A non-asymptotic approach for model selection via penalization in high-dimensional mixture of experts models

TrungTin Nguyen

Seminar at Department of Statistical Sciences, University of Padova

Outline

1. Collection of GLoME and BLoME models
   - Context and motivating example
   - Conditional density estimation
   - Graphical model representation of MoE models
   - Gaussian gating networks
2. Model selection in GLoME amd BLoME models
   - Model selection in standard MoE regression models
   - Penalized maximum likelihood estimator
   - Asymptotic approach
   - Non-asymptotic approach with oracle inequalities
3. Numerical experiments
4. Main positive messages and perspectives

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**[1] **(MoE)** models due to their flexibility and effectiveness (several universal approximation theorems [2] [3] [4] [5] with good convergence rates).

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

[2] Ho, N., Yang, C.-Y., and Jordan, M. I. (2022). Convergence rates for Gaussian mixtures of experts. Journal of Machine Learning Research.

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

[4] 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*.

[5] **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**[1] **(MoE)** models due to their flexibility and effectiveness (several universal approximation theorems [2] [3] [4] [5] with good convergence rates).

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

[2] Ho, N., Yang, C.-Y., and Jordan, M. I. (2022). Convergence rates for Gaussian mixtures of experts. Journal of Machine Learning Research.

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

[4] 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.*

[5] **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**[1] **(MoE)** models due to their flexibility and effectiveness (several universal approximation theorems [2] [3] [4] [5] with good convergence rates).

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

[2] Ho, N., Yang, C.-Y., and Jordan, M. I. (2022). Convergence rates for Gaussian mixtures of experts. Journal of Machine Learning Research.

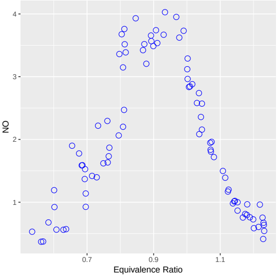
[3] Nguyen, H. D., Chamroukhi, F., and Forbes, F. (2019). Approximation results regarding the multiple-output Gaussian gated mixture of linear experts model. Neurocomputing.

[4] 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.*

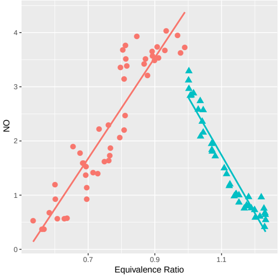
[5] **Nguyen, T.**, Chamroukhi, F., Nguyen, H. D., and McLachlan, G. J. (2022). Approximation of probability density functions via location-scale finite mixtures in Lebesgue spaces. *Communications in Statistics - Theory and Method.*

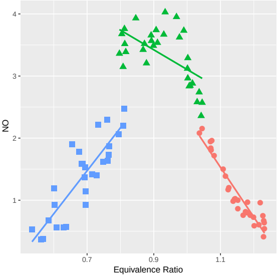
Motivating example: Ethanol data set 88 observations

(a) Raw Ethanol data set

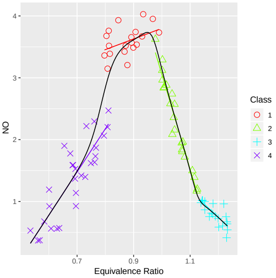


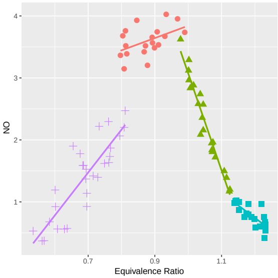
Collection of MoE models with linear mean functions characterized by 2-5 clusters

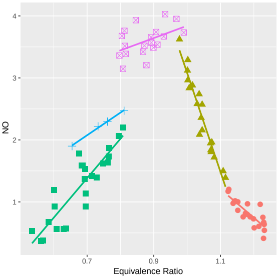




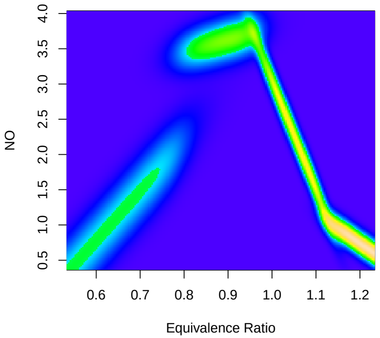
(b) Our best data-driven MoE model



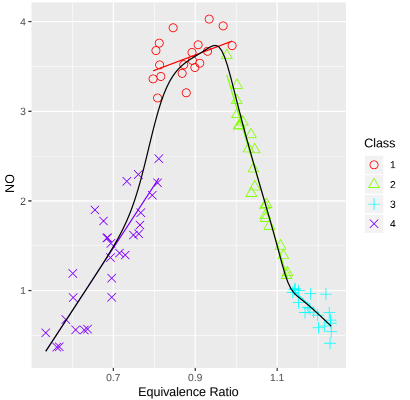




Outline



(a) 2D view of our estimate conditional density with 4 clusters.



(b) Our nonlinear regression and clustering using MoE models.

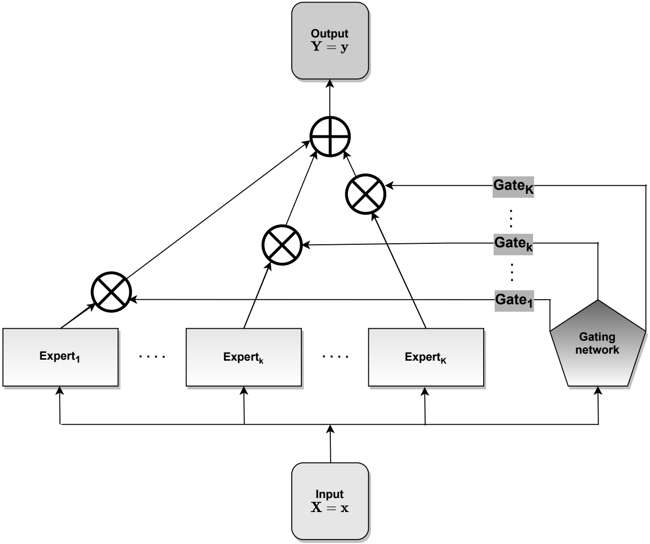
Regression and clustering have been recasted as a task of estimating the true but unknown conditional density estimation s_0 using MoE distribution

$\Rightarrow$ It makes sense to select models from the conditional density point of view.

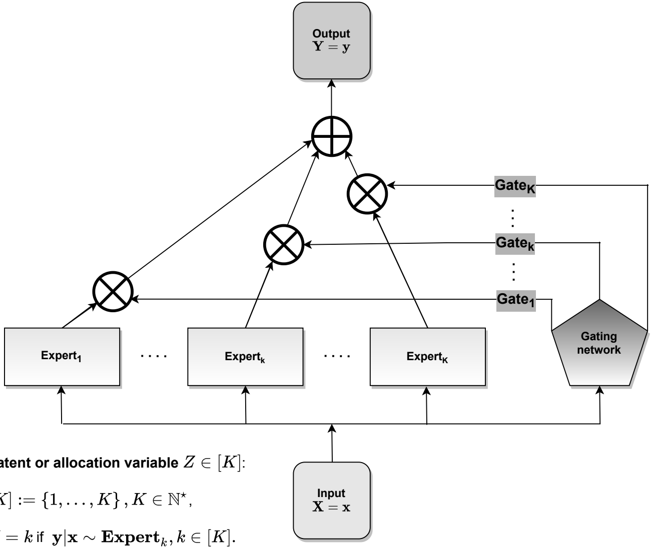
Flexibility and effectiveness of MoE models

Originally introduced as neural network architectures in [Jacobs et al., 1991, Jordan and Jacobs, 1994]:

- Modeling more complex data generating processes than classical finite mixtures or finite mixtures of regression models [McLachlan and Peel, 2000].
- Universal approximation properties with good convergence rates [Mendes and Jiang, 2012, Norets, 2010, Ho et al., 2022, Nguyen et al., 2019, **Nguyen et al., 2020b**, **Nguyen et al., 2021a**, **Nguyen et al., 2020a**].
- Applied to numerous areas of business, science, and technology for the tasks: clustering, regression analysis, conditional density estimation and classification [Yuksel et al., 2012, Masoudnia and Ebrahimpour, 2014, Nguyen and Chamroukhi, 2018, Chamroukhi & Huynh, 2019].



Schematic diagram of the neural network architecture of a K-component MoE model.

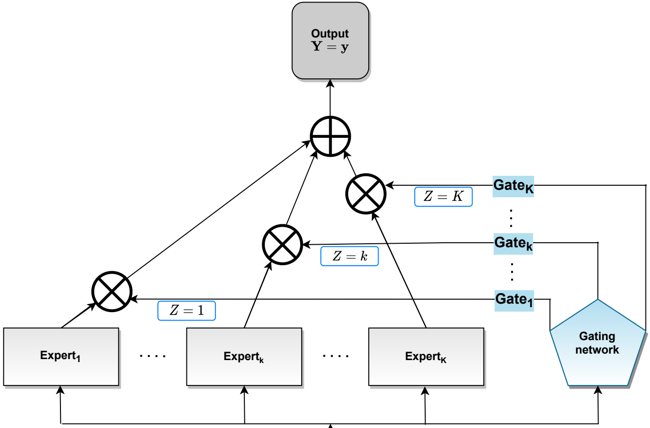


Latent or allocation variable $Z \in [K]$:

$[K] := \{1, \dots, K\}, K \in \mathbb{N}^\star$,

$Z = k$ if $\mathbf{y}|\mathbf{x} \sim \mathbf{Expert}_k, k \in [K]$.

Schematic diagram of the neural network architecture of a K-component MoE model.



Latent or allocation variable $Z \in [K]$:

$[K] := \{1, \dots, K\}, K \in \mathbb{N}^\star$,

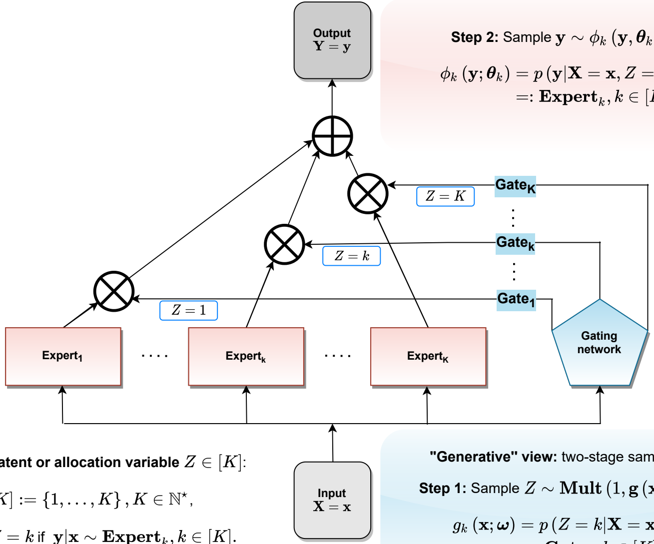
$Z = k$ if $\mathbf{y}|\mathbf{x} \sim \mathbf{Expert}_k, k \in [K]$.

"Generative" view: two-stage sampling

Step 1: Sample $Z \sim \mathbf{Mult}\left(1, \mathbf{g}\left(\mathbf{x}, \boldsymbol{\omega}\right)\right)$,

$$g_k\left(\mathbf{x}; \boldsymbol{\omega}\right) = p\left(Z = k | \mathbf{X} = \mathbf{x}\right) =: \mathbf{Gate}_k, k \in [K].$$

Schematic diagram of the neural network architecture of a K-component MoE model.



Step 2: Sample $\mathbf{y} \sim \phi_k\left(\mathbf{y}, \boldsymbol{\theta}_k\right)$,

$$\begin{aligned}\phi_k\left(\mathbf{y}; \boldsymbol{\theta}_k\right) &= p\left(\mathbf{y} | \mathbf{X} = \mathbf{x}, Z = k\right)\\ &=: \mathbf{Expert}_k, k \in [K].\end{aligned}$$

Latent or allocation variable $Z \in [K]$:

$[K] := \{1, \ldots, K\}, K \in \mathbb{N}^\star$,

$Z = k$ if $\mathbf{y} | \mathbf{x} \sim \mathbf{Expert}_k, k \in [K]$.

"Generative" view: two-stage sampling

Step 1: Sample $Z \sim \mathbf{Mult}\left(1, \mathbf{g}\left(\mathbf{x}, \boldsymbol{\omega}\right)\right)$,

$$\begin{aligned}g_k\left(\mathbf{x}; \boldsymbol{\omega}\right) &= p\left(Z = k | \mathbf{X} = \mathbf{x}\right)\\ &=: \mathbf{Gate}_k, k \in [K].\end{aligned}$$

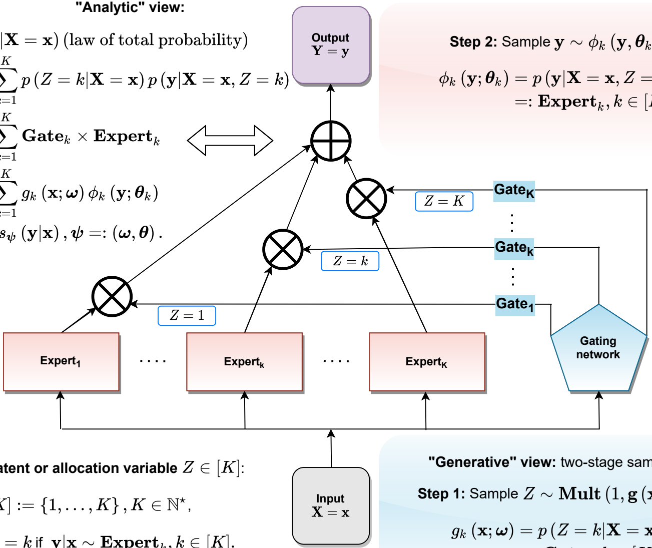
Schematic diagram of the neural network architecture of a K-component MoE model.

"Analytic" view:

$$
\begin{aligned}
&p(\mathbf{y}|\mathbf{X}=\mathbf{x}) \text{ (law of total probability)}\\
&= \sum_{k=1}^{K} p(Z=k|\mathbf{X}=\mathbf{x})\, p(\mathbf{y}|\mathbf{X}=\mathbf{x}, Z=k)\\
&= \sum_{k=1}^{K} \mathbf{Gate}_k \times \mathbf{Expert}_k\\
&= \sum_{k=1}^{K} g_k(\mathbf{x};\boldsymbol{\omega})\, \phi_k(\mathbf{y};\boldsymbol{\theta}_k)\\
&=: s_{\boldsymbol{\psi}}(\mathbf{y}|\mathbf{x}), \boldsymbol{\psi} =: (\boldsymbol{\omega}, \boldsymbol{\theta}).
\end{aligned}
$$

Step 2: Sample $\mathbf{y} \sim \phi_k(\mathbf{y}, \boldsymbol{\theta}_k)$,

$$
\begin{aligned}
\phi_k(\mathbf{y};\boldsymbol{\theta}_k) &= p(\mathbf{y}|\mathbf{X}=\mathbf{x}, Z=k)\\
&=: \mathbf{Expert}_k, k \in [K].
\end{aligned}
$$



Latent or allocation variable $Z \in [K]$:

$[K] := \{1, \dots, K\}, K \in \mathbb{N}^\star$,

$Z = k$ if $\mathbf{y}|\mathbf{x} \sim \mathbf{Expert}_k, k \in [K]$.

