Model Selection and Approximation in High-dimensional Mixtures of Experts Models: From Theory to Practice

TrungTin Nguyen

29 October 2021

Oral Presentation at

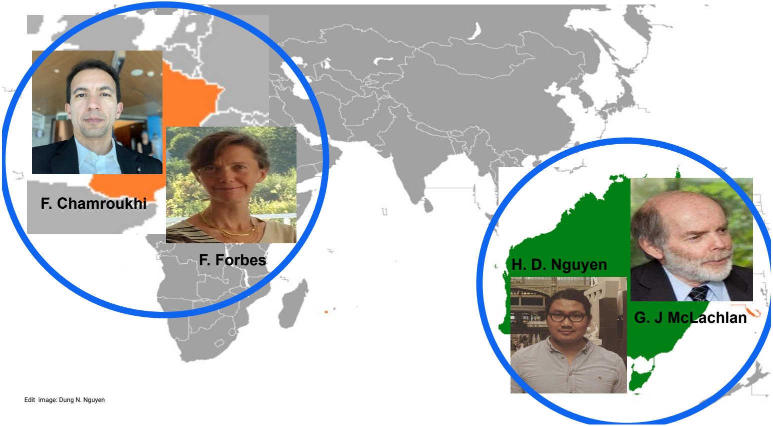
JED 2021 LE HAVRE

A mixture of French and Australian teams...

Edit image: Dung N. Nguyen

Joint work with

Faicel Chamroukhi (Université de Caen Normandie, France),
Hien Duy Nguyen (University of Queensland, Australia),
Geoffrey J McLachlan (University of Queensland, Australia),
Florence Forbes (Inria Grenoble-Rhône-Alpes, France).



Outline

Outline

1. Collection of mixture of experts models
2. Non-asymptotic oracle inequality
3. Numerical experiment
4. Future work on non-asymptotic oracle inequality

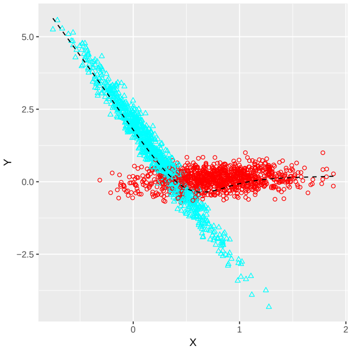
Outline

1. Collection of mixture of experts models
2. Non-asymptotic oracle inequality
3. Numerical experiment
4. Future work on non-asymptotic oracle inequality

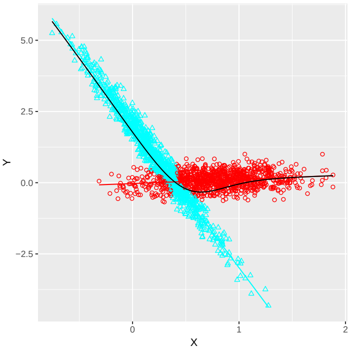
Nonlinear regression models for heterogeneous data

- We have: n random samples $(\mathrm{X}_i, \mathrm{Y}_i)_{i \in [n]} \in (\mathcal{X} \times \mathcal{Y})^n \subset \left(\mathbb{R}^D \times \mathbb{R}^L\right)^n$ with the corresponding observed values $(\mathrm{x}_i, \mathrm{y}_i)_{i \in [n]}$, $[n] := \{1, \ldots, n\}$, arising from an unknown conditional density s_0.
- Objective: learning **nonlinear regression models for heterogeneous data** between the multivariate response $\mathrm{Y} = (\mathrm{Y}_j)_{j \in [L]}$, and the set of covariates $\mathrm{X} = (\mathrm{X}_j)_{j \in [D]}$.
- Our proposal: approximating s_0 by a Gaussian-gated localized mixture of experts (GLoME) model due to its flexibility and effectiveness.
  - **Model selection problem**: estimating the number of mixture components via penalized maximum likelihood estimators.
  - **Non-asymptotic oracle inequality**: providing a lower bound on the penalty that ensures a weak oracle inequality is satisfied by our estimator.

