Dendrograms of Mixing Measures for Softmax-Gated Gaussian Mixture of Experts: Consistency without Model Sweeps

TienHai Do[1,2], Trung Nguyen Mai[2,3], TrungTin Nguyen[4,5], Nhat Ho[6], Binh T. Nguyen[1,2], Christopher Drovandi[4,5]

[1]Faculty of Mathematics and Computer Science, University of Science, Ho Chi Minh City, Vietnam.
[2]Vietnam National University Ho Chi Minh City, Vietnam.
[3]Faculty of Information Technology, University of Science, Ho Chi Minh City, Vietnam.
[4]ARC Centre of Excellence for the Mathematical Analysis of Cellular Systems.
[5]School of Mathematical Sciences, Queensland University of Technology, Australia.
[6]Department of Statistics and Data Science, University of Texas at Austin, Austin, USA.

AISTATS 2026

Problem Setting: Softmax-Gated MoE (SGMoE)

- **Model:** $p_{G_0}(y|\boldsymbol{x}) = \sum_{k=1}^{K_0} \frac{\exp(\boldsymbol{\omega}_{1k}^{0\top}\boldsymbol{x}+\omega_{0k}^0)}{\sum_{j=1}^{K_0}\exp(\boldsymbol{\omega}_{1j}^{0\top}\boldsymbol{x}+\omega_{0k}^0)} \times \mathcal{N}(y|\boldsymbol{a}_k^{0\top}\boldsymbol{x} + b_k^0, \sigma_k^0)$
- **Mixing measure:** $G_0 := \sum_{k=1}^{K_0} \exp(\omega_{0k}^0)\,\delta_{\boldsymbol{\theta}_k^0}$ where $\boldsymbol{\theta}_k^0 = (\boldsymbol{\omega}_{1k}^0, \boldsymbol{a}_k^0, b_k^0, \sigma_k^0)$.
- **Motivation:** Classical criteria such as AIC and BIC typically require fitting models for all candidate values $K = 1, 2, \ldots, K_{\max}$ in order to select K_0. To avoid this exhaustive sweep, we propose a sweep-free dendrogram selection criterion (DSC) based on dendrograms. Our approach reveals the hierarchical structure of the model while recovering the fast $N^{-1/2}$ convergence rate for parameter estimation.

Dissimilarity and Aggregation Formulas

Dissimilarity Metric: We define a dissimilarity metric d between two atoms.

$$\mathrm{d} = \frac{\exp(\omega_{0\ell_1} + \omega_{0\ell_2})}{\exp(\omega_{0\ell_1}) + \exp(\omega_{0\ell_2})}\left(\|(\boldsymbol{\omega}_{1\ell_1}, b_{\ell_1}) - (\boldsymbol{\omega}_{1\ell_2}, b_{\ell_2})\|^2 + \|(\boldsymbol{a}_{\ell_1}, \sigma_{\ell_1}) - (\boldsymbol{a}_{\ell_2}, \sigma_{\ell_2})\|\right)$$

Softmax-Weighted Merging: Pick pair $(i, j) = \arg\min_{\ell_1 \neq \ell_2 \in [K]} \mathrm{d}(\cdot, \cdot)$ and replace with expert $(*)$:

- $\omega_{0*} = \log(\exp \omega_{0i} + \exp \omega_{0j})$, $\boldsymbol{\omega}_{1*} = \sum_{k \in \{i,j\}} \exp(\omega_{0k} - \omega_{0*})\boldsymbol{\omega}_{1k}$
- $b_* = \sum_{k \in \{i,j\}} \exp(\omega_{0k} - \omega_{0*}) b_k$
- $\boldsymbol{a}_* = \sum_{k \in \{i,j\}} \frac{\exp(\omega_{0k})}{\exp(\omega_{0*})}[(\boldsymbol{\omega}_{1k} - \boldsymbol{\omega}_{1*})(b_k - b_*) + \boldsymbol{a}_k]$
- $\sigma_* = \sum_{k \in \{i,j\}} \frac{\exp(\omega_{0k})}{\exp(\omega_{0*})}[(b_k - b_*)^2 + \sigma_k]$

Aggregating Procedure

Alg 1: SGMoE Merge Step

Input: $G^{(\kappa)} = \sum_{k=1}^{\kappa} \exp(\omega_{0k})\delta_{\boldsymbol{\theta}_k}$

1. **Find** $(i,j) = \arg\min_{\ell_1 \neq \ell_2} \mathcal{D}_{clus}(\delta_{\boldsymbol{\theta}_{\ell_1}}, \delta_{\boldsymbol{\theta}_{\ell_2}})$
2. **Aggregate** parameters $\boldsymbol{\theta}_* = (\omega_{0*}, \boldsymbol{\omega}_{1*}, a_*, b_*, \sigma_*)$
3. **Update** mixing measure:

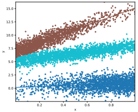
$G^{(\kappa-1)} \leftarrow \exp(\omega_{0*})\delta_{\boldsymbol{\theta}_*} + \sum_{k \neq i,j} \exp(\omega_{0k})\delta_{\boldsymbol{\theta}_k}$