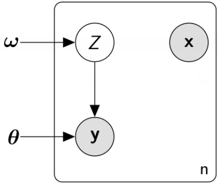
"Generative" view: two-stage sampling

Step 1: Sample $Z \sim \mathbf{Mult}(1, \mathbf{g}(\mathbf{x}, \boldsymbol{\omega}))$,

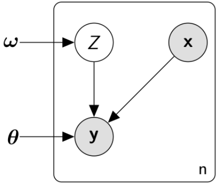
$$
\begin{aligned}
g_k(\mathbf{x};\boldsymbol{\omega}) &= p(Z=k|\mathbf{X}=\mathbf{x})\\
&=: \mathbf{Gate}_k, k \in [K].
\end{aligned}
$$

Schematic diagram of the neural network architecture of a K-component MoE model.

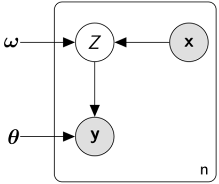
Outline



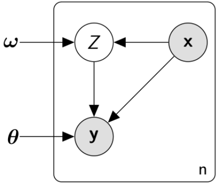
(a) Mixture model



(b) MoE regression model

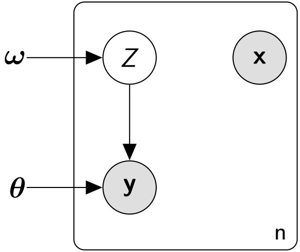


(c) Simple MoE regression model

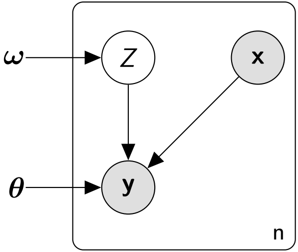


(d) Standard MoE regression model

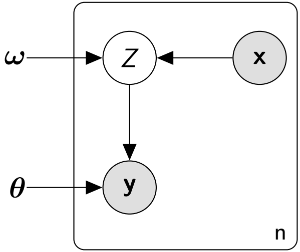
Presence or absence of edges between the inputs $\mathbf{x}$, the latent variable Z and the output $\mathbf{y}$ $\implies$ four special cases of MoE regression models.



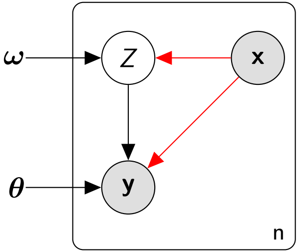
(a) Mixture model



(b) MoE regression model



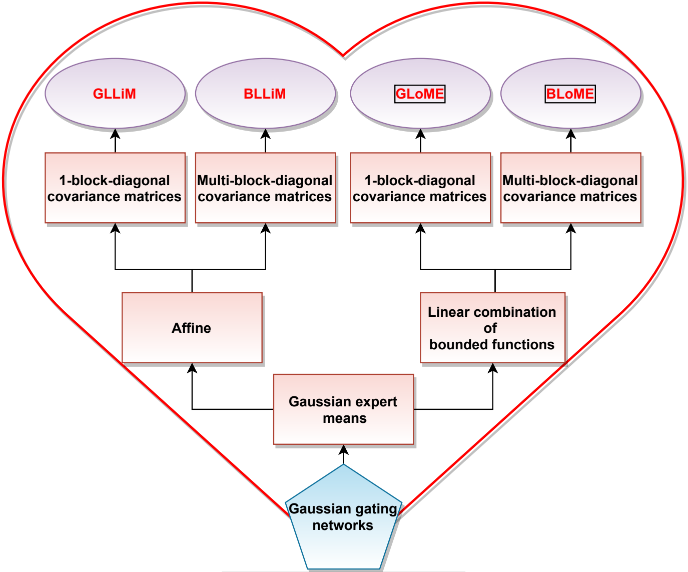
(c) Simple MoE regression model



(d) **Standard MoE regression model**

Model selection and approximation for **standard MoE regression models**.

Outline



GLoME: Gaussian-gated Localized MoE

GLLiM: Gaussian Locally-Linear Mapping

BLoME: Block-diagonal covariance Gaussian-gated Localized MoE

BLLiM: Block-diagonal covariance Gaussian Locally-Linear Mapping

Definition: GLLiM, BLLiM, GLoME and BLoME models

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

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

High-dimensional data using inverse regression frameworks (GLLiM models [Deleforge et al., 2015]): $\mathbf{Y} \equiv$ input, $\mathbf{X} \equiv$ output, $\mathcal{X} \subset \mathbb{R}^D$, $\mathcal{Y} \subset \mathbb{R}^L$, with $D \gg L$.

Establishing non-asymptotic oracle inequalities $\leftarrow$ Boundedness conditions on model parameters $\psi_{K,d,\mathbf{B}}$.

Definition: GLLiM, BLLiM, GLoME and BLoME models

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

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

High-dimensional data using inverse regression frameworks (GLLiM models [Deleforge et al., 2015]): $\mathbf{Y} \equiv$ input, $\mathbf{X} \equiv$ output, $\mathcal{X} \subset \mathbb{R}^D$, $\mathcal{Y} \subset \mathbb{R}^L$, with $D \gg L$.

Establishing non-asymptotic oracle inequalities $\leftarrow$ Boundedness conditions on model parameters $\psi_{K,d,\mathbf{B}}$.

Definition: GLLiM, BLLiM, GLoME and BLoME models

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

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

High-dimensional data using inverse regression frameworks (GLLiM models [Deleforge et al., 2015]): $\mathbf{Y} \equiv$ input, $\mathbf{X} \equiv$ output, $\mathcal{X} \subset \mathbb{R}^D$, $\mathcal{Y} \subset \mathbb{R}^L$, with $D \gg L$.

Establishing non-asymptotic oracle inequalities $\leftarrow$ Boundedness conditions on model parameters $\psi_{K,d,\mathbf{B}}$.

Definition: Collection of models (GLLiM, BLLiM, GLoME and BLoME)

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

- $\mathbf{m} \in \mathcal{M} = [K_{\max}] \times [d_{\max}] \times (\mathcal{B}_k)_{k\in[K]}, K_{\max}, d_{\max} \in \mathbb{N}^\star$.
- $\mathcal{B}_k$ = all possible partitions of the covariables indexed by $[D]$.
- $\widetilde{\mathcal{M}} = [K_{\max}] \times [d_{\max}] \times (\mathcal{B}_{k,\Lambda})_{k\in[K]} \subset \mathcal{M}$: high-dimensional data.

↑

[Devijver et al., 2017, Devijver et al., 2018] BLLiM procedure: trade-off complexity and sparsity ← Prediction on gene expression data with heterogeneous observations and hidden graph-structured interactions between small modules of correlated genes.

Definition: Collection of models (GLLiM, BLLiM, GLoME and BLoME)

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

- $\mathbf{m} \in \mathcal{M} = [K_{\max}] \times [d_{\max}] \times (\mathcal{B}_k)_{k\in[K]}, K_{\max}, d_{\max} \in \mathbb{N}^\star$.
- $\mathcal{B}_k$ = all possible partitions of the covariables indexed by $[D]$.
- $\widetilde{\mathcal{M}} = [K_{\max}] \times [d_{\max}] \times (\mathcal{B}_{k,\boldsymbol{\Lambda}})_{k\in[K]} \subset \mathcal{M}$: high-dimensional data.

$\uparrow$

[Devijver et al., 2017, Devijver et al., 2018] BLLiM procedure: trade-off complexity and sparsity $\leftarrow$ Prediction on gene expression data with heterogeneous observations and hidden graph-structured interactions between small modules of correlated genes.

Definition: Collection of models (GLLiM, BLLiM, GLoME and BLoME)

$$S_{\mathbf{m}} = \Big\{ (\mathbf{x}, \mathbf{y}) \mapsto s_{\psi_{K,d,\mathbf{B}}}(\mathbf{x}|\mathbf{y}) = s_{\mathbf{m}}(\mathbf{x}|\mathbf{y}) : \mathbf{m} = (K, d, \mathbf{B}),$$

$$\psi_{K,d,\mathbf{B}} = (\boldsymbol{\omega}, \boldsymbol{\upsilon}, \boldsymbol{\Sigma}(\mathbf{B})) \in \widetilde{\boldsymbol{\Omega}}_K \times \boldsymbol{\Upsilon}_{K,d} \times \mathbf{V}_K(\mathbf{B}) = \widetilde{\boldsymbol{\Psi}}_{K,d,\mathbf{B}} \Big\}.$$

- $\mathbf{m} \in \mathcal{M} = [K_{\max}] \times [d_{\max}] \times (\mathcal{B}_k)_{k\in[K]}, K_{\max}, d_{\max} \in \mathbb{N}^\star$.
- $\mathcal{B}_k$ = all possible partitions of the covariables indexed by $[D]$.
- $\widetilde{\mathcal{M}} = [K_{\max}] \times [d_{\max}] \times (\mathcal{B}_{k,\boldsymbol{\Lambda}})_{k\in[K]} \subset \mathcal{M}$: high-dimensional data.

$\Uparrow$

[Devijver et al., 2017, Devijver et al., 2018] BLLiM procedure: trade-off complexity and sparsity $\leftarrow$ Prediction on gene expression data with heterogeneous observations and hidden graph-structured interactions between small modules of correlated genes.

Outline

Outline

Model selection in standard MoE regression models

- **Best data-driven model**: selecting from a collection of MoE models characterized by hyperparameters,
  - GLoME models: $\mathbf{m} = (K, d)$,
  - BLoME models: $\mathbf{m} = (K, d, \mathbf{B})$,

→ **Penalized maximum likelihood estimator (PMLE)**:
  - **MLE is not sufficient**: underestimation of the risk of the estimate $\Rightarrow$ choosing models too complex.
  - **PMLE via adding pen(m)**: compensate bias (too simple model) and variance (too complex model).

- **Our contributions**: establishing **non-asymptotic risk bounds** that take the form of **weak oracle inequalities**, provided that **lower bounds on the penalties** hold true.
  1. Deterministic collection of MoE models characterized by $\mathcal{M}$: GLLiM, GLoME.
  2. **Random** collection of MoE models characterized by $\widetilde{\mathcal{M}}$: **BLLiM, BLoME**.

Model selection in standard MoE regression models

- **Best data-driven model**: selecting from a collection of MoE models characterized by hyperparameters,
  - GLoME models: $\mathbf{m} = (K, d)$,
  - BLoME models: $\mathbf{m} = (K, d, \mathbf{B})$,

→ **Penalized maximum likelihood estimator (PMLE)**:
  - **MLE is not sufficient**: underestimation of the risk of the estimate $\Rightarrow$ choosing models too complex.
  - **PMLE via adding $\text{pen}(\mathbf{m})$**: compensate bias (too simple model) and variance (too complex model).

- **Our contributions**: establishing **non-asymptotic risk bounds** that take the form of **weak oracle inequalities**, provided that **lower bounds on the penalties** hold true.
  1. Deterministic collection of MoE models characterized by $\mathcal{M}$: GLLiM, GLoME.
  2. **Random** collection of MoE models characterized by $\widetilde{\mathcal{M}}$: **BLLiM, BLoME**.

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 $\mathrm{KL}^{\otimes n}$ **(conditional densities and random covariate variables)**:

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

  if $sdy \ll tdy$, $+\infty$ otherwise.

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

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

Fixed predictors $\Rightarrow$ no $\mathbb{E}_{\mathbf{Y}_{[n]}}[\cdot]$.

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.

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

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

Fixed predictors $\Rightarrow$ no $\mathbb{E}_{\mathbf{Y}_{[n]}}[\cdot]$.

Outline

Some asymptotic approaches for model selection in MoE models

- Akaike information criterion (AIC), Bayesian information criterion[6] (BIC) 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].

[6] Forbes, F., Nguyen, H. D., **Nguyen, T.**, & Arbel, J. (2022). Summary statistics and discrepancy measures for approximate Bayesian computation via surrogate posteriors. *Statistics and Computing*.

- Akaike information criterion (AIC), Bayesian information criterion[6] (BIC) 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].

6 Forbes, F., Nguyen, H. D., **Nguyen, T.**, & Arbel, J. (2022). Summary statistics and discrepancy measures for approximate Bayesian computation via surrogate posteriors. *Statistics and Computing*.

Some asymptotic approaches for model selection in MoE models

- Akaike information criterion (AIC), Bayesian information criterion[6] (BIC) 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 $\mathbb{E}\left[\text{KL}^{\otimes n}\left(s_0, \widehat{s}_{\mathbf{m}}\right)\right]$:
  - ✔ Finite sample guarantee.
  - ✖ Strong regularity assumptions of [White, 1982].

[6] Forbes, F., Nguyen, H. D., **Nguyen, T.**, & Arbel, J. (2022). Summary statistics and discrepancy measures for approximate Bayesian computation via surrogate posteriors. *Statistics and Computing*.

Outline

Theorem (Non-asymptotic oracle inequality for deterministic collection of MoE models[a])

[a] **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 models. Electronic Journal of Statistics.

- **Assumptions**: *We are given:* $(S_{\mathbf{m}})_{\mathbf{m}\in\mathcal{M}}$, $\rho \in (0,1)$, $C_1 > 1$, $\Xi = \sum_{\mathbf{m}\in\mathcal{M}} e^{-z_m} < \infty, z_m \in \mathbb{R}^+, \forall m \in \mathcal{M}$.

- **Non-asymptotic upper bound:** *There exist constants* C *and* $\kappa(\rho, C_1) > 0$ *such that whenever for all* $m \in \mathcal{M}$,

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

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

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

Theorem (Non-asymptotic oracle inequality for deterministic collection of MoE models[a])

[a] **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 models. Electronic Journal of Statistics.