Typical realization and regression clustering results



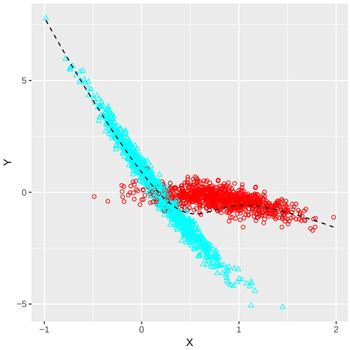
(a) Typical realization: WS case



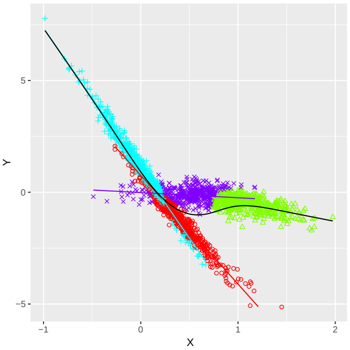
(b) Clustering by GLoME

Clustering deduced from the estimated conditional density of GLoME by a MAP principle with 2000 data points of 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) Clustering by GLoME

Clustering deduced from the estimated conditional density of GLoME by a MAP principle with 2000 data points of example MS. The dash and solid black curves present the true and estimated mean functions.

Flexibility and effectiveness of MoE models

GLoME is a mixture of experts (MoE) model [Jacobs et al., 1991, Xu et al., 1995, Nguyen and Chamroukhi, 2018]:

- Generalizing the classical finite mixtures and finite mixtures regression models [McLachlan and Peel, 2000].
- Containing a supervised Gaussian locally-linear mapping (GLLiM) model for high-dimensional regression data ($D \gg L$) [Deleforge et al., 2015].
- Approximation capabilities of MoE [Mendes and Jiang, 2012, Ho et al., 2019, Nguyen et al., 2021a].

GLoME models

Motivated by GLLiM models, an inverse regression framework, in GLoME models, Y becomes the covariates and X plays the role of a multivariate response.

Definition

$$s_{\psi_K}(\mathrm{x}|\mathrm{y}) = \sum_{k=1}^{K} \underbrace{\mathrm{g}_k(\mathrm{y};\boldsymbol{\omega})}_{\text{Gaussian gating function}} \underbrace{\Phi_D(\mathrm{x};\boldsymbol{\upsilon}_k(\mathrm{y}),\Sigma_k)}_{\text{Gaussian expert}},$$

- $K \in \mathbb{N}^{\star}$: number of mixture components,
- $\psi_K = (\boldsymbol{\omega}, \boldsymbol{\upsilon}, \Sigma) \in \Omega_K \times \Upsilon_K \times \mathrm{V}_K =: \Psi_K$: model parameter.

Gaussian gating functions

Definition

$$\mathrm{g}_k(\mathrm{y};\boldsymbol{\omega}) = \frac{\boldsymbol{\pi}_k \Phi_L(\mathrm{y};\mathrm{c}_k,\Gamma_k)}{\sum_{j=1}^K \boldsymbol{\pi}_j \Phi_L(\mathrm{y};\mathrm{c}_j,\Gamma_j)}, \text{ for every } k \in [K],$$

- $\boldsymbol{\omega} = (\boldsymbol{\pi}, \boldsymbol{c}, \Gamma) \in (\Pi_{K-1} \times \mathrm{C}_K \times \boldsymbol{V}'_K) =: \Omega_K$,
- $\Pi_{K-1} = \left\{ (\boldsymbol{\pi}_k)_{k\in[K]} \in (\mathbb{R}^+)^K, \sum_{k=1}^K \boldsymbol{\pi}_k = 1 \right\}$,
- C_K: K-tuples of mean vectors of size $L \times 1$,
- $\boldsymbol{V}'_K$: K-tuples of elements in $\mathcal{S}_L^{++}$,
- $\mathcal{S}_L^{++}$: collection of symmetric positive definite matrices on $\mathbb{R}^L$.

Boundedness conditions on the gating parameters

Assumptions

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

- $a_{\boldsymbol{\pi}}, A_{\boldsymbol{c}}, a_\Gamma, A_\Gamma$: *positive constants,*
- $m(\mathrm{A})$ *and* $M(\mathrm{A})$*: the modulus of the smallest and largest eigenvalues of a matrix* A*, respectively.*

Boundedness conditions on the Gaussian expert means

Assumptions ([Montuelle and Pennec, 2014])

- *Linear combination of bounded functions:* $d_\Upsilon \in \mathbb{N}^\star$, $T_\Upsilon \in \mathbb{R}^+$, $\boldsymbol{\upsilon} = (\boldsymbol{\upsilon}_k)_{k\in[K]} \in \Upsilon_K = \Upsilon_b^K$,

$$\Upsilon_b = \left\{ \mathrm{y} \mapsto \left(\sum_{i=1}^{d_\Upsilon} \boldsymbol{\alpha}_i^{(j)} \varphi_{\Upsilon,i}(\mathrm{y}) \right)_{j\in[D]} =: (\boldsymbol{\upsilon}_j(\mathrm{y}))_{j\in[D]} : \|\boldsymbol{\alpha}\|_\infty \leq T_\Upsilon \right\},$$

$(\varphi_{\Upsilon,i})_{i\in[d_\Upsilon]}$ *is a collection of bounded functions on* $\mathcal{Y}$.

