statistics.

Main positive messages and perspectives

- **Our contributions**: establishing **non-asymptotic risk bounds** that take the form of **weak oracle inequalities**, provided that **lower bounds on the penalties** hold true.
- Partially answering important questions:
  1. **Which value of** K should be chosen, given the sample size n.
  2. Whether it is better to use a **few complex experts** or **combine many simple experts**, given the total number of parameters.
- Minimax lower bounds: only known for mixture models[8].

[8] Maugis-Rabusseau, C. and Michel, B. (2013). Adaptive density estimation for clustering with Gaussian mixtures. ESAIM: Probability and Statistics.

Main positive messages and perspectives

- **Our contributions**: establishing **non-asymptotic risk bounds** that take the form of **weak oracle inequalities**, provided that **lower bounds on the penalties** hold true.
- Partially answering important questions:
  1. **Which value of** K should be chosen, given the sample size n.
  2. Whether it is better to use a **few complex experts** or **combine many simple experts**, given the total number of parameters.
- Minimax lower bounds: only known for mixture models[8].

[8] Maugis-Rabusseau, C. and Michel, B. (2013). Adaptive density estimation for clustering with Gaussian mixtures. ESAIM: Probability and Statistics.

My Coauthors $\in$ Mixture of French and Australian Experts

UNIVERSITÉ CAEN NORMANDIE

THE UNIVERSITY OF QUEENSLAND AUSTRALIA

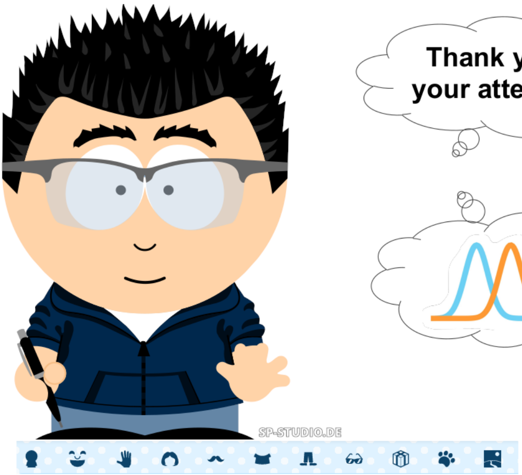
Inria

Faicel Chamroukhi

Hien Duy Nguyen

Florence Forbes

“Essentially, all models are wrong, but some are useful”. George E.P. Box (1987).



➚ This is my best data-driven model to approximate myself.