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

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

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

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

Theorem: Oracle inequality for random collection [Nguyen et al., 2022]

- Given random (defined on datasets) subcollection $(S_{\mathbf{m}})_{\mathbf{m}\in\widetilde{\mathcal{M}}}$, $\widetilde{\mathcal{M}} \subset \mathcal{M}$, $\rho \in (0,1)$, $C_1 > 1$, we define $\Xi = \sum_{\mathbf{m}\in\mathcal{M}} e^{-z_m} < \infty, z_m \in \mathbb{R}^+, \forall m \in \mathcal{M}$. Assume that there exists $\tau > 0$ and $\epsilon_{KL} > 0$ such that, for all $\mathbf{m} \in \mathcal{M}$, one can find $\bar{s}_{\mathbf{m}} \in S_{\mathbf{m}}$, such that $\bar{s}_{\mathbf{m}} \geq e^{-\tau} s_0$, and

$$\mathrm{KL}^{\otimes n}(s_0, \bar{s}_{\mathbf{m}}) \leq \inf_{s_{\mathbf{m}}\in S_{\mathbf{m}}} \mathrm{KL}^{\otimes n}(s_0, s_{\mathbf{m}}) + \frac{\epsilon_{KL}}{n}.$$

- There exist constant C, $\kappa(\rho, C_1) > 0$, $C_2(\rho, C_1) > 0$ such that if

$$\mathrm{pen}(\mathbf{m}) \geq \kappa(\rho, C_1)\left[(C + \ln n)\dim(S_{\mathbf{m}}) + (1 \vee \tau) z_m\right], \forall m \in \mathcal{M},$$

then η'-PMLE $\widehat{s}_{\widehat{\mathbf{m}}}$ on $\widetilde{\mathcal{M}}$ satisfies

$$\mathbb{E}\left[\mathrm{JKL}_{\rho}^{\otimes n}(s_0, \widehat{s}_{\widehat{\mathbf{m}}})\right] \leq C_1 \mathbb{E}\left[\inf_{\mathbf{m}\in\widetilde{\mathcal{M}}}\left(\inf_{s_{\mathbf{m}}\in S_{\mathbf{m}}} \mathrm{KL}^{\otimes n}(s_0, s_{\mathbf{m}}) + 2\frac{\mathrm{pen}(\mathbf{m})}{n}\right)\right]$$

$$+ C_2(\rho, C_1)(1 \vee \tau)\frac{\Xi^2}{n} + \frac{\eta + \eta'}{n}.$$

Theorem: Oracle inequality for random collection [Nguyen et al., 2022]

Given random (defined on datasets) subcollection $(S_\mathbf{m})_{\mathbf{m}\in\widetilde{\mathcal{M}}}$, $\widetilde{\mathcal{M}} \subset \mathcal{M}$, $\rho \in (0,1)$, $C_1 > 1$, we define $\Xi = \sum_{\mathbf{m}\in\mathcal{M}} e^{-z_m} < \infty, z_m \in \mathbb{R}^+, \forall m \in \mathcal{M}$. Assume that there exists $\tau > 0$ and $\epsilon_{KL} > 0$ such that, for all $\mathbf{m} \in \mathcal{M}$, one can find $\bar{s}_\mathbf{m} \in S_\mathbf{m}$, such that $\bar{s}_\mathbf{m} \geq e^{-\tau} s_0$, and

$$\mathrm{KL}^{\otimes n}(s_0, \bar{s}_\mathbf{m}) \leq \inf_{s_\mathbf{m}\in S_\mathbf{m}} \mathrm{KL}^{\otimes n}(s_0, s_\mathbf{m}) + \frac{\epsilon_{KL}}{n}.$$

There exist constant C, $\kappa(\rho, C_1) > 0$, $C_2(\rho, C_1) > 0$ such that if

$$\mathrm{pen}(\mathbf{m}) \geq \kappa(\rho, C_1)\left[(C + \ln n)\dim(S_\mathbf{m}) + (1 \vee \tau) z_m\right], \forall m \in \mathcal{M},$$

then η'-PMLE $\widehat{s}_{\widehat{\mathbf{m}}}$ on $\widetilde{\mathcal{M}}$ satisfies

$$\mathbb{E}\left[\mathrm{JKL}_\rho^{\otimes n}(s_0, \widehat{s}_{\widehat{\mathbf{m}}})\right] \leq C_1 \mathbb{E}\left[\inf_{\mathbf{m}\in\widetilde{\mathcal{M}}}\left(\inf_{s_\mathbf{m}\in S_\mathbf{m}} \mathrm{KL}^{\otimes n}(s_0, s_\mathbf{m}) + 2\frac{\mathrm{pen}(\mathbf{m})}{n}\right)\right] + C_2(\rho, C_1)(1 \vee \tau)\frac{\Xi^2}{n} + \frac{\eta + \eta'}{n}.$$

Outline

Procedure for deterministic collection of GLLiM models

Goal: seek for the best data-driven model among $(S_m^*)_{\mathbf{m}\in\mathcal{M}}$, $\mathcal{M} = [K_{\max}]$ based on $(\mathbf{x}_i, \mathbf{y}_i)_{i\in[n]}$ arising from an forward conditional density s_0^*:

1. Each $m \in \mathcal{M}$: estimate the forward MLE $\widehat{s}_{\mathbf{m}}^*(\mathbf{y}_i|\mathbf{x}_i)$ by inverse MLE $\widehat{s}_{\mathbf{m}}$ via an inverse regression trick by GLLiM-EM algorithm (xLLiM package).
2. Calculate η'-PMLE $\widehat{s}_{\widehat{\mathbf{m}}}$ with a "simplified" $\mathrm{pen}(\mathbf{m}) = \kappa \dim(S_m^*)$.

➞ Data-driven non-asymptotic approach for choosing κ.

- ➞ Our oracle inequality partially suggests **shape of penalty function** in a finite sample setting.
- ➞ Slope heuristic approach (capushe package) works well with our simplified $\mathrm{pen}(\mathbf{m}) = \kappa \dim(S_m^*)$ [Birgé and Massart, 2007, Baudry et al., 2012, Arlot et al., 2016, Arlot, 2019].

Numerical protocols

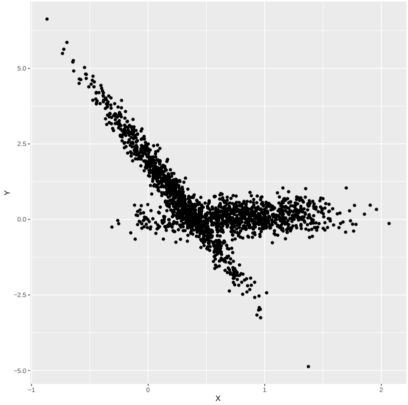
- $L = D = 1$: behavior of $\mathrm{JKL}_{\rho}^{\otimes n}\left(s_0^*, \widehat{s}_{\widehat{\mathbf{m}}}^*\right)$ and convergence rates of error terms $\frac{1}{n}$.
- $D \gg L$: dimensionality reduction capability of GLLiM in high-dimensional regression data [Deleforge et al., 2015].

➠**Well-Specified (WS)**: $s_0^* \in S_m^*$,

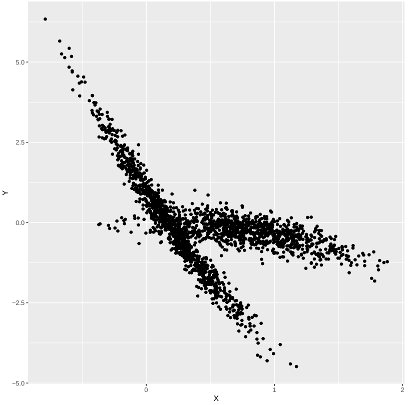
$$s_0^*(y|x) = \frac{\Phi(x; 0.2, 0.1)\Phi(y; -5x + 2, 0.09) + \Phi(x; 0.8, 0.15)\Phi(y; 0.1x, 0.09)}{\Phi(x; 0.2, 0.1) + \Phi(x; 0.8, 0.15)}.$$

➠**Misspecified (MS)**: $s_0^* \notin S_m^*$,

$$s_0^*(y|x) = \frac{\Phi(x; 0.2, 0.1)\Phi(y; x^2 - 6x + 1, 0.09) + \Phi(x; 0.8, 0.15)\Phi(y; -0.4x^2, 0.09)}{\Phi(x; 0.2, 0.1) + \Phi(x; 0.8, 0.15)}.$$



(a) WS case

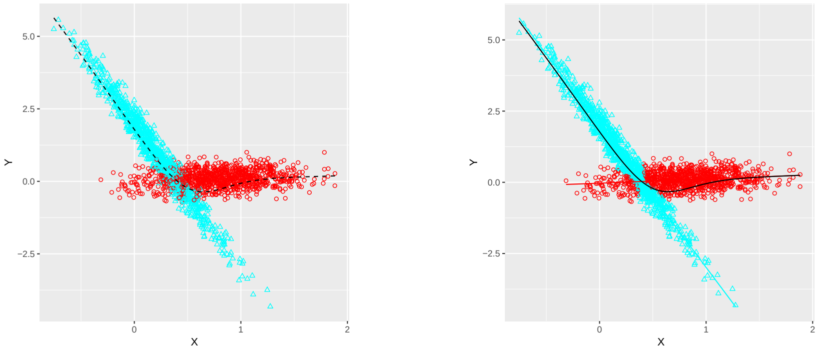


(b) MS case

Typical realizations of heterogeneous data from nonlinear regression models.

Perform multiple tasks simultaneously: regression analysis, clustering, conditional density estimation and model selection (*e.g.*, number of clusters, degree of polynomials).

Typical realization and regression clustering results

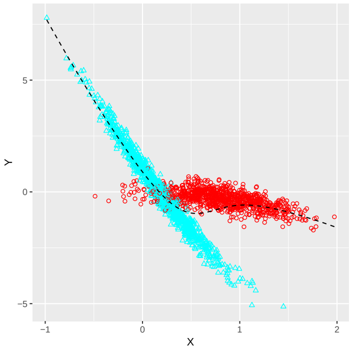


(a) Typical realization: WS case

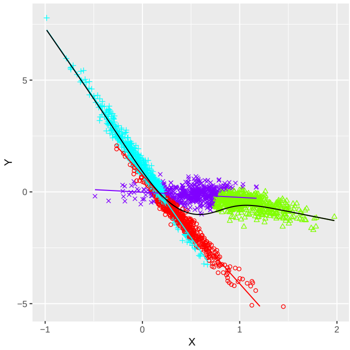
(b) Regression clustering by GLLiM

Regression and clustering deduced from the estimated conditional density of GLLiM with $n = 2000$ in example WS. The dash and solid black curves present the true and estimated mean functions.

Typical realization and regression clustering results

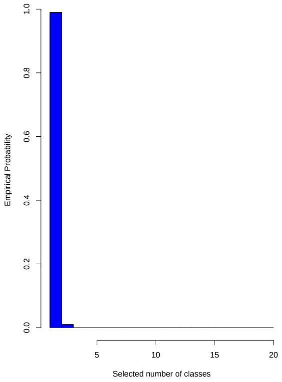


(a) Typical realization: MS case

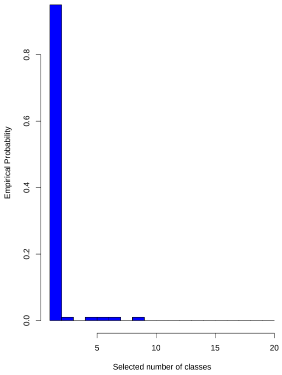


(b) Regression clustering by GLLiM

Regression clustering deduced from the estimated conditional density of GLLiM with $n = 2000$ in example MS. The dash and solid black curves present the true and estimated mean functions.

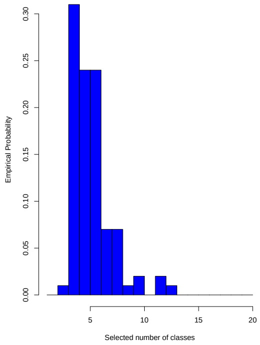


(a) $n = 2000$

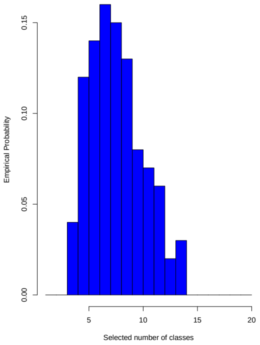


(b) $n = 10000$

Comparison histograms of selected K in WS case using jump criterion over 100 trials between $n = 2000$ and $n = 10000$.



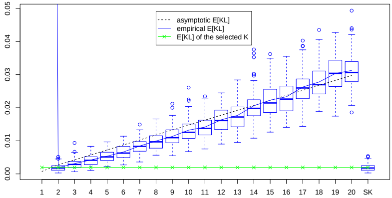
(a) 2000 data points



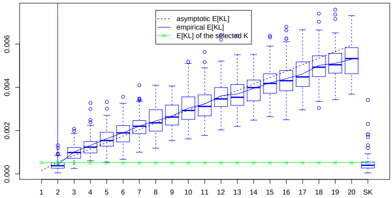
(b) 10000 data points

Comparison histograms of selected K in MS case using jump criterion over 100 trials between $n = 2000$ and $n = 10000$.

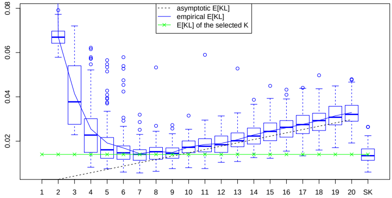
Box-plot of Kullback-Leibler divergence over 100 trials



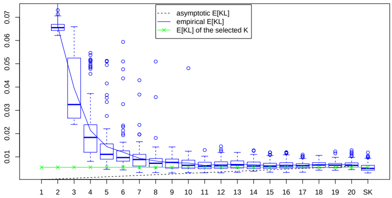
(a) WS with $n = 2000$



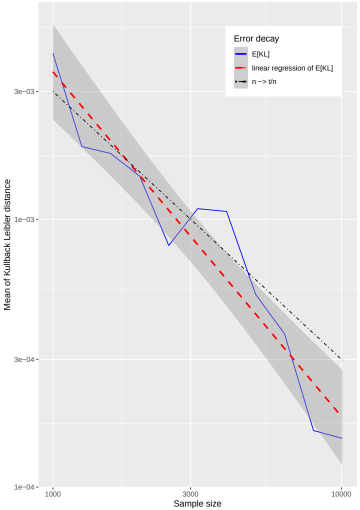
(b) WS with $n = 10000$



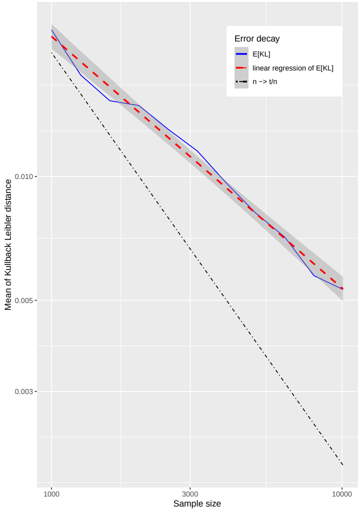
(c) MS with $n = 2000$



(d) MS with $n = 10000$



(a) WS: free regression's slope ≈ -1.287 and $t = 3$.



(b) MS: free regression's slope ≈ -0.6120, $t = 20$.

Rate of error decay, $\mathbb{E}\left[\mathrm{KL}^{\otimes n}\left(s_0, \widehat{s_{\mathbf{m}}}\right)\right]$, is represented in a log-log scale, using 30 trials. A free least-square regression (black dashed line) with standard error and a regression with slope -1 were added for presenting rate of convergence $1/n$.

Approximation error and variance term

The bias-variance trade-off differs between the two examples:

- **WS case:** since the true density belongs to the model, the best data-driven choice is $K = 2$ even for large n.
- **MS case:** best data-driven choice K should balance a model **approximation error term** and a **variance** one, *i.e.*, the larger n the more complex the model and thus K.