- *Polynomial means:* $d'_\Upsilon \in \mathbb{N}^\star$, $\mathcal{Y} = [0,1]^L$, $\Upsilon_K = \Upsilon_p^K$,

$$\Upsilon_p = \left\{ \mathrm{y} \mapsto \left(\sum_{|\mathrm{r}|=0}^{d'_\Upsilon} \boldsymbol{\alpha}_{\mathrm{r}}^{(j)} \mathrm{y}^{\mathrm{r}} \right)_{j\in[D]} =: (\boldsymbol{\upsilon}_j(\mathrm{y}))_{j\in[D]} : \|\boldsymbol{\alpha}\|_\infty \leq T_\Upsilon \right\}.$$

Boundedness conditions on the Gaussian expert covariance matrices

Assumptions ([Celeux and Govaert, 1995])

$$V_K = \Big\{ \Sigma = (\Sigma_k)_{k \in [K]} = \left(B_k \mathrm{P}_k \mathrm{A}_k \mathrm{P}_k^\top \right)_{k \in [K]} : \forall k \in [K],$$
$$B_- \leq B_k \leq B_+,\ \mathrm{P}_k \in SO(D),\ \mathrm{A}_k \in \mathcal{A}(\lambda_-, \lambda_+) \Big\},$$

- $B_k = |\Sigma_k|^{1/D}$: *volume,* $B_- \in \mathbb{R}^+, B_+ \in \mathbb{R}^+$,
- P_k: *eigenvectors of* Σ_k *belongs to the special orthogonal* $SO(D)$,
- A_k: *diagonal matrix of normalized eigenvalues of* Σ_k, *such that* $|\mathrm{A}_k| = 1$ *and* $0 < \forall i \in [D], \lambda_- \leq (\mathrm{A}_k)_{i,i} \leq \lambda_+$.

Collection of GLoME models

Objective: estimating s_0 by conditional densities belonging to the collection of GLoME models $(S_m)_{m\in\mathcal{M}}$, defined by

Definition

$$S_m = \Big\{ \mathcal{X} \times \mathcal{Y} \ni (\mathrm{x}, \mathrm{y}) \mapsto s_{\psi_K}(\mathrm{x}|\mathrm{y}) =: s_m(\mathrm{x}|\mathrm{y}) : \\ \psi_K = (\boldsymbol{\omega}, \boldsymbol{\upsilon}, \Sigma) \in \widetilde{\Omega}_K \times \Upsilon_K \times \mathrm{V}_K =: \widetilde{\Psi}_K \Big\},$$

- $\mathcal{M} = \{K \in [K_{\max}], K_{\max} \in \mathbb{N}^\star\}$,
- $\dim(S_m) = \dim\left(\widetilde{\Psi}_K\right) = \dim\left(\widetilde{\Omega}_K\right) + \dim(\Upsilon_K) + \dim(\mathrm{V}_K)$.

Outline

Penalized maximum likelihood estimator (PMLE)

Definition

An η'-**PMLE** $\widehat{m}$ (corresponding **the best model** $S_{\widehat{m}}$ among $(S_m)_{m \in \mathcal{M}}$) is defined as follows

$$\widehat{m} = \operatorname*{argmin}_{m \in \mathcal{M}} \left(\sum_{i=1}^{n} - \ln \left(\widehat{s}_m \left(\mathrm{x}_i | \mathrm{y}_i \right) \right) + \mathrm{pen}(m) \right) + \eta' :$$

- $\widehat{s}_m$ is an η-**minimizer** of the negative log-likelihood (NLL) (infimum may not be unique or reached), defined by

$$\widehat{s}_m = \operatorname*{argmin}_{s_m \in S_m} \sum_{i=1}^{n} - \ln \left(s_m \left(\mathrm{x}_i | \mathrm{y}_i \right) \right) + \eta,$$

- $\mathrm{pen}(m)$: compensating variance and bias.

Losses

Definition

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

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

  if $sdy \ll tdy$ and $+\infty$ otherwise. Fixed predictors $\Rightarrow$ no $\mathbb{E}_{\mathrm{Y}}\left[\cdot\right]$.

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

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

Main result on an non-asymptotic oracle inequality

Theorem ([Nguyen et al., 2021b])

✍ **Assumptions**: *given a collection* $(S_m)_{m\in\mathcal{M}}$ *of GLoME models,* $\rho \in (0,1)$, $C_1 > 1$, $\Xi = \sum_{m\in\mathcal{M}} e^{-z_m} < \infty, z_m \in \mathbb{R}^+, \forall m \in \mathcal{M}$.

☞ **Conclusion**: *there exist constants* C *and* $\kappa(\rho, C_1) > 0$ *such that whenever for all* $m \in \mathcal{M}$,

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

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

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

Relationship: soft-max and Gaussian gating functions

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

$$\Gamma_S = \left\{ \gamma = \left((\mathrm{a}_k)_{k\in[K]}, (\mathrm{b}_k)_{k\in[K]} \right) \in \mathbb{R}^K \times \left(\mathbb{R}^L\right)^K \right\},$$

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

Lemma ([Nguyen et al., 2021a])

In general, $\mathcal{P}_S \subset \mathcal{P}_G$. If all Γ_k, $k \in [K]$, are identical, then $\mathcal{P}_G = \mathcal{P}_S$.

➠ Obtaining finite-sample oracle inequality for GLoME model is much more challenging compared to soft-max-gated mixture of experts (SGaME) model [Montuelle and Pennec, 2014].

Reparameterization the space of Gaussian gating functions

- Make use of the results for bracketing entropy of logistic weights from SGaME models [Montuelle and Pennec, 2014].

☞ Reparameterization trick:

$$
\mathrm{W}_K = \left\{ \mathrm{y} \mapsto \left(\ln\left(\boldsymbol{\pi}_k \Phi_L\left(\mathrm{y}; \mathrm{c}_k, \Gamma_k\right)\right)\right)_{k\in[K]} =: \mathrm{w}\left(\mathrm{y}; \boldsymbol{\omega}\right) : \boldsymbol{\omega} \in \widetilde{\Omega}_K \right\},
$$

$$
\mathcal{P}_K = \left\{ \mathrm{y} \mapsto \left(\frac{e^{\mathrm{w}_k(\mathrm{y})}}{\sum_{l=1}^{K} e^{\mathrm{w}_l(\mathrm{y})}} \right)_{k\in[K]} =: \left(g_{\mathrm{w},k}\left(\mathrm{y}\right)\right)_{k\in[K]}, \mathrm{w} \in \mathrm{W}_K \right\}.
$$

Asymptotic theory of a single parametric model

Misspecified case: $s_0 \notin S_m$, $\psi_m^\star = \mathrm{argmin}_{\psi_m \in \Psi_m} \mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, s_{\psi_m}\right)$,

$$S_m = \left\{ (\mathrm{x}, \mathrm{y}) \mapsto s_{\psi_m}(\mathrm{x}|\mathrm{y}) =: s_m(\mathrm{x}|\mathrm{y}) : \psi_m \in \Psi_m \subset \mathbb{R}^{\dim(S_m)} \right\}.$$

Theorem ([White, 1982, Cohen and Pennec, 2011])

✍ *Assumptions:* S_m *is identifiable, some strong regularity assumptions on* $\psi_m \mapsto s_{\psi_m}$, $\exists\ \mathrm{A}\left(\psi_m\right)$ *and* $\mathrm{B}\left(\psi_m\right)$*:*

$$[\mathrm{A}\left(\psi_m\right)]_{k,l} = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\int \frac{-\partial^2 \ln s_{\psi_m}}{\partial \psi_{m,k}\partial \psi_{m,l}}\left(\mathrm{x}|\mathrm{Y}_i\right) s_0\left(\mathrm{x}|\mathrm{Y}_i\right) d\lambda\right],$$

$$[\mathrm{B}\left(\psi_m\right)]_{k,l} = \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\int \frac{\partial \ln s_{\psi_m}}{\partial \psi_{m,k}}\left(\mathrm{x}|\mathrm{Y}_i\right) \frac{\partial \ln s_{\psi_m}}{\partial \psi_{m,l}}\left(\mathrm{x}|\mathrm{Y}_i\right) s_0\left(\mathrm{x}|\mathrm{Y}_i\right) d\lambda\right].$$

☞ *Conclusion:* $\mathbb{E}\left[\mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_m\right)\right]$ *is asymptotically equivalent to*

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

Asymptotic theory of a single parametric model

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

Theorem ([White, 1982, Cohen and Pennec, 2011])

- *It holds that*

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

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

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

Drawbacks of asymptotic theory

Well-specified case: $s_0 \in S_m$.