Empirical behavior of weak oracle inequality

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

- No closed form formula for $\mathrm{JKL}_{\rho}^{\otimes n}\left(s_0, \widehat{s}_{\widehat{\mathbf{m}}}\right)$ ➠ Monte Carlo method.
- Empirical mean $\mathrm{JKL}_{\rho}^{\otimes n}\left(s_0, \widehat{s}_{\widehat{\mathbf{m}}}\right) \leq$ Empirical mean $\mathrm{KL}^{\otimes n}\left(s_0, \widehat{s}_{\mathbf{m}}\right)$, $m \in \mathcal{M} = [20]$ over 55 trials.
- Empirical mean $\mathrm{KL}^{\otimes n}\left(s_0, \widehat{s}_{\mathbf{m}}\right) \sim \frac{\dim(S_{\mathbf{m}})}{2n}$ (shown by a dotted line): **expected behavior in asymptotic theory in WS case!**

Empirical behavior of weak oracle inequality

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

- No closed form formula for $\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_{\widehat{\mathbf{m}}}\right)$ ➠ Monte Carlo method.
- Empirical mean $\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_{\widehat{\mathbf{m}}}\right) \leq$ Empirical mean $\mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_{\mathbf{m}}\right)$, $m \in \mathcal{M} = [20]$ over 55 trials.
- Empirical mean $\mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_{\mathbf{m}}\right) \sim \frac{\dim(S_{\mathrm{m}})}{2n}$ (shown by a dotted line): **expected behavior in asymptotic theory in WS case!**

Empirical behavior of weak oracle inequality

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

- No closed form formula for $\mathrm{JKL}_{\rho}^{\otimes n}\left(s_0, \widehat{s}_{\widehat{\mathbf{m}}}\right)$ ➠ Monte Carlo method.
- Empirical mean $\mathrm{JKL}_{\rho}^{\otimes n}\left(s_0, \widehat{s}_{\widehat{\mathbf{m}}}\right) \leq$ Empirical mean $\mathrm{KL}^{\otimes n}\left(s_0, \widehat{s}_{\mathbf{m}}\right)$, $m \in \mathcal{M} = [20]$ over 55 trials.
- Empirical mean $\mathrm{KL}^{\otimes n}\left(s_0, \widehat{s}_{\mathbf{m}}\right) \sim \frac{\dim(S_{\mathbf{m}})}{2n}$ (shown by a dotted line): **expected behavior in asymptotic theory in WS case!**

Outline

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 on model selection:
  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.
- **Numerical experiment** for **high-dimensional data**.
- **Minimax lower bounds**: only known for mixture models[7].
- **Mathematically justifying the slope heuristic** in MoE models as in least-squares regression on a random (or fixed) design with regressogram (projection) estimators, respectively, [Birgé and Massart, 2007, Arlot and Massart, 2009, Arlot and Bach, 2009, Arlot, 2019].

[7] 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 on model selection:
  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.
- **Numerical experiment** for **high-dimensional data**.
- **Minimax lower bounds**: only known for mixture models[7].
- **Mathematically justifying the slope heuristic** in MoE models as in least-squares regression on a random (or fixed) design with regressogram (projection) estimators, respectively, [Birgé and Massart, 2007, Arlot and Massart, 2009, Arlot and Bach, 2009, Arlot, 2019].

[7] 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 on model selection:
  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.
- **Numerical experiment** for **high-dimensional data**.
- **Minimax lower bounds**: only known for mixture models[7].
- **Mathematically justifying the slope heuristic** in MoE models as in least-squares regression on a random (or fixed) design with regressogram (projection) estimators, respectively, [Birgé and Massart, 2007, Arlot and Massart, 2009, Arlot and Bach, 2009, Arlot, 2019].

[7] 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 on model selection:
  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.
- **Numerical experiment** for **high-dimensional data**.
- **Minimax lower bounds**: only known for mixture models[7].
- **Mathematically justifying the slope heuristic** in MoE models as in least-squares regression on a random (or fixed) design with regressogram (projection) estimators, respectively, [Birgé and Massart, 2007, Arlot and Massart, 2009, Arlot and Bach, 2009, Arlot, 2019].

[7] 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 on model selection:
  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.
- **Numerical experiment** for **high-dimensional data**.
- **Minimax lower bounds**: only known for mixture models[7].
- **Mathematically justifying the slope heuristic** in MoE models as in least-squares regression on a random (or fixed) design with regressogram (projection) estimators, respectively, [Birgé and Massart, 2007, Arlot and Massart, 2009, Arlot and Bach, 2009, Arlot, 2019].

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

References I

Arlot, S., & Bach, F. (2009). Data-driven calibration of linear estimators with minimal penalties. *Advances in Neural Information Processing Systems*, (Vol. 22). Curran Associates.
(Cited on pages 59, 60, 61, 62, and 63.)

Arlot, S., & Massart, P. (2009). Data-driven Calibration of Penalties for Least-Squares Regression. *Journal of Machine Learning Research*, 10(10), 245–279.
(Cited on pages 59, 60, 61, 62, and 63.)

Arlot, S. (2019). Minimal penalties and the slope heuristics: a survey. *Journal de la Société Française de Statistique*, 160(3), 1–106.
(Cited on pages 45, 59, 60, 61, 62, and 63.)

Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7), 719–725.
(Cited on pages 36, 37, and 38.)

Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, 138(1):33–73.
(Cited on pages 45, 59, 60, 61, 62, and 63.)

Deleforge, A., Forbes, F., and Horaud, R. (2015). High-dimensional regression with gaussian mixtures and partially-latent response variables. *Statistics and Computing*, 25(5):893–911.
(Cited on pages 21, 22, 23, and 46.)

References II

Devijver, E., Gallopin, M. (2018). Block-diagonal covariance selection for high-dimensional Gaussian graphical models. *Journal of the American Statistical Association*.

(Cited on pages 24, 25, and 26.)

Devijver, E., Gallopin, M., and Perthame, E. (2017). Nonlinear network-based quantitative trait prediction from transcriptomic data. *arXiv preprint arXiv:1701.07899*.

(Cited on pages 24, 25, and 26.)

White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, 50(1):1–25.

(Cited on pages 36, 37, 38, 81, 82, 83, 84, 85, 86, and 87.)

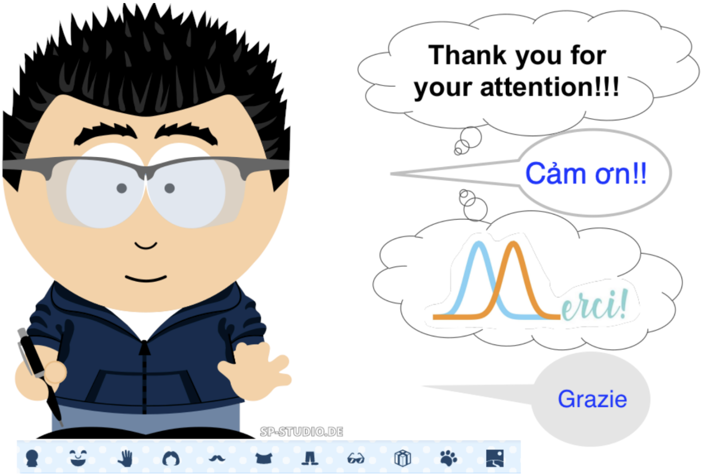
My Coauthors ∈ Mixture of French and Australian Experts

Hien Duy Nguyen

Faicel Chamroukhi

Florence Forbes

“Essentially, all models are wrong, but some are useful”.[8]



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

[8] Box, G. E.P. (1979). “Robustness in the strategy of scientific model building”. In Robustness in Statistics (pp. 201-236). Academic Press.

Supplementary material

Outline

Mild assumption: Boundedness conditions

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

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

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

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

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

Classical covariance matrix parameterization[9]

Boundedness conditions on Gaussian expert covariance matrices

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

- $B_k = |\boldsymbol{\Sigma}_k|^{1/D}$: 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 $\lambda_-, \lambda_+ \in \mathbb{R}$.

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

Mild assumption: Boundedness conditions on eigenvalues of Gaussian expert block-diagonal covariance matrices

$$0 < \lambda_m \leq m\left(\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)\right) \leq M\left(\boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right)\right) \leq \lambda_M, \text{ for every } k \in [K].$$

↑ constant ↑ smallest ↑ largest eigenvalue

$$\left\{ \boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right) \in \mathcal{S}_q^{++} \;\middle|\; \begin{array}{l} \boldsymbol{\Sigma}_k\left(\mathbf{B}_k\right) = \mathbf{P}_k \begin{pmatrix} \boldsymbol{\Sigma}_k^{[1]} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{\Sigma}_k^{[2]} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \boldsymbol{\Sigma}_k^{[G_k]} \end{pmatrix} \mathbf{P}_k^{-1}, \\ \boldsymbol{\Sigma}_k^{[g]} \in \mathcal{S}^{++}_{\text{card}\left(d_k^{[g]}\right)}, \forall g \in [G_k] \end{array} \right\}$$

$\mathbf{V}_K\left(\mathbf{B}\right) = \left(\mathbf{V}_k\left(\mathbf{B}_k\right)\right)_{k\in[K]}$, $\mathbf{P}_k$: permutation, $\mathbf{B}_k = \left(d_k^{[g]}\right)_{g\in[G_k]}$: block structure, $d_k^{[g]}$: set of variables in gth group, $\mathbf{B} = \left(\mathbf{B}_k\right)_{k\in[K]}$.

Outline

Solution for different divergences

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

- In general: $\mathrm{JKL}_{\rho}^{\otimes n}\left(s_0, \widehat{s}_{\widehat{\mathbf{m}}}\right) \leq \mathrm{KL}^{\otimes n}\left(s_0, \widehat{s}_{\widehat{\mathbf{m}}}\right)$.
- If $\sup_{\mathbf{m}\in\mathcal{M}} \sup_{s_{\mathbf{m}}\in S_{\mathbf{m}}} \left\|s_0/s_{\mathbf{m}}\right\|_{\infty} < \infty \Leftarrow \mathcal{Y}$ is compact, s_0 is compactly supported, the regression functions are uniformly bounded, and a uniform lower bound on the eigenvalues of the covariance matrices, Proposition 1 from [Cohen and Le Pennec, 2011] implies that

$$\frac{C_{\rho}}{2+\ln\left\|s_0/\widehat{s}_{\widehat{\mathbf{m}}}\right\|_{\infty}} \mathrm{KL}^{\otimes n}\left(s_0, \widehat{s}_{\widehat{\mathbf{m}}}\right) \leq \mathrm{JKL}_{\rho}^{\otimes n}\left(s_0, \widehat{s}_{\widehat{\mathbf{m}}}\right).$$

Solution for different divergences

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

- In general: $\mathrm{JKL}_{\rho}^{\otimes n}\left(s_0, \widehat{s}_{\widehat{\mathbf{m}}}\right) \leq \mathrm{KL}^{\otimes n}\left(s_0, \widehat{s}_{\widehat{\mathbf{m}}}\right)$.
- If $\sup_{\mathbf{m}\in\mathcal{M}} \sup_{s_{\mathbf{m}}\in S_{\mathbf{m}}} \left\| s_0 / s_{\mathbf{m}} \right\|_{\infty} < \infty \Leftarrow \mathcal{Y}$ is compact, s_0 is compactly supported, the regression functions are uniformly bounded, and a uniform lower bound on the eigenvalues of the covariance matrices, Proposition 1 from [Cohen and Le Pennec, 2011] implies that

$$\frac{C_{\rho}}{2 + \ln \left\| s_0 / \widehat{s}_{\widehat{\mathbf{m}}} \right\|_{\infty}} \mathrm{KL}^{\otimes n}\left(s_0, \widehat{s}_{\widehat{\mathbf{m}}}\right) \leq \mathrm{JKL}_{\rho}^{\otimes n}\left(s_0, \widehat{s}_{\widehat{\mathbf{m}}}\right).$$

Solution for misspecified models

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

- $C_1 = 1$ with $\mathrm{KL}^{\otimes n}$ loss: still an open question, only known for aggregation of a finite number of densities as [Dalalyan & Sebbar, 2018, Rigollet, 2012].
- Bias $\inf_{s_{\mathbf{m}}\in S_{\mathbf{m}}} \mathrm{KL}^{\otimes n}\left(s_0, s_{\mathbf{m}}\right)$: small for $\mathcal{M}$ well-chosen via approximation capabilities of MoE and GMM models [Nguyen et al., 2019, **Nguyen et al., 2020b**, **Nguyen et al., 2020a**, **Nguyen et al., 2021a**].

Solution for misspecified models

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

- $C_1 = 1$ with $\mathrm{KL}^{\otimes n}$ loss: still an open question, only known for aggregation of a finite number of densities as [Dalalyan & Sebbar, 2018, Rigollet, 2012].

- Bias $\inf_{s_{\mathbf{m}}\in S_{\mathbf{m}}} \mathrm{KL}^{\otimes n}\left(s_0, s_{\mathbf{m}}\right)$: small for $\mathcal{M}$ well-chosen via approximation capabilities of MoE and GMM models [Nguyen et al., 2019, **Nguyen et al., 2020b**, **Nguyen et al., 2020a**, **Nguyen et al., 2021a**].

Lemma: Relationship between linear-weight softmax and Gaussian gating networks [Nguyen et al., 2021a]

In general, $\mathcal{P}_G \supset \mathcal{P}_S$.
Furthermore, if all $\boldsymbol{\Gamma}_k$, $k \in [K]$, are identical, then $\mathcal{P}_G = \mathcal{P}_S$.

$$\mathcal{P}_S = \left\{ \mathbf{y} \mapsto \left(\mathbf{g}_k\left(\mathbf{y};\gamma\right)\right)_{k\in[K]} = \left(\frac{\exp\left(\mathbf{a}_k + \mathbf{b}_k^\top \mathbf{y}\right)}{\sum_{l=1}^K \exp\left(\mathbf{a}_l + \mathbf{b}_l^\top \mathbf{y}\right)} \right)_{k\in[K]}, \gamma \in \boldsymbol{\Gamma}_S \right\}.$$

$$\boldsymbol{\Gamma}_S = \left\{ \gamma = \left(\left(\mathbf{a}_k\right)_{k\in[K]}, \left(\mathbf{b}_k\right)_{k\in[K]} \right) \in \mathbb{R}^K \times \left(\mathbb{R}^L\right)^K \right\}.$$

$$\mathcal{P}_G = \left\{ \mathbf{y} \mapsto \left(\mathbf{g}_k\left(\mathbf{y};\boldsymbol{\omega}\right)\right)_{k\in[K]} = \left(\frac{\boldsymbol{\pi}_k \Phi_L\left(\mathbf{y}; \mathbf{c}_k, \boldsymbol{\Gamma}_k\right)}{\sum_{j=1}^K \boldsymbol{\pi}_j \Phi_L\left(\mathbf{y}; \mathbf{c}_j, \boldsymbol{\Gamma}_j\right)} \right)_{k\in[K]}, \boldsymbol{\omega} \in \boldsymbol{\Omega} \right\}.$$

Reparameterization of the space of Gaussian gating networks

☼ Reparameterization trick:

$$\mathbf{W}_K = \left\{ \mathbf{w} : \mathbf{y} \mapsto \left(\underbrace{\ln\left(\pi_k \Phi_L\left(\mathbf{y}; \mathbf{c}_k, \boldsymbol{\Gamma}_k\right)\right)}_{\text{Nonlinear}} \right)_{k \in [K]} = \mathbf{w}\left(\mathbf{y}; \boldsymbol{\omega}\right) : \boldsymbol{\omega} \in \boldsymbol{\Omega} \right\}.$$

$$\mathcal{P}_G = \left\{ \mathbf{y} \mapsto \left(\frac{\exp\left(w_k(\mathbf{y})\right)}{\sum_{l=1}^{K} \exp\left(w_l(\mathbf{y})\right)} \right)_{k \in [K]}, \mathbf{w} \in \mathbf{W}_K \right\}.$$

- We obtain non-asymptotic oracle inequality for BLoME models, which is much more challenging compared to LinBoSGaME models [?] since we controlled several difficult bracketing entropies from:
  1. Nonlinear weight function softmax gating networks.
  2. Multi-block-diagonal structures for Gaussian expert covariance matrices.

Reparameterization of the space of Gaussian gating networks

- Reparameterization trick:

$$\mathbf{W}_K = \left\{ \mathbf{w} : \mathbf{y} \mapsto \left(\underbrace{\ln\left(\pi_k \Phi_L\left(\mathbf{y}; \mathbf{c}_k, \boldsymbol{\Gamma}_k\right)\right)}_{\text{Nonlinear}} \right)_{k \in [K]} = \mathbf{w}\left(\mathbf{y}; \omega\right) : \omega \in \boldsymbol{\Omega} \right\}.$$

$$\mathcal{P}_G = \left\{ \mathbf{y} \mapsto \left(\frac{\exp\left(w_k(\mathbf{y})\right)}{\sum_{l=1}^{K} \exp\left(w_l(\mathbf{y})\right)} \right)_{k \in [K]}, \mathbf{w} \in \mathbf{W}_K \right\}.$$

- We obtain non-asymptotic oracle inequality for BLoME models, which is much more challenging compared to LinBoSGaME models [?] since we controlled several difficult bracketing entropies from:

  1. Nonlinear weight function softmax gating networks.
  2. Multi-block-diagonal structures for Gaussian expert covariance matrices.

Asymptotic theory of a single parametric model

Misspecified case: $s_0 \notin S_{\mathbf{m}}$, $\psi^{\star}_m = \operatorname{argmin}_{\psi_{\mathbf{m}} \in \boldsymbol{\Psi}_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, s_{\psi_{\mathbf{m}}}\right)$,

$$S_{\mathbf{m}} = \left\{ (\mathbf{x}, \mathbf{y}) \mapsto s_{\psi_{\mathbf{m}}}(\mathbf{x}|\mathbf{y}) = s_{\mathbf{m}}(\mathbf{x}|\mathbf{y}) : \psi_{\mathbf{m}} \in \boldsymbol{\Psi}_{\mathbf{m}} \subset \mathbb{R}^{\dim(S_{\mathbf{m}})} \right\}.$$

Theorem: [White, 1982, Cohen and Le Pennec, 2011]

- Assumptions: $S_{\mathbf{m}}$ is identifiable and there are some strong regularity assumptions on $\psi_{\mathbf{m}} \mapsto s_{\psi_{\mathbf{m}}}$, $\exists$ $\mathbf{A}\left(\psi_{\mathbf{m}}\right)$ and $\mathbf{B}\left(\psi_{\mathbf{m}}\right)$:
  $\left[\mathbf{A}\left(\psi_{\mathbf{m}}\right)\right]_{k,l} = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\int \frac{-\partial^2 \ln s_{\psi_{\mathbf{m}}}}{\partial \psi_{m,k} \partial \psi_{m,l}}\left(\mathbf{x}|\mathbf{Y}_i\right) s_0\left(\mathbf{x}|\mathbf{Y}_i\right) d\lambda\right]$,
  $\left[\mathbf{B}\left(\psi_{\mathbf{m}}\right)\right]_{k,l} = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\int \frac{\partial \ln s_{\psi_{\mathbf{m}}}}{\partial \psi_{m,k}}\left(\mathbf{x}|\mathbf{Y}_i\right) \frac{\partial \ln s_{\psi_{\mathbf{m}}}}{\partial \psi_{m,l}}\left(\mathbf{x}|\mathbf{Y}_i\right) s_0\left(\mathbf{x}|\mathbf{Y}_i\right) d\lambda\right]$.
- Conclusion: $\mathbb{E}\left[\mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_{\mathbf{m}}\right)\right]$ is asymptotically equivalent to

$$\mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, s_{\psi^{\star}_m}\right) + \frac{1}{2n} tr\left(\mathbf{B}\left(\psi^{\star}_m\right) \mathbf{A}\left(\psi^{\star}_m\right)^{-1}\right).$$

Asymptotic theory of a single parametric model

Misspecified case: $s_0 \notin S_{\mathbf{m}}$, $\psi_m^\star = \operatorname{argmin}_{\psi_{\mathbf{m}} \in \boldsymbol{\Psi}_{\mathbf{m}}} \mathrm{KL}^{\otimes n}\left(s_0, s_{\psi_{\mathbf{m}}}\right)$,

$$S_{\mathbf{m}} = \left\{(\mathbf{x}, \mathbf{y}) \mapsto s_{\psi_{\mathbf{m}}}(\mathbf{x}|\mathbf{y}) = s_{\mathbf{m}}(\mathbf{x}|\mathbf{y}) : \psi_{\mathbf{m}} \in \boldsymbol{\Psi}_{\mathbf{m}} \subset \mathbb{R}^{\dim(S_{\mathbf{m}})}\right\}.$$

Theorem: [White, 1982, Cohen and Le Pennec, 2011]

- Assumptions: $S_{\mathbf{m}}$ is identifiable and there are some strong regularity assumptions on $\psi_{\mathbf{m}} \mapsto s_{\psi_{\mathbf{m}}}$, $\exists$ $\mathbf{A}\left(\psi_{\mathbf{m}}\right)$ and $\mathbf{B}\left(\psi_{\mathbf{m}}\right)$:
  $\left[\mathbf{A}\left(\psi_{\mathbf{m}}\right)\right]_{k,l} = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \int \frac{-\partial^2 \ln s_{\psi_{\mathbf{m}}}}{\partial\psi_{m,k}\partial\psi_{m,l}}\left(\mathbf{x}|\mathbf{Y}_i\right) s_0\left(\mathbf{x}|\mathbf{Y}_i\right) d\lambda\right]$,
  $\left[\mathbf{B}\left(\psi_{\mathbf{m}}\right)\right]_{k,l} = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^n \int \frac{\partial \ln s_{\psi_{\mathbf{m}}}}{\partial\psi_{m,k}}\left(\mathbf{x}|\mathbf{Y}_i\right) \frac{\partial \ln s_{\psi_{\mathbf{m}}}}{\partial\psi_{m,l}}\left(\mathbf{x}|\mathbf{Y}_i\right) s_0\left(\mathbf{x}|\mathbf{Y}_i\right) d\lambda\right]$.
- Conclusion: $\mathbb{E}\left[\mathrm{KL}^{\otimes n}\left(s_0, \widehat{s}_{\mathbf{m}}\right)\right]$ is asymptotically equivalent to

$$\mathrm{KL}^{\otimes n}\left(s_0, s_{\psi_m^\star}\right) + \frac{1}{2n} tr\left(\mathbf{B}\left(\psi_m^\star\right)\mathbf{A}\left(\psi_m^\star\right)^{-1}\right).$$

Asymptotic theory of a single parametric model

Well-specified case: $s_0 \in S_{\mathbf{m}}$, $\psi_m^\star = \operatorname{argmin}_{\psi_{\mathbf{m}} \in \Psi_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, s_{\psi_{\mathbf{m}}}\right)$.

Theorem: [White, 1982, Cohen and Le Pennec, 2011]

- It holds that

$$s_0 = s_{\psi_m^\star}, \mathbf{A}\left(\psi_m^\star\right) = \mathbf{B}\left(\psi_m^\star\right).$$

- The same assumption in misspecified case: $\mathbb{E}\left[\mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_{\mathbf{m}}\right)\right]$ is asymptotically equivalent to

$$\underbrace{\mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, s_{\psi_m^\star}\right)}_{=0} + \frac{1}{2n} \dim\left(S_{\mathbf{m}}\right).$$

Asymptotic theory of a single parametric model

Well-specified case: $s_0 \in S_{\mathbf{m}}$, $\psi^\star_m = \operatorname{argmin}_{\psi_{\mathbf{m}} \in \boldsymbol{\Psi}_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}} \left(s_0, s_{\psi_{\mathbf{m}}}\right)$.

Theorem: [White, 1982, Cohen and Le Pennec, 2011]

- It holds that

$$s_0 = s_{\psi^\star_m}, \mathbf{A}\left(\psi^\star_m\right) = \mathbf{B}\left(\psi^\star_m\right).$$

- The same assumption in misspecified case: $\mathbb{E}\left[\mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_{\mathbf{m}}\right)\right]$ is asymptotically equivalent to

$$\underbrace{\mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, s_{\psi^\star_m}\right)}_{=0} + \frac{1}{2n} \dim\left(S_{\mathbf{m}}\right).$$

Drawbacks of asymptotic theory

Drawbacks: **asymptotic normality** of $\sqrt{n}\left(\widehat{\psi}_m - \psi_m^\star\right)$ is required!

➡ Some previous ideas to handle **non-asymptotic normality**:

- Extension in non parametric case or non-identifiable model, Wilk's phenomenon, [Wilks, 1938].
- Generalization of the corresponding Chi-Square goodness-of-fit test [Fan et al., 2001].
- Finite sample deviation of the corresponding empirical quantity in a bounded loss setting [Boucheron and Massart, 2011].

Our targets: Obtain an upper bound on similar expected loss:

✔ Holds for finite sample.

✖ Strong regularity assumptions of [White, 1982].

Drawbacks of asymptotic theory

Drawbacks: **asymptotic normality** of $\sqrt{n}\left(\widehat{\psi}_m - \psi_m^\star\right)$ is required!

➡ Some previous ideas to handle **non-asymptotic normality**:

- Extension in non parametric case or non-identifiable model, Wilk's phenomenon, [Wilks, 1938].
- Generalization of the corresponding Chi-Square goodness-of-fit test [Fan et al., 2001].
- Finite sample deviation of the corresponding empirical quantity in a bounded loss setting [Boucheron and Massart, 2011].

Our targets: Obtain an upper bound on similar expected loss:

✔ Holds for finite sample.

✖ Strong regularity assumptions of [White, 1982].

Drawbacks of asymptotic theory

Drawbacks: **asymptotic normality** of $\sqrt{n}\left(\widehat{\psi}_m - \psi^\star_m\right)$ is required!

Some previous ideas to handle **non-asymptotic normality**:

- Extension in non parametric case or non-identifiable model, Wilk's phenomenon, [Wilks, 1938].
- Generalization of the corresponding Chi-Square goodness-of-fit test [Fan et al., 2001].
- Finite sample deviation of the corresponding empirical quantity in a bounded loss setting [Boucheron and Massart, 2011].

Our targets: Obtain an upper bound on similar expected loss:

- ✔ Holds for finite sample.
- ✖ Strong regularity assumptions of [White, 1982].

Weak oracle inequalities

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

Potential issues

1. Different divergences: $\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_{\widehat{\mathbf{m}}}\right) \leq \mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_{\widehat{\mathbf{m}}}\right)$.
2. $C_1 > 1$ and misspecified case: Given fixed collection $\mathcal{M}$, as $n \to \infty$, the error bound $\to C_1 \inf_{s_{\mathbf{m}}\in S_{\mathbf{m}}} \mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, s_{\mathbf{m}}\right)$ (potentially large!).

Weak oracle inequalities

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

Potential issues

1. Different divergences: $\mathrm{JKL}_{\rho}^{\otimes n}\left(s_0, \widehat{s}_{\widehat{\mathbf{m}}}\right) \leq \mathrm{KL}^{\otimes n}\left(s_0, \widehat{s}_{\widehat{\mathbf{m}}}\right)$.
2. $C_1 > 1$ and misspecified case: Given fixed collection $\mathcal{M}$, as $n \to \infty$, the error bound $\to C_1 \inf_{s_{\mathbf{m}}\in S_{\mathbf{m}}} \mathrm{KL}^{\otimes n}\left(s_0, s_{\mathbf{m}}\right)$ (potentially large!).

Outline

Universal approximation of finite location-scale mixtures

Given a PDF φ (*e.g.*, standard multivariate normal distribution (MND)),

$$\mathcal{S}^{\varphi} = \bigcup_{K\in\mathbb{N}^\star} \mathcal{S}_K^{\varphi}, \text{ where } \mathcal{S}_K^{\varphi} = \Big\{\mathcal{Y} \ni \mathbf{y} \mapsto s_K^{\varphi}(\mathbf{y}) = \sum_{k=1}^{K} \frac{\boldsymbol{\pi}_k}{\sigma_k^L} \varphi\left(\frac{\mathbf{y}-\boldsymbol{\upsilon}_k}{\sigma_k}\right),$$
$$\boldsymbol{\upsilon}_k \in \mathbb{R}^L, \sigma_k \in \mathbb{R}^+, \boldsymbol{\pi} \in \boldsymbol{\Pi}_{K-1}\Big\}.$$

$\mathcal{L}_\infty$: essentially bounded function, $\mathcal{L}_p$: Lebesgue PDF.

Theorem: [Nguyen et al., 2020b, Nguyen et al., 2020a]

- Given any PDFs $s_0, \varphi \in \mathcal{C}$ and a compact set $\mathcal{Y} \subset \mathbb{R}^L$, there exists $\{s_K^{\varphi}\}_{K\in\mathbb{N}^\star} \subset \mathcal{S}^{\varphi}$, $\lim_{K\to\infty} \sup_{\mathbf{y}\in\mathcal{Y}} |s_0(\mathbf{y}) - s_K^{\varphi}(\mathbf{y})| = 0$.
- For $p \in [1,\infty)$, if $s_0 \in \mathcal{L}_p$ and $\varphi \in \mathcal{L}_\infty$, there exists $\{s_K^{\varphi}\}_{K\in\mathbb{N}^\star} \subset \mathcal{S}^{\varphi}$, $\lim_{K\to\infty} \|s_0 - s_K^{\varphi}\|_{\mathcal{L}_p} = 0$.

✸ $\mathcal{S}_K^{\varphi} \not\subseteq S_{\mathbf{m}}$!

Universal approximation of finite location-scale mixtures

Given a PDF φ (*e.g.*, standard multivariate normal distribution (MND)),

$$\mathcal{S}^{\varphi} = \bigcup_{K \in \mathbb{N}^\star} \mathcal{S}_K^{\varphi}, \text{ where } \mathcal{S}_K^{\varphi} = \Big\{ \mathcal{Y} \ni \mathbf{y} \mapsto s_K^{\varphi}(\mathbf{y}) = \sum_{k=1}^{K} \frac{\boldsymbol{\pi}_k}{\sigma_k^L} \varphi\left(\frac{\mathbf{y} - \boldsymbol{\upsilon}_k}{\sigma_k}\right), \\ \boldsymbol{\upsilon}_k \in \mathbb{R}^L, \sigma_k \in \mathbb{R}^+, \boldsymbol{\pi} \in \boldsymbol{\Pi}_{K-1} \Big\}.$$

$\mathcal{L}_\infty$: essentially bounded function, $\mathcal{L}_p$: Lebesgue PDF.

Theorem: [Nguyen et al., 2020b, Nguyen et al., 2020a]

- Given any PDFs $s_0, \varphi \in \mathcal{C}$ and a compact set $\mathcal{Y} \subset \mathbb{R}^L$, there exists $\{s_K^{\varphi}\}_{K \in \mathbb{N}^\star} \subset \mathcal{S}^{\varphi}$, $\lim_{K \to \infty} \sup_{\mathbf{y} \in \mathcal{Y}} |s_0(\mathbf{y}) - s_K^{\varphi}(\mathbf{y})| = 0$.
- For $p \in [1, \infty)$, if $s_0 \in \mathcal{L}_p$ and $\varphi \in \mathcal{L}_\infty$, there exists $\{s_K^{\varphi}\}_{K \in \mathbb{N}^\star} \subset \mathcal{S}^{\varphi}$, $\lim_{K \to \infty} \|s_0 - s_K^{\varphi}\|_{\mathcal{L}_p} = 0$.