✷ Problem: **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].

Non-asymptotic upper bound of a single model

- **Initial target**:

$$\mathbb{E}\left[\mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_m\right)\right] \leq \left(\inf_{\psi_m \in \Psi_m} \mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, s_{\psi_m}\right) + \frac{1}{2n} \dim\left(S_m\right)\right) + C_2 \frac{1}{n}.$$

- **Our contribution**: weaker than expected!

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

  1. $\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_m\right) \leq \mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_m\right)$,
  2. $C_1 > 1$, κ is a constant that depends on C_1,
  3. Model complexity: $\mathfrak{D}_m \leftrightarrow \dim\left(S_m\right)$.

- **Existence of a corresponding lower bound for GLoME models: still an open question!** (Gaussian mixture models (GMM) [Maugis-Rabusseau and Michel, 2013]).

Remarks on our weak oracle inequality

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

Potential issues

1. Different divergences: $\mathrm{JKL}_{\rho}^{\otimes\mathrm{n}}\left(s_0, \widehat{s}_{\widehat{m}}\right) \leq \mathrm{KL}^{\otimes\mathrm{n}}\left(s_0, \widehat{s}_{\widehat{m}}\right)$.
2. $C_1 > 1$ and misspecified case: as $n \to \infty$, the error bound $\to C_1 \inf_{s_m\in S_m} \mathrm{KL}^{\otimes\mathrm{n}}\left(s_0, s_m\right)$ (potentially large!).
3. $\frac{\mathrm{pen}(m)}{n}$ is not directly related to the variance (asymptotic variance in the parametric case $\dim\left(S_m\right)/n$).

Solution for different divergences

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

- In general: $\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_{\widehat{m}}\right) \leq \mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_{\widehat{m}}\right)$.
- If $\sup_{m \in \mathcal{M}} \sup_{s_m \in S_m} \|s_0/s_m\|_{\infty} < \infty \Leftrightarrow \mathcal{Y}$ 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 Pennec, 2011] implies that

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

Solution for misspecified model

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

- ❄ $C_1 = 1$ with $\mathrm{KL}^{\otimes \mathrm{n}}$ loss: still an open question!
- ☞ Bias $\inf_{s_m \in S_m} \mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, s_m\right)$: small for $\mathcal{M}$ well-chosen via approximation capabilities of MoE and GMM models [Nguyen et al., 2019, Nguyen et al., 2020c, Nguyen et al., 2020b, Nguyen et al., 2021a].

Outline

Procedure for collection of supervised GLLiM models

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

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

✷ **Large enough but not explicit value** for κ!

- Asymptotic criteria: AIC ($\kappa = 1$) and BIC ($\kappa = \frac{\ln n}{2}$) [Akaike, 1974, Schwarz et al., 1978].
- Non-asymptotic criterion: our finite-sample oracle inequality, strong justification for slope heuristic approach (capushe package) in a finite sample setting [Birgé and Massart, 2007, Baudry et al., 2012].

Numerical schemes

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

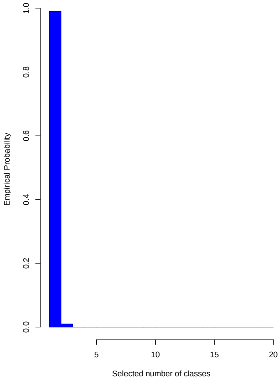
➠**Well-Specified (WS)**: $s_0^* \in S_m^*$,

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

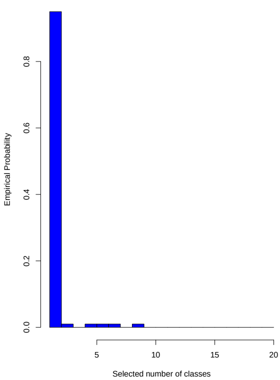
➠**Misspecified (MS)**: $s_0^* \notin S_m^*$,

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

Histogram of selected K using slope heuristic over 100 trials



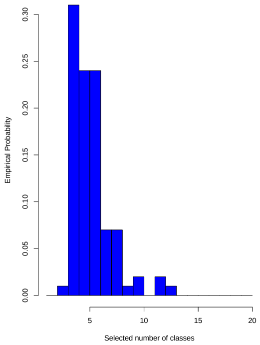
(a) 2000 data points



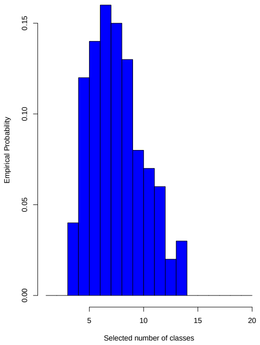
(b) 10000 data points

Comparison histograms of selected K in WS case using jump criterion over 100 trials between 2000 and 10000 data points.