✸ $\mathcal{S}_K^{\varphi} \not\subseteq S_{\mathbf{m}}$!

Outline

5 Boundedness conditions

6 Non-asymptotic approach for model selection in MoE models

7 Universal approximation theorems of MoE models
- Finite location-scale MoE models
- Inverse regression trick

8 Softmax gating networks

Definition: Isotropic SGaME and GLLiM models

- Location-scale family: given a PDF φ, $\boldsymbol{\upsilon} \in \mathcal{X}$, $\sigma \in R^{+}$,

$$\mathcal{E}_{\varphi} = \left\{ \mathbf{x} \mapsto \frac{1}{\sigma^{D}} \varphi \left(\frac{\mathbf{x} - \boldsymbol{\upsilon}}{\sigma} \right) = \Phi_{D}(\mathbf{x}; \boldsymbol{\upsilon}, \sigma) \right\}.$$

- Isotropic SGaME models ($\subset S_{\mathbf{m}}$):

$$\mathcal{S}_{S}^{\varphi} = \Big\{ (\mathbf{x}, \mathbf{y}) \mapsto s_{K}^{\varphi}(\mathbf{x}|\mathbf{y}) = \sum_{k=1}^{K} \mathbf{g}_{k} \left(\mathbf{y}; \gamma \right) \Phi_{D} \left(\mathbf{x}; \boldsymbol{\upsilon}_{k}, \sigma_{k} \right),$$
$$\Phi_{D} \in \mathcal{E}_{\varphi} \cap \mathcal{L}_{\infty}, \mathbf{g}_{k} \left(\cdot ; \gamma \right) \in \mathcal{P}_{S}^{K}, \in \mathbb{N}^{\star} \Big\}.$$

- Isotropic GLLiM models ($\subset S_{\mathbf{m}}$):

$$\mathcal{S}_{G}^{\varphi} = \Big\{ (\mathbf{x}, \mathbf{y}) \mapsto s_{K}^{\varphi}(\mathbf{x}|\mathbf{y}) = \sum_{k=1}^{K} \mathbf{g}_{k} \left(\mathbf{y}; \omega \right) \Phi_{D} \left(\mathbf{x}; \boldsymbol{\upsilon}_{k}, \sigma_{k} \right),$$
$$\Phi_{D} \in \mathcal{E}_{\varphi} \cap \mathcal{L}_{\infty}, \mathbf{g}_{k} \left(\cdot ; \omega \right) \in \mathcal{P}_{G}^{K}, K \in \mathbb{N}^{\star} \Big\}.$$

Definition: Isotropic SGaME and GLLiM models

- Location-scale family: given a PDF φ, $\boldsymbol{\upsilon} \in \mathcal{X}$, $\sigma \in R^{+}$,

$$\mathcal{E}_{\varphi} = \left\{ \mathbf{x} \mapsto \frac{1}{\sigma^{D}} \varphi\left(\frac{\mathbf{x} - \boldsymbol{\upsilon}}{\sigma} \right) = \Phi_{D}(\mathbf{x}; \boldsymbol{\upsilon}, \sigma) \right\}.$$

- Isotropic SGaME models ($\subset S_{\mathbf{m}}$):

$$\mathcal{S}_{S}^{\varphi} = \left\{ (\mathbf{x}, \mathbf{y}) \mapsto s_{K}^{\varphi}(\mathbf{x}|\mathbf{y}) = \sum_{k=1}^{K} \mathbf{g}_{k}\left(\mathbf{y}; \gamma\right) \Phi_{D}\left(\mathbf{x}; \boldsymbol{\upsilon}_{k}, \sigma_{k}\right), \right.$$

$$\left. \Phi_{D} \in \mathcal{E}_{\varphi} \cap \mathcal{L}_{\infty}, \mathbf{g}_{k}\left(\cdot; \gamma\right) \in \mathcal{P}_{S}^{K}, \in \mathbb{N}^{\star} \right\}.$$

- Isotropic GLLiM models ($\subset S_{\mathbf{m}}$):

$$\mathcal{S}_{G}^{\varphi} = \left\{ (\mathbf{x}, \mathbf{y}) \mapsto s_{K}^{\varphi}(\mathbf{x}|\mathbf{y}) = \sum_{k=1}^{K} \mathbf{g}_{k}\left(\mathbf{y}; \boldsymbol{\omega}\right) \Phi_{D}\left(\mathbf{x}; \boldsymbol{\upsilon}_{k}, \sigma_{k}\right), \right.$$

$$\left. \Phi_{D} \in \mathcal{E}_{\varphi} \cap \mathcal{L}_{\infty}, \mathbf{g}_{k}\left(\cdot; \boldsymbol{\omega}\right) \in \mathcal{P}_{G}^{K}, K \in \mathbb{N}^{\star} \right\}.$$

Theorem: Approximation capabilities of isotropic SGaME and GLLiM models [Nguyen et al., 2021a]

(a) Given $p \in [1, \infty)$, a compact set $\mathcal{Y} \subset \mathbb{R}^L$, $\varphi \in \mathcal{F} \cap \mathcal{C}$, for target $s_0 \in \mathcal{F}_p \cap \mathcal{C}_b^u$, there exist sequences $\{s_K^{\varphi}\}_{K \in \mathbb{N}^\star} \subset \mathcal{S}_S^{\varphi}$ and $\{s_K'^{\varphi}\}_{K \in \mathbb{N}^\star} \subset \mathcal{S}_G^{\varphi}$ such that

$$\lim_{K \to \infty} \|s_0 - s_K^{\varphi}\|_{\mathcal{L}_p} = 0,$$

$$\lim_{K \to \infty} \left\|s_0 - s_K'^{\varphi}\right\|_{\mathcal{L}_p} = 0.$$

(b) Given $L = 1$, and $0 < \lambda(\mathcal{X}) < \infty$, $\varphi \in \mathcal{F} \cap \mathcal{C}_b^u$, for any target $s_0 \in \mathcal{F} \cap \mathcal{C}_b^u$, there exist sequences $\{s_K^{\varphi}\}_{K \in \mathbb{N}^\star} \subset \mathcal{S}_S^{\varphi}$ and $\{s_K'^{\varphi}\}_{K \in \mathbb{N}^\star} \subset \mathcal{S}_G^{\varphi}$ such that

$$\lim_{K \to \infty} s_K^{\varphi} = s_0 \text{ almost uniformly},$$

$$\lim_{K \to \infty} s_K'^{\varphi} = s_0 \text{ almost uniformly}.$$

Theorem: Approximation capabilities of isotropic SGaME and GLLiM models [Nguyen et al., 2021a]

(a) Given $p \in [1, \infty)$, a compact set $\mathcal{Y} \subset \mathbb{R}^L$, $\varphi \in \mathcal{F} \cap \mathcal{C}$, for target $s_0 \in \mathcal{F}_p \cap \mathcal{C}_b^u$, there exist sequences $\{s_K^\varphi\}_{K \in \mathbb{N}^\star} \subset \mathcal{S}_S^\varphi$ and $\{s_K'^\varphi\}_{K \in \mathbb{N}^\star} \subset \mathcal{S}_G^\varphi$ such that

$$\lim_{K \to \infty} \|s_0 - s_K^\varphi\|_{\mathcal{L}_p} = 0,$$

$$\lim_{K \to \infty} \|s_0 - s_K'^\varphi\|_{\mathcal{L}_p} = 0.$$

(b) Given $L = 1$, and $0 < \lambda(\mathcal{X}) < \infty$, $\varphi \in \mathcal{F} \cap \mathcal{C}_b^u$, for any target $s_0 \in \mathcal{F} \cap \mathcal{C}_b^u$, there exist sequences $\{s_K^\varphi\}_{K \in \mathbb{N}^\star} \subset \mathcal{S}_S^\varphi$ and $\{s_K'^\varphi\}_{K \in \mathbb{N}^\star} \subset \mathcal{S}_G^\varphi$ such that

$$\lim_{K \to \infty} s_K^\varphi = s_0 \text{ almost uniformly,}$$

$$\lim_{K \to \infty} s_K'^\varphi = s_0 \text{ almost uniformly.}$$

Mild assumption: Linear combination of bounded functions whose coefficients belong to a compact set

- Weights: $\mathbf{W}_{K,d_{\mathbf{W}}} = \{0\} \otimes \mathbf{W}_{Bo,d_{\mathbf{W}}}^{K-1}$ (identifiability condition),

$$\mathbf{W}_{Bo,d_{\mathbf{W}}} = \left\{ \mathbf{x} \mapsto \sum_{d=1}^{d_{\mathbf{W}}} \omega_d \theta_{\mathbf{W},d}(\mathbf{x}) : \max_{d \in [d_{\mathbf{W}}]} |\omega_d| \leq T_{\mathbf{W}} \right\}.$$

- Gaussian expert means: $\Upsilon_{K,d_\Upsilon} = \Upsilon_{Bo,d_\Upsilon}^{K}$,

$$\Upsilon_{Bo,d_\Upsilon} = \left\{ \mathbf{x} \mapsto \left(\sum_{d=1}^{d_\Upsilon} \beta_d^{(z)} \theta_{\Upsilon,d}(\mathbf{x}) \right)_{z \in [q]} : \max_{d \in [d_\Upsilon], z \in [q]} \left| \beta_d^{(z)} \right| \leq T_\Upsilon \right\}.$$

- Collections of bounded basis functions: $\mathbf{x} \mapsto \left(\theta_{\mathbf{W},d}(\mathbf{x}), \theta_{\Upsilon,d}(\mathbf{x}) \right)$, $T_{\mathbf{W}} \in \mathbb{R}^+$, $T_\Upsilon \in \mathbb{R}^+$.

Mild assumption: Polynomials on compact sets to **simplify the interpretation of sparsity in high-dimensional data**

- Weights: $\mathbf{W}_{K,d_{\mathbf{W}}} = \{0\} \otimes \mathbf{W}_{Po,d_{\mathbf{W}}}^{K-1}$ (identifiability condition), $T_{\mathbf{W}} \in \mathbb{R}^{+}$,

$$\mathbf{W}_{Po,d_{\mathbf{W}}} = \left\{ \mathbf{x} \mapsto \sum_{|\boldsymbol{\alpha}|=0}^{d_{\mathbf{W}}} \omega_{\boldsymbol{\alpha}} \mathbf{x}^{\boldsymbol{\alpha}} \in \mathbb{R} : \max_{\boldsymbol{\alpha} \in \mathcal{A}} |\omega_{\boldsymbol{\alpha}}| \leq T_{\mathbf{W}} \right\}.$$

- Gaussian expert means: $|||\mathbf{A}|||_{\infty} = \max_{i \in [q], j \in [p]} \left| [\mathbf{A}]_{i,j} \right|$, $T_{\boldsymbol{\Upsilon}} \in \mathbb{R}^{+}$,

$$\boldsymbol{\Upsilon}_{K,d_{\boldsymbol{\Upsilon}}} = \Bigg\{ \mathbf{x} \mapsto \left(\beta_{k0} + \sum_{d=1}^{d_{\boldsymbol{\Upsilon}}} \beta_{kd} \mathbf{x}^{d} \right)_{k \in [K]} : \max \left\{ |||\beta_{kd}|||_{\infty} : k \in [K], d \in (\{0\} \cup [d_{\boldsymbol{\Upsilon}}]) \right\} \leq T_{\boldsymbol{\Upsilon}} \Bigg\},$$

High-dimensional regression data: $\mathcal{X} = [0,1]^{p}$, $\mathcal{Y} \subset \mathbb{R}^{q}$, with $p, q \gg n$, see [Nguyen et al., 2020c] for example.

Mild assumption: Boundedness conditions on eigenvalues of Gaussian expert block-diagonal covariance matrices

$$0 < \lambda_m \leq m\left(\Sigma_k\left(\mathbf{B}_k\right)\right) \leq M\left(\Sigma_k\left(\mathbf{B}_k\right)\right) \leq \lambda_M, \text{ for every } k \in [K].$$

↑ constant ↑ smallest ↑ largest eigenvalue

$$\left\{ \Sigma_k\left(\mathbf{B}_k\right) \in \mathcal{S}_q^{++} \;\middle|\; \begin{array}{l} \Sigma_k\left(\mathbf{B}_k\right) = \mathbf{P}_k \begin{pmatrix} \Sigma_k^{[1]} & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & \Sigma_k^{[2]} & \dots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \ddots & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \dots & \Sigma_k^{[G_k]} \end{pmatrix} \mathbf{P}_k^{-1}, \\ \Sigma_k^{[g]} \in \mathcal{S}^{++}_{\mathrm{card}\left(d_k^{[g]}\right)}, \forall g \in [G_k] \end{array} \right\}$$

↑

$\mathbf{V}_K\left(\mathbf{B}\right) = \left(\mathbf{V}_k\left(\mathbf{B}_k\right)\right)_{k \in [K]}$, $\mathbf{P}_k$: permutation, $\mathbf{B}_k = \left(d_k^{[g]}\right)_{g \in [G_k]}$: block structure, $d_k^{[g]}$: set of variables in gth group, $\mathbf{B} = \left(\mathbf{B}_k\right)_{k \in [K]}$.

Definition: Variable selection via selecting relevant variables

- A couple $(\mathbf{Y}_z, \mathbf{X}_j)$ and its corresponding indices $(z, j) \in [q] \times [p]$:
  - irrelevant if $\mathbf{X}_j$ does not explain the variable $\mathbf{Y}_z$, *i.e.*,

$$[\beta_{kd}]_{z,j} = 0,\ \omega_k^{[j,l]} = \mathbf{0}, \text{ for all } k \in [K],\ d \in [d_{\Upsilon}],\ l \in [d_{\mathbf{W}}]:$$

    - β_{kd}, $k \in [K]$, $d \in [d_{\Upsilon}]$: matrices of d-th term regression coefficients of k-th mixture component.
    - $\omega_k^{[j,l]} = \{\omega_{k\boldsymbol{\alpha}} \in \mathbb{R} : \boldsymbol{\alpha} \in \mathcal{A}_l, \boldsymbol{\alpha}_j > 0\}$, $\forall l \in [d_{\mathbf{W}}]$, $j \in [p]$: vector of monomial coefficients of l-th degree involving $\mathbf{x}_j$,
    - $\mathcal{A}_l = \left\{\boldsymbol{\alpha} = (\alpha_t)_{t \in [p]} \in \mathbb{N}^p, |\boldsymbol{\alpha}| = l\right\}$, $|\boldsymbol{\alpha}|$ degree of monomials $\mathbf{x}^{\boldsymbol{\alpha}}$.
  - relevant if they are not irrelevant.
- Sparse model: few of relevant variables.
- $\mathbf{J} = \{(z, j) \in [q] \times [p] : (\mathbf{Y}_z, \mathbf{X}_j) \text{ are relevant couples}\}$.
- The set of **relevant variables (columns)**:
  $\mathbf{J}_\omega = \{j \in [p] : \exists z \in [q], (z, j) \in \mathbf{J}\}$.

Definition: Variable selection via selecting relevant variables

- A couple $(\mathbf{Y}_z, \mathbf{X}_j)$ and its corresponding indices $(z, j) \in [q] \times [p]$:
  - irrelevant if $\mathbf{X}_j$ does not explain the variable $\mathbf{Y}_z$, *i.e.*,

$$[\beta_{kd}]_{z,j} = 0,\ \omega_k^{[j,l]} = \mathbf{0},\ \text{for all } k \in [K],\ d \in [d_\Upsilon],\ l \in [d_W];$$

  - relevant if they are not irrelevant.
- Sparse model: few of relevant variables.
- $\mathbf{J} = \{(z, j) \in [q] \times [p] : (\mathbf{Y}_z, \mathbf{X}_j) \text{ are relevant couples}\}$.
- The set of **relevant variables (columns)**:
  $\mathbf{J}_\omega = \{j \in [p] : \exists z \in [q], (z, j) \in \mathbf{J}\}$.

Definition: Variable selection via selecting relevant variables

- A couple $(\mathbf{Y}_z, \mathbf{X}_j)$ and its corresponding indices $(z, j) \in [q] \times [p]$:
  - irrelevant if $\mathbf{X}_j$ does not explain the variable $\mathbf{Y}_z$, *i.e.*,

$$[\beta_{kd}]_{z,j} = 0,\ \omega_k^{[j,l]} = \mathbf{0},\ \text{for all } k \in [K],\ d \in [d_\Upsilon],\ l \in [d_W];$$

  - relevant if they are not irrelevant.
- Sparse model: few of relevant variables.
- $\mathbf{J} = \{(z, j) \in [q] \times [p] : (\mathbf{Y}_z, \mathbf{X}_j) \text{ are relevant couples}\}$.
- The set of **relevant variables (columns)**:
  $\mathbf{J}_\omega = \{j \in [p] : \exists z \in [q], (z, j) \in \mathbf{J}\}$.

Definition: Low-rank regression matrices

- **Low-rank matrices**: Regression coefficients β_{kd}, $k \in [K]$, $d \in [L]$: can be **well approximated by low-rank matrices**.
  - β_{kd} is fully determined by $R_{kd}\left(p+q-R_{kd}\right)$ coefficients if $\operatorname{rank}\left(\beta_{kd}\right) = R_{kd}$.
  - The total parameters to be estimate **may be smaller** than the sample size nq.
- **Vector of ranks**: **Combining rank and column sparsity**. $\mathbf{R} = \left(R_{kd}\right)_{k\in[K],d\in[d_\Upsilon]} \in \left[\operatorname{card}\left(\mathbf{J}_\omega\right) \wedge q\right]^{d_\Upsilon K}$, where $\operatorname{rank}\left(\beta_{kd}\right) = R_{kd}$, $a \wedge b = \min\left(a, b\right)$.

- $\mathbf{J} \subset \mathcal{P}\left([q] \times [p]\right)$ and $\mathbf{J}_\omega \subset \mathcal{P}\left([q]\right)$, where $\mathcal{P}\left([q] \times [p]\right)$ contains all subsets of $[q] \times [p]$.
- If for all $k \in [K]$, $d \in [d_\Upsilon]$, a matrix β_{kd} has $\operatorname{card}\left(\mathbf{J}_\omega\right)$ relevant columns ➔ there are $q \operatorname{card}\left(\mathbf{J}_\omega\right)$ coefficients $\ll qp$ per clusters if when $\operatorname{card}\left(\mathbf{J}_\omega\right) \ll p$.

Definition: Low-rank regression matrices

- **Low-rank matrices**: Regression coefficients β_{kd}, $k \in [K]$, $d \in [L]$: can be **well approximated by low-rank matrices**.
  - β_{kd} is fully determined by $R_{kd}\left(p+q-R_{kd}\right)$ coefficients if $\operatorname{rank}\left(\beta_{kd}\right) = R_{kd}$.
  - The total parameters to be estimate **may be smaller** than the sample size nq.
- **Vector of ranks**: **Combining rank and column sparsity**. $\mathbf{R} = \left(R_{kd}\right)_{k\in[K],d\in[d_\Upsilon]} \in \left[\operatorname{card}\left(\mathbf{J}_\omega\right) \wedge q\right]^{d_\Upsilon K}$, where $\operatorname{rank}\left(\beta_{kd}\right) = R_{kd}$, $a \wedge b = \min\left(a, b\right)$.

- $\mathbf{J} \subset \mathcal{P}\left([q] \times [p]\right)$ and $\mathbf{J}_\omega \subset \mathcal{P}\left([q]\right)$, where $\mathcal{P}\left([q] \times [p]\right)$ contains all subsets of $[q] \times [p]$.
- If for all $k \in [K]$, $d \in [d_\Upsilon]$, a matrix β_{kd} has $\operatorname{card}\left(\mathbf{J}_\omega\right)$ relevant columns ➔ there are $q \operatorname{card}\left(\mathbf{J}_\omega\right)$ coefficients $\ll qp$ per clusters if when $\operatorname{card}\left(\mathbf{J}_\omega\right) \ll p$.

Definition: Essentially bounded and Lebesgue conditional PDF

- Essentially bounded function on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$:

$$\mathcal{L}_\infty(\mathcal{Z}) = \left\{ f : \underbrace{\inf \{ a \geq 0 : \lambda (\{ \mathbf{z} \in \mathcal{Z} : |f(\mathbf{z})| > a \}) = 0 \}}_{= \|f\|_{\infty, \mathcal{Z}}} < \infty \right\}.$$

- Lebesgue conditional PDF: $\mathcal{F}_p = \mathcal{F} \cap \mathcal{L}_p$, $p \in [1, \infty)$,

$$\mathcal{F} = \left\{ f : \mathcal{Z} \to [0, \infty), \int_{\mathcal{Y}} f(\mathbf{x}|\mathbf{y}) d\lambda(\mathbf{x}) = 1 \right\},$$

$$\mathcal{L}_p(\mathcal{Z}) = \left\{ f := \underbrace{\left(\int_{\mathcal{Z}} |f(\mathbf{z})|^p \, d\lambda(\mathbf{z}) \right)^{1/p}}_{= \|f\|_{p, \mathcal{Z}}} < \infty \right\}.$$

Definition: Essentially bounded and Lebesgue conditional PDF

- Essentially bounded function on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$:

$$\mathcal{L}_\infty(\mathcal{Z}) = \left\{ f : \underbrace{\inf \left\{ a \geq 0 : \lambda \left(\left\{ \mathbf{z} \in \mathcal{Z} : |f(\mathbf{z})| > a \right\} \right) = 0 \right\}}_{=\|f\|_{\infty,\mathcal{Z}}} < \infty \right\}.$$

- Lebesgue conditional PDF: $\mathcal{F}_p = \mathcal{F} \cap \mathcal{L}_p$, $p \in [1, \infty)$,

$$\mathcal{F} = \left\{ f : \mathcal{Z} \to [0, \infty), \int_{\mathcal{Y}} f(\mathbf{x}|\mathbf{y}) d\lambda(\mathbf{x}) = 1 \right\},$$

$$\mathcal{L}_p(\mathcal{Z}) = \left\{ f := \underbrace{\left(\int_{\mathcal{Z}} |f(\mathbf{z})|^p \, d\lambda(\mathbf{z}) \right)^{1/p}}_{=\|f\|_{p,\mathcal{Z}}} < \infty \right\}.$$

Definition: softmax gating networks

$$g_{\mathbf{w},k,d_W}(\mathbf{x}) = \underbrace{\frac{\exp\left(w_{k,d_W}(\mathbf{x})\right)}{\sum_{l=1}^{K}\exp\left(w_{l,d_W}(\mathbf{x})\right)}}_{\text{normalized exponential function}}, \text{ for every } k \in [K],$$

- $\mathbf{w} = \left(w_{k,d_W}\right)_{k\in[K]} \in \mathbf{W}_{K,d_W}$: functions defined in logistic schemes (weights),
- For every $\mathbf{x} \in \mathcal{X}$, $\left(g_{\mathbf{w},k,d_W}(\mathbf{x})\right)_{k\in[K]} \in \mathbf{\Pi}_{K-1}$,
- $\mathbf{\Pi}_{K-1} = \left\{\left(\pi_k\right)_{k\in[K]} \in \left(\mathbb{R}^+\right)^K, \sum_{k=1}^{K}\pi_k = 1\right\}$.

Appendix: GLLiM model hierarchical definition

$$\mathbf{Y} = \sum_{k=1}^{K} \mathbb{I}\left(Z = k\right)\left(\mathbf{A}_k^*\mathbf{X} + \mathbf{b}_k^* + \mathbf{E}_k^*\right),$$

$\mathbf{Y} \in \mathbb{R}^L$, $\mathbf{X}$(or $\boldsymbol{\theta}$) $\in \mathbb{R}^D$ with $D \gg L$, $\mathbb{I}$ Indicator function, $\mathbf{A}_k^* \in \mathbb{R}^{L\times D}$, $\mathbf{b}_k^* \in \mathbb{R}^L$.
$\boldsymbol{E}_k'$: capturing both observation noise in $\mathbb{R}^L$ and reconstruction error due to the local affine approximations, Gaussian, centered:

$$p\left(\mathbf{Y} = \mathbf{y}|\mathbf{X} = \mathbf{x}, Z = k; \psi_K^*\right) = \mathcal{N}_L\left(\mathbf{y}; \mathbf{A}_k^*\mathbf{x} + \mathbf{b}_k^*, \boldsymbol{\Sigma}_k^*\right),$$

- Affine transformations are local: mixture of K Gaussians

$$p\left(\mathbf{X} = \mathbf{x}|Z = k; \psi_K^*\right) = \mathcal{N}_D\left(\mathbf{x}; \mathbf{c}_k^*, \boldsymbol{\Gamma}_k^*\right), p\left(Z = k; \psi_K^*\right) = \boldsymbol{\pi}_k^*,$$

  where $\mathbf{c}_k^* \in \mathbb{R}^D, \boldsymbol{\Gamma}_k^* \in \mathbb{R}^{D\times D}$, $\boldsymbol{\pi}^* = \left(\boldsymbol{\pi}_k^*\right)_{k\in[K]} \in \boldsymbol{\Pi}_{K-1}$.
- The set of all model parameters is:

$$\psi_K^* = \left(\boldsymbol{\pi}_k^*, \mathbf{c}_k^*, \boldsymbol{\Gamma}_k^*, \mathbf{A}_k^*, \mathbf{b}_k^*, \boldsymbol{\Sigma}_k^*\right)_{k\in[K]}$$

  Usually $\boldsymbol{\Sigma}_k^* = \sigma^2\boldsymbol{I}_L$ for $k = 1\ldots K$ (isotropic reconstruction error).

Outline

Appendix : GLLiM link between ϕ and ϕ'

Idea: $\mathbf{Y} \equiv$ input, $\mathbf{X} \equiv$ output, $\mathbf{X} \subset \mathbb{R}^D$, $\mathbf{Y} \subset \mathbb{R}^L$:

$$p\left(\mathbf{X}=\mathbf{x} | \mathbf{Y}=\mathbf{y}, Z=k ; \psi_K\right)=\phi_D\left(\mathbf{x} ; \mathbf{A}_k \mathbf{y}+\mathbf{b}_k, \boldsymbol{\Sigma}_k\right),$$

$$p\left(\mathbf{Y}=\mathbf{y} | Z=k ; \psi_K\right)=\phi_L\left(\mathbf{y} ; \mathbf{c}_k, \boldsymbol{\Gamma}_k\right), p\left(Z=k ; \psi_k\right)=\pi_k,$$

$$p\left(\mathbf{X}=\mathbf{x} | \mathbf{Y}=\mathbf{y} ; \psi_K\right)=\sum_{k=1}^K \frac{\pi_k \phi_L\left(\mathbf{y} ; \mathbf{c}_k, \boldsymbol{\Gamma}_k\right)}{\sum_{j=1}^K \pi_j \phi_L\left(\mathbf{y} ; \mathbf{c}_j, \boldsymbol{\Gamma}_j\right)} \phi_D\left(\mathbf{x} ; \mathbf{A}_k \mathbf{y}+\mathbf{b}_k, \boldsymbol{\Sigma}_k\right).$$

Link function:

$$\theta_K=\begin{pmatrix}\mathbf{c}_k \\ \boldsymbol{\Gamma}_k \\ \mathbf{A}_k \\ \mathbf{b}_k \\ \boldsymbol{\Sigma}_k\end{pmatrix}_{k \in[K]} \mapsto\begin{pmatrix}\mathbf{c}_k^* \\ \boldsymbol{\Gamma}_k^* \\ \mathbf{A}_k^* \\ \mathbf{b}_k^* \\ \boldsymbol{\Sigma}_k^*\end{pmatrix}_{k \in[K]}=\begin{pmatrix}\mathbf{A}_k \mathbf{c}_k+\mathbf{b}_k \\ \boldsymbol{\Sigma}_k+\mathbf{A}_k \boldsymbol{\Gamma}_k \mathbf{A}_k^{\top} \\ \boldsymbol{\Sigma}_k^* \mathbf{A}_k^{\top} \boldsymbol{\Sigma}_k^{-1} \\ \boldsymbol{\Sigma}_k^*\left(\boldsymbol{\Gamma}_k^{-1} \mathbf{c}_k-\mathbf{A}_k^{\top} \boldsymbol{\Sigma}_k^{-1} \mathbf{b}_k\right) \\ \left(\boldsymbol{\Gamma}_k^{-1}+\mathbf{A}_k^{\top} \boldsymbol{\Sigma}_k^{-1} \mathbf{A}_k\right)^{-1}\end{pmatrix}_{k \in[K]} \in \boldsymbol{\Theta}_K^*,$$

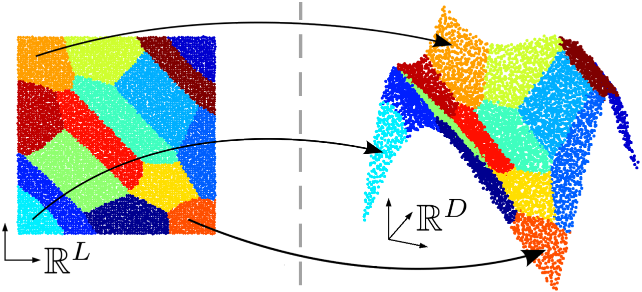
with the note that $\boldsymbol{\pi}^* \equiv \boldsymbol{\pi}$.

Reduce the number of parameters: The number of parameters depends on the GLLiM variant but is in $\mathcal{O}(DKL)$

If diagonal covariances $\boldsymbol{\Sigma}_k$, the number of parameters is $K-1+K(L+L(L+1)/2+DL+2D)$ $\rightarrow$ for $K=100$, $L=4$ and $D=10$ leads to 7499 parameters and to 61499 parameters if $D=100$.