Histogram of selected K using slope heuristic over 100 trials



(a) 2000 data points



(b) 10000 data points

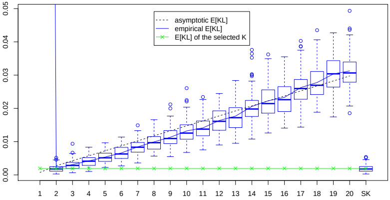
Comparison histograms of selected K in MS case using jump criterion over 100 trials between 2000 and 10000 data points.

Approximation error and variance term

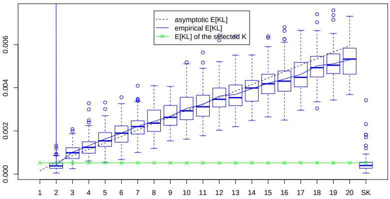
The bias-variance trade-off differs between the two examples:

- **WS case:** since the true density belongs to the model, the best choice is $K = 2$ even for large n.
- **MS case:** best 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.

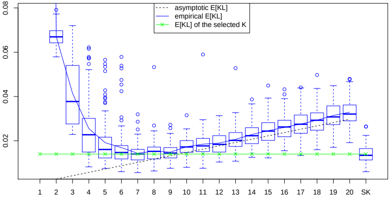
Box-plot of Kullback-Leibler divergence over 100 trials



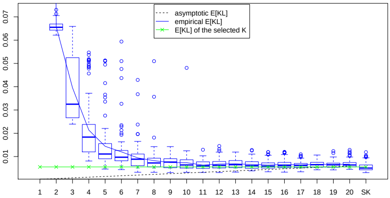
(a) WS with $n = 2000$



(b) WS with $n = 10000$



(c) MS with $n = 2000$



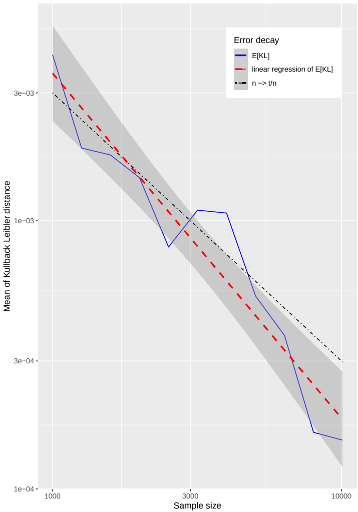
(d) MS with $n = 10000$

Empirical behavior of weak oracle inequality

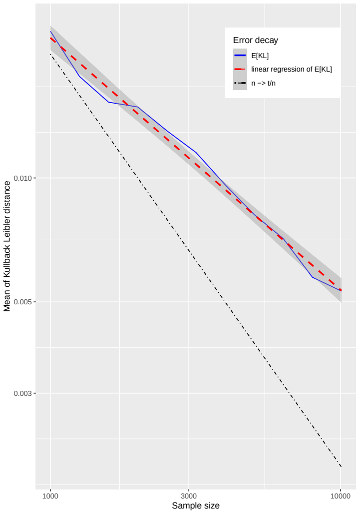
$$\mathbb{E}\left[\mathrm{JKL}_{\rho}^{\otimes \mathrm{n}}\left(s_0, \widehat{s}_{\widehat{m}}\right)\right] \leq C_1 \inf_{m \in \mathcal{M}} \left(\inf_{s_m \in S_m} \mathrm{KL}^{\otimes \mathrm{n}}\left(s_0, s_m\right) + \frac{\mathrm{pen}(m)}{n} \right) + \frac{\kappa\left(\rho, C_1\right) C_1 \Xi}{n}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

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

Rate of error decay in a log-log scale, using 30 trials



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



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

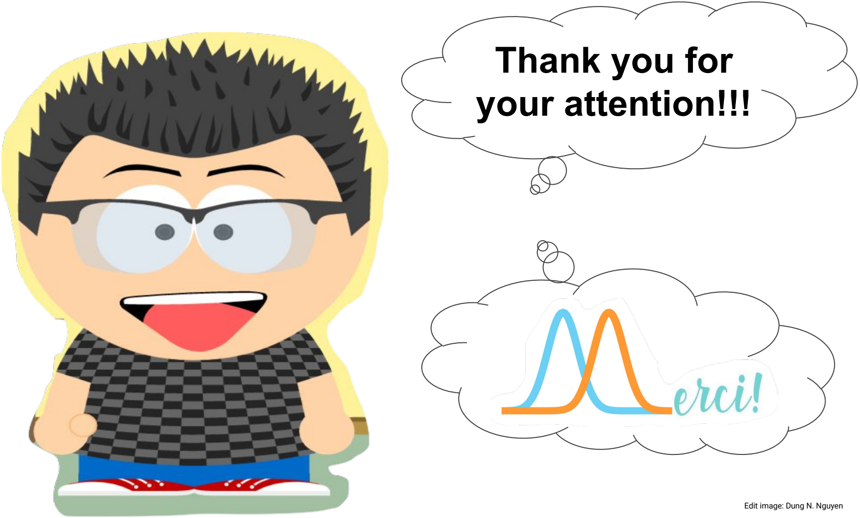
Outline

1. Collection of mixture of experts models
2. Non-asymptotic oracle inequality
3. Numerical experiment
4. Future work on non-asymptotic oracle inequality

❄ Non-asymptotic oracle inequalities for MoE models ❄

1. **Minimax analysis** (lower bound) for GLoME and SGaME models (GMM [Maugis-Rabusseau and Michel, 2013]).
2. **Improvement on upper bound**: $C_1 > 1 \to C_1 = 1$.

This is actually a realistic model selection problem...



Edit image: Dung N. Nguyen

Universal approximation theorem of MO-MoLE models

Multiple-output Gaussian gated mixture of linear experts (MO-MoLE) models and the class of MO-continuous functions, given $\mathcal{Y}$ is a compact set,

$$\mathcal{M}_K(\mathcal{Y}) = \left\{ \mathcal{Y} \ni \mathrm{y} \mapsto \sum_{k=1}^{K} \mathrm{g}_k(\mathrm{y};\boldsymbol{\omega}) \left[\mathrm{a}_k + \mathrm{B}_k^\top \mathrm{y}\right] \right\},$$

$$\mathcal{C}_L(\mathcal{Y}) = \left\{ \mathcal{Y} \ni \mathrm{y} \mapsto \mathrm{m}(\mathrm{y}) = (\mathrm{m}_j(\mathrm{y}))_{j\in[L]} : \mathrm{m}_j \in \mathcal{C}(\mathcal{Y}),\ j \in [L] \right\}.$$

Theorem ([Nguyen et al., 2019])

For all $\mathrm{m}^0 \in \mathcal{C}_L(\mathcal{Y})$, *there exists* $\{\mathrm{m}_K\}_{K\in\mathbb{N}^\star} \subset \bigcup_{K\in\mathbb{N}^\star} \mathcal{M}_K(\mathcal{Y})$,

$$\sum_{j=1}^{L} \sup_{\mathrm{y}\in\mathcal{Y}} \left|\mathrm{m}_j^0(\mathrm{y}) - \mathrm{m}_{K,j}(\mathrm{y})\right| \xrightarrow{K\to\infty} 0.$$

✸ $\mathcal{M}_K(\mathcal{Y}) \not\subseteq S_m$!

Universal approximation theorem of GMMs

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

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

Theorem ([Nguyen et al., 2020c, Nguyen et al., 2020b])

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

✸ $\mathcal{S}_K^{\varphi} \not\subseteq S_m$!

Essentially bounded and Lebesgue conditional PDF

Definition

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

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

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

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

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

Approximation class: isotropic SGaME and GLLiM

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

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

- Isotropic SGaME models:

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

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

- Isotropic GLLiM model ($\subset S_m$) when φ is standard MND:

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

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

Approximation capabilities: isotropic SGaME and GLLiM

Theorem ([Nguyen et al., 2021a])

(a) *Given* $\varphi \in \mathcal{F} \cap \mathcal{C}$, *for any target* $s_0 \in \mathcal{F}_p \cap \mathcal{C}_b^u$, *there exist sequences* $\left\{s_K^{\varphi}\right\}_{K \in \mathbb{N}^\star} \subset \mathcal{S}_S^{\varphi}$ *and* $\left\{s_K'^{\varphi}\right\}_{K \in \mathbb{N}^\star} \subset \mathcal{S}_G^{\varphi}$ *such that*

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

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

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

References I

Akaike, H. (1974).
A new look at the statistical model identification.
IEEE transactions on automatic control, 19(6):716–723.
(Cited on page 31.)