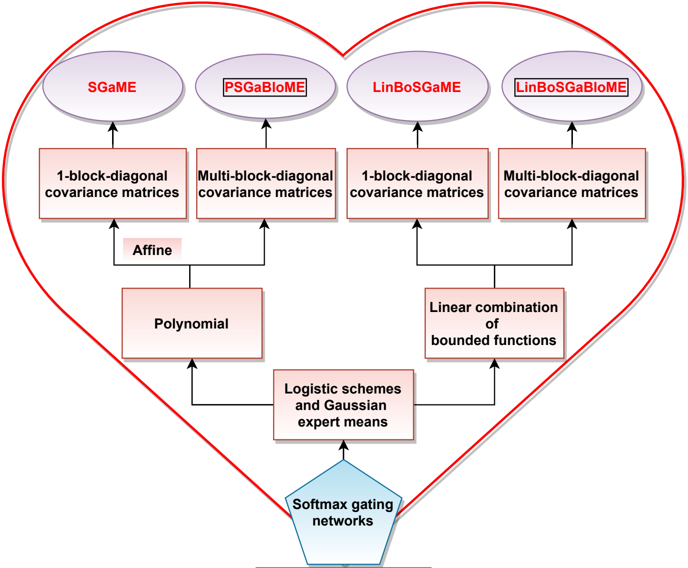
Appendix: GLLiM Geometric Interpretation

Interpretation: This model induces a partition of $\mathbb{R}^L$ into K regions $\mathcal{R}_k$ where the transformation is the most probable. If $|\boldsymbol{\Gamma}_1| = \cdots = |\boldsymbol{\Gamma}_K|$: $\{\mathcal{R}_k, k = 1 \dots K\}$ define a Voronoi diagram of centroids $\{\boldsymbol{c}_k, k = 1 \dots K\}$ (Mahalanobis distance $||.||_{\boldsymbol{\Gamma}^*}$).



$L = 2, D = 3, K = 15$. The low-to-high regression (shown here) may well be interpreted as a parameterization of an L-dimensional manifold embedded in a D-dimensional space.

Outline



SGaME: Softmax-Gated MoE

PSGaBloME: Polynomial Softmax-Gated Block-diagonal MoE

LinBoSGaME: Linear-combination-of-Bounded-functions Softmax-Gated MoE

LinBoSGaBloME: Linear-combination-of-Bounded-functions Softmax-Gated Block-diagonal MoE

Definition: SGaME, PSGaBloME, LinBoSGaME and LinBoSGaBloME

$$s_{\psi_{K,d_{\mathbf{W}},d_{\mathbf{\Upsilon}},\mathbf{B}}}(\mathbf{y}|\mathbf{x}) = \sum_{k=1}^{K} \underbrace{\frac{\exp\left(w_{k,d_{\mathbf{W}}}(\mathbf{x})\right)}{\sum_{l=1}^{K}\exp\left(w_{l,d_{\mathbf{W}}}(\mathbf{x})\right)}}_{\text{Softmax gating network}} \underbrace{\mathcal{N}_q\left(\mathbf{y}; \upsilon_{k,d_{\mathbf{\Upsilon}}}(\mathbf{x}), \mathbf{\Sigma}_k\left(\mathbf{B}_k\right)\right)}_{\text{Gaussian expert}}.$$

- $\mathbf{w} = \left(w_{k,d_{\mathbf{W}}}\right)_{k\in[K]} \in \mathbf{W}_{K,d_{\mathbf{W}}}$: functions defined in logistic schemes (weights),
- $K \in \mathbb{N}^{\star}$: number of mixture components,
- $d_{\mathbf{W}}, d_{\mathbf{\Upsilon}} \in \mathbb{N}^{\star}$: gating networks and mean functions' hyperparameters, *e.g.*, degrees of polynomials,
- $\mathbf{B} = \left(\mathbf{B}_k\right)_{k\in[K]}$: block-diagonal structures for covariance matrices,
- $\psi_{K,d_{\mathbf{W}},d_{\mathbf{\Upsilon}},\mathbf{B}} = \left(\mathbf{w}, \upsilon, \mathbf{\Sigma}\left(\mathbf{B}\right)\right) \in \mathbf{W}_{K,d_{\mathbf{W}}} \times \mathbf{\Upsilon}_{K,d} \times \mathbf{V}_K\left(\mathbf{B}\right) = \psi_{K,d_{\mathbf{W}},d_{\mathbf{\Upsilon}},\mathbf{B}}$: model parameter.

Definition: Collection of models (SGaME, PSGaBloME, LinBoSGaME and LinBoSGaBloME)

$$S_{\mathbf{m}} = \Big\{ (\mathbf{x}, \mathbf{y}) \mapsto s_{\psi_{(K, L_{\mathbf{W}}, d_{\Upsilon}, \mathbf{B}, \mathbf{J}, \mathbf{R})}}(\mathbf{y}|\mathbf{x}) = s_{\mathbf{m}}(\mathbf{y}|\mathbf{x}) : \\ \mathbf{m} = (K, L_{\mathbf{W}}, d_{\Upsilon}, \mathbf{J}, \mathbf{R}), \psi_{(K, L_{\mathbf{W}}, d_{\Upsilon}, \mathbf{B}, \mathbf{J}, \mathbf{R})} \in \boldsymbol{\Psi}_{(K, L_{\mathbf{W}}, d_{\Upsilon}, \mathbf{B}, \mathbf{J}, \mathbf{R})} \Big\}.$$

- $\mathbf{J} = \{(z, j) \in [q] \times [p] : (\mathbf{Y}_z, \mathbf{X}_j) \text{ are relevant couples}\}$.
- $\mathbf{R}$: vector ranks regression matrices.
- $\mathbf{m} \in \mathcal{M}$.
- $\mathcal{M} = [K_{\max}] \times [L_{\max}] \times [D_{\max}] \times (\mathcal{B}_k)_{k \in [K]} \times \mathcal{P}([q] \times [p]) \times [\min(q, p)]^{D_{\max} K}$.
- $K_{\max}, L_{\max}, D_{\max} \in \mathbb{N}^\star$ may depend on n, $\mathcal{P}(\mathbf{A})$ all subsets of $\mathbf{A}$.
- $\mathcal{B}_k$ = all possible partitions of the covariables indexed by $[q]$.

High-dimensional regression data: $\mathcal{X} \subset \mathbb{R}^p$, $\mathcal{Y} \subset \mathbb{R}^q$, with $p, q \gg n$.

Definition: Random subcollection of models ⟵ large number of models

$$S_{\mathbf{m}} = \Big\{(\mathbf{x},\mathbf{y}) \mapsto s_{\psi_{(K,L_{\mathbf{W}},d_{\Upsilon},\mathbf{B},\mathbf{J},\mathbf{R})}}(\mathbf{y}|\mathbf{x}) = s_{\mathbf{m}}(\mathbf{y}|\mathbf{x}) : \mathbf{m} = (K, L_{\mathbf{W}}, d_{\Upsilon}, \mathbf{B}, \mathbf{J}, \mathbf{R}), \psi_{(K,L_{\mathbf{W}},d_{\Upsilon},\mathbf{B},\mathbf{J},\mathbf{R})} \in \boldsymbol{\Psi}_{(K,L_{\mathbf{W}},d_{\Upsilon},\mathbf{B},\mathbf{J},\mathbf{R})}\Big\}.$$

- $\mathbf{m} \in \widetilde{\mathcal{M}} = [K_{\max}] \times [L_{\max}] \times [D_{\max}] \times (\mathcal{B}_{k,\boldsymbol{\Lambda}})_{k\in[K]} \times \mathcal{J} \times \mathcal{R}_{(K,\mathbf{J},D_{\max})} \subset \mathcal{M}$.
- $\mathcal{B}_{k,\boldsymbol{\Lambda}} = (\mathcal{B}_{k,\lambda})_{\lambda\in\boldsymbol{\Lambda}}$ **(potentially random)** $\subset \mathcal{B}_k$: partition of the variables corresponding to the block-diagonal structure of the adjacency matrix $\mathbf{E}_{k,\lambda} = \left[\mathbb{I}\left\{\left|[\mathbf{S}_k]_{z,z'}\right| > \lambda\right\}\right]_{z\in[q],z'\in[q]}$, based on the thresholded absolute value of the sample covariance matrix $\mathbf{S}_k$ in each cluster $k \in [K]$.
- $\mathcal{J}$ (**relevant couples, potentially random**) $\subset \mathcal{P}([q]\times[p])$.
  $\mathbf{J}_\omega$ relevant variable: $\mathbf{J}_\omega = \{j \in [p] : \exists\, z \in [q], (z,j) \in \mathcal{J}\}$.
  $\uparrow\downarrow$ **Our Lasso+l_2-Rank procedure**
- $\mathcal{R}_{(K,\mathbf{J},D_{\max})}$(**vector ranks regression matrices**) $\subset [\min(\mathrm{card}(\mathbf{J}_\omega), q)]^{D_{\max}K}$.

Key references

TrungTin Nguyen, 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 Methods*.

(Cited on pages 10, 76, 77, 91, and 92.)

TrungTin Nguyen, Nguyen, H. D., Chamroukhi, F., and Forbes, F. (2022). A non-asymptotic approach for model selection via penalization in high-dimensional mixture of experts models. *Electronic Journal of Statistics*.

(Cited on pages 42 and 43.)

TrungTin Nguyen, Nguyen, H. D., Chamroukhi, F., and McLachlan, G. J. (2020). Approximation by finite mixtures of continuous density functions that vanish at infinity. *Cogent Mathematics & Statistics*, 7(1):1750861.

(Cited on pages 10, 76, 77, 91, and 92.)

Nguyen, H. D., TrungTin Nguyen, 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*, 8(1):13.

(Cited on pages 10, 76, 77, 78, 96, and 97.)

Forbes, F., Nguyen, H. D.,TrungTin Nguyen, & Arbel, J. (2022). Summary statistics and discrepancy measures for approximate Bayesian computation via surrogate posteriors. *Statistics and Computing*.

(Not cited.)

Further references I

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19(6):716–723.
(Not cited.)

Arlot, S., & Bach, F. (2009). Data-driven calibration of linear estimators with minimal penalties. *Advances in Neural Information Processing Systems*, (Vol. 22). Curran Associates.
(Cited on pages 59, 60, 61, 62, and 63.)

Arlot, S., & Massart, P. (2009). Data-driven Calibration of Penalties for Least-Squares Regression. *Journal of Machine Learning Research*, 10(10), 245–279.
(Cited on pages 59, 60, 61, 62, and 63.)

Arlot, S., Vincent, B., Baudry, J.-P., Maugis, C., & Michel, B. (2016). capushe: CAlibrating Penalities Using Slope HEuristics. *R package version 1.1.1*, 1(1).
(Cited on page 45.)

Arlot, S. (2019). Minimal penalties and the slope heuristics: a survey. *Journal de la Société Française de Statistique*, 160(3), 1–106.
(Cited on pages 45, 59, 60, 61, 62, and 63.)

Baudry, J.-P., Maugis, C., and Michel, B. (2012). Slope heuristics: overview and implementation. *Statistics and Computing*, 22(2):455–470.
(Cited on page 45.)

Further references II

Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(7), 719–725.
(Cited on pages 36, 37, and 38.)

Birgé, L. and Massart, P. (2007). Minimal penalties for Gaussian model selection. *Probability Theory and Related Fields*, 138(1):33–73.
(Cited on pages 45, 59, 60, 61, 62, and 63.)

Chamroukhi, F. & Huynh, B.-T. (2019). Regularized maximum likelihood estimation and feature selection in mixtures-of-experts models. *Journal de la Société Française de Statistique*, 160(1), 57–85.
(Cited on page 10.)

Deleforge, A., Forbes, F., and Horaud, R. (2015). High-dimensional regression with gaussian mixtures and partially-latent response variables. *Statistics and Computing*, 25(5):893–911.
(Cited on pages 21, 22, 23, and 46.)

Devijver, E., Gallopin, M. (2018). Block-diagonal covariance selection for high-dimensional Gaussian graphical models. *Journal of the American Statistical Association*.
(Cited on pages 24, 25, and 26.)

Devijver, E., Gallopin, M., and Perthame, E. (2017). Nonlinear network-based quantitative trait prediction from transcriptomic data. *arXiv preprint arXiv:1701.07899*.
(Cited on pages 24, 25, and 26.)

Further references III

Ho, N., Yang, C.-Y., and Jordan, M. I. (2022). Convergence rates for gaussian mixtures of experts. *Journal of Machine Learning Research*.

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991). Adaptive mixtures of local experts. *Neural computation*, 3(1):79–87.

(Cited on page 10.)

Jordan, M. I. and Jacobs, R. A. (1994). Hierarchical mixtures of experts and the EM algorithm. *Neural computation*, 6(2):181–214.

Masoudnia, S. and Ebrahimpour, R. (2014). Mixture of experts: a literature survey. *Artificial Intelligence Review*, 42(2):275–293.

McLachlan, G. J. and Peel, D. (2000). *Finite Mixture Models*. John Wiley & Sons.

(Cited on page 10.)

Mendes, E. F. and Jiang, W. (2012). On convergence rates of mixtures of polynomial experts. *Neural computation*, 24(11):3025–3051.

Nguyen, H. D. and Chamroukhi, F. (2018). Practical and theoretical aspects of mixture-of-experts modeling: An overview. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(4):e1246.

Further references IV

Nguyen, H. D., Chamroukhi, F., and Forbes, F. (2019). Approximation results regarding the multiple-output Gaussian gated mixture of linear experts model. *Neurocomputing*, 366:208–214.

(Cited on page 10.)

(Cited on page 10.)

(Cited on page 10.)

(Cited on page 10.)

(Cited on page 10.)

(Cited on pages 10, 76, and 77.)

Norets, A. (2010). Approximation of conditional densities by smooth mixtures of regressions. *The Annals of statistics*, 38(3):1733–1766.

(Cited on page 10.)

Schwarz, G. et al. (1978). Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464.

(Not cited.)

White, H. (1982). Maximum Likelihood Estimation of Misspecified Models. *Econometrica*, 50(1):1–25.

(Cited on pages 36, 37, 38, 81, 82, 83, 84, 85, 86, and 87.)

Yuksel, S. E., Wilson, J. N., and Gader, P. D. (2012). Twenty Years of Mixture of Experts. *IEEE Transactions on Neural Networks and Learning Systems*, 23(8):1177–1193.

(Cited on page 10.)

Further references V

Boucheron, S. and Massart, P. (2011). A high-dimensional Wilks phenomenon. *Probability Theory and Related Fields*, 150(3):405–433.
(Cited on pages 85, 86, and 87.)

Cohen, S. and Le Pennec, E. (2011). Conditional density estimation by penalized likelihood model selection and applications. *Technical report, INRIA*.
(Cited on pages 74, 75, 81, 82, 83, and 84.)

Dalalyan, A. S. & Sebbar, M. (2018). Optimal Kullback–Leibler aggregation in mixture density estimation by maximum likelihood. *Mathematical Statistics and Learning*, 1(1), 1–35.
(Cited on pages 76 and 77.)

Fan, J., Zhang, C., and Zhang, J. (2001). Generalized Likelihood Ratio Statistics and Wilks Phenomenon. *The Annals of Statistics*, 29(1):153–193.
(Cited on pages 85, 86, and 87.)

Nguyen, T., Nguyen, H. D., Chamroukhi, F., & McLachlan, G. J. (2020c). An l_1-oracle inequality for the Lasso in mixture-of-experts regression models. *arXiv preprint arXiv:2009.10622*.
(Cited on page 99.)

Rigollet, P. (2012). Kullback–Leibler aggregation and misspecified generalized linear models. *The Annals of Statistics*, 40(2), 639–665.
(Cited on pages 76 and 77.)

Further references VI

Wilks, S. S. (1938). The Large-Sample Distribution of the Likelihood Ratio for Testing Composite Hypotheses. *The Annals of Mathematical Statistics*, 9(1):60–62.
(Cited on pages 85, 86, and 87.)