Baudry, J.-P., Maugis, C., and Michel, B. (2012).
Slope heuristics: overview and implementation.
Statistics and Computing, 22(2):455–470.
(Cited on page 31.)

Birgé, L. and Massart, P. (2007).
Minimal penalties for Gaussian model selection.
Probability theory and related fields, 138(1-2):33–73.
(Cited on page 31.)

References II

Boucheron, S. and Massart, P. (2011).
A high-dimensional Wilks phenomenon.
Probability Theory and Related Fields, 150(3):405–433.
(Cited on page 25.)

Celeux, G. and Govaert, G. (1995).
Gaussian parsimonious clustering models.
Pattern recognition, 28(5):781–793.
(Cited on page 15.)

Cohen, S. and Pennec, E. L. (2011).
Conditional density estimation by penalized likelihood model selection and applications.
Technical report, INRIA.
(Cited on pages 23, 24, and 28.)

References III

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 10 and 32.)

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 page 25.)

Ho, N., Yang, C.-Y., and Jordan, M. I. (2019).
Convergence Rates for Gaussian Mixtures of Experts.
arXiv preprint arXiv:1907.04377.
(Cited on page 10.)

References IV

Jacobs, R. A., Jordan, M. I., Nowlan, S. J., and Hinton, G. E. (1991).
Adaptive Mixtures of Local Experts.
Neural Computation, 3:79–87.
(Cited on page 10.)

Maugis-Rabusseau, C. and Michel, B. (2013).
Adaptive density estimation for clustering with Gaussian mixtures.
ESAIM: Probability and Statistics, 17:698–724.
(Cited on pages 26 and 40.)

McLachlan, G. J. and Peel, D. (2000).
Finite Mixture Models.
John Wiley & Sons.
(Cited on page 10.)

References V

Mendes, E. F. and Jiang, W. (2012).
On Convergence Rates of Mixtures of Polynomial Experts.
Neural Computation, 24(11):3025–3051.
(Cited on page 10.)

Montuelle, L. and Pennec, E. L. (2014).
Mixture of Gaussian regressions model with logistic weights, a penalized maximum likelihood approach.
Electronic Journal of Statistics, 8(1):1661–1695.
(Cited on pages 14, 21, and 22.)

Nguyen, H. D. and Chamroukhi, F. (2018).
Practical and theoretical aspects of mixture-of-experts modeling: An overview.
WIREs Data Mining and Knowledge Discovery, 8(4):e1246.
(Cited on page 10.)

References VI

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 pages 29 and 42.)

Nguyen, H. D., Nguyen, T., Chamroukhi, F., and McLachlan, G. (2020a).
Approximations of conditional probability density functions in Lebesgue spaces via mixture of experts models.
arXiv preprint arXiv:2012.02385.
(Not cited.)

References VII

Nguyen, H. D., Nguyen, T., Chamroukhi, F., and McLachlan, G. J. (2021a).
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, 21, 29, and 46.)

Nguyen, T., Chamroukhi, F., Nguyen, H. D., and McLachlan, G. J. (2020b).
Approximation of probability density functions via location-scale finite mixtures in Lebesgue spaces.
arXiv preprint arXiv:2008.09787.
(Cited on pages 29 and 43.)

References VIII

Nguyen, T. T., Nguyen, H. D., Chamroukhi, F., and Forbes, F. (2021b).
A non-asymptotic penalization criterion for model selection in mixture of experts models.
arXiv preprint arXiv:2104.02640.
(Cited on page 20.)

Nguyen, T. T., Nguyen, H. D., Chamroukhi, F., and McLachlan, G. J. (2020c).
Approximation by finite mixtures of continuous density functions that vanish at infinity.
Cogent Mathematics & Statistics, 7(1):1750861.
(Cited on pages 29 and 43.)

References IX

Schwarz, G. et al. (1978).
Estimating the dimension of a model.
The annals of statistics, 6(2):461–464.
(Cited on page 31.)

White, H. (1982).
Maximum Likelihood Estimation of Misspecified Models.
Econometrica, 50(1):1–25.
(Cited on pages 23 and 24.)

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 page 25.)

References X

Xu, L., Jordan, M., and Hinton, G. E. (1995).
An Alternative Model for Mixtures of Experts.
In Tesauro, G., Touretzky, D., and Leen, T., editors, *Advances in Neural Information Processing Systems*, volume 7. MIT Press.
(Cited on page 10.)