Return $G^{(\kappa-1)}$

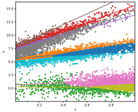
Alg 2: Aggregation Path

Input: Over-fitted MLE $G^{(K)} = \sum_{k=1}^{K} \exp(\omega_{0k})\delta_{\boldsymbol{\theta}_k}$

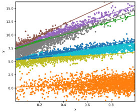
1. **Initialize** Tree $\mathcal{T} = (\mathcal{V}, \mathcal{E}, \mathcal{H})$ with atoms of $G^{(K)}$
2. **For** $\kappa = K, \ldots, 2$ **do:**
   - Apply $G^{(\kappa-1)} \leftarrow$ **Alg 1**$(G^{(\kappa)})$
   - Update tree nodes $\mathcal{V}$ and links $\mathcal{E}$
   - Record merge height $h^{(\kappa)} \in \mathcal{H}$
3. **Return** Path $\{G^{(\kappa)}\}_{\kappa=1}^{K}$ and Tree $\mathcal{T}$



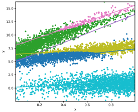
(a) True Regression



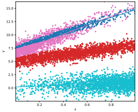
(b) $K = 10$



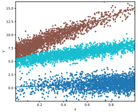
(c) $K = 8$



(d) $K = 6$



(e) $K = 4$



(f) $K_0 = 3$

Dendrogram Selection Criterion (DSC)

- For each level κ in the aggregation path $\{\widehat{G}_N^{(\kappa)}\}_{\kappa=2}^{K}$, define:

$$\mathrm{DSC}_N^{(\kappa)} := -\left(\mathrm{h}_N^{(\kappa)} + \epsilon_N\,\bar{\ell}_N\big(p_{\widehat{G}_N^{(\kappa)}}\big)\right)$$

- **Terms:**
  - $\mathrm{h}_N^{(\kappa)}$: Merging height at level κ.
  - $\bar{\ell}_N$: Average log-likelihood of the aggregated model.
  - ϵ_N: Tuning parameter satisfying $1 \ll \epsilon_N \ll (N/\log N)^{1/(2\bar{r})}$.
- **Decision Rule:** $\widehat{K}_N = \mathrm{argmin}_{\kappa\in[2,K]}\mathrm{DSC}_N^{(\kappa)}$.
- This rule favors levels where atoms are well-separated (heights) and fit the data well (likelihood).

Theoretical Results

Theorem 1 (Fast convergence rates). There exist universal constants $C_1', c_1, C_2', c_2 > 0$ such that for all $\kappa \in [K_0 + 1, K]$ and $\kappa' \in [K_0]$, we have

$$\mathbb{P}\big(\mathrm{D}_{\mathrm{FRA}}(\widehat{G}_N^{(\kappa)}, G_0) > C_2'(\log N/N)^{1/2}\big) \lesssim e^{-c_2 \log N},$$
$$\mathbb{P}\big(\mathrm{D}_{\mathrm{E}}(\widehat{G}_N^{(\kappa')}, G_0^{(\kappa')}) > C_1'(\log N/N)^{1/2}\big) \lesssim e^{-c_1 \log N}.$$

Theorem 2 (Height control). For all $\kappa \in [K_0 + 1, K]$ and $\kappa' \in [K_0]$,

$$\mathrm{h}_N^{(\kappa)} \lesssim (\log N/N)^{1/\bar{r}(\widehat{G}_N)}, \ \left|\mathrm{h}_N^{(\kappa')} - \mathrm{h}_0^{(\kappa')}\right| \lesssim (\log N/N)^{1/2},$$

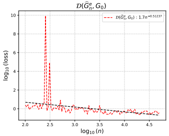
with constants depending only on G_0, $\boldsymbol{\theta}$, and κ.

Theorem 3 (Likelihood concentration on the path). Under regularity conditions, for any $\kappa \in [K_0 + 1, K]$, $\left|\bar{\ell}_N(p_{\widehat{G}_N^{(\kappa)}}) - \mathcal{L}(p_{G_0})\right| \lesssim (\log N/N)^{1/(2\bar{r}(\widehat{G}_N))}$. Moreover, for $\kappa' \in [K_0]$, $\bar{\ell}_N(p_{\widehat{G}_N^{(\kappa')}}) \to \mathcal{L}(p_{G_0^{(\kappa')}})$ in $\mathbb{P}_{G_0}$-probability as $N \to \infty$.

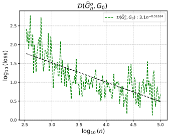
Theorem 4 (Consistency of model selection). Under regularity conditions, if $K_0 \geq 2$, then $\widehat{K}_N \to K_0$ in $\mathbb{P}_{G_0}$-probability as $N \to \infty$.

Experiments

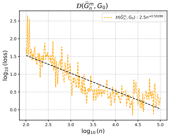
Exp 1: Restoring Parametric Convergence Rates



(a) Exact $\hat{G}_n^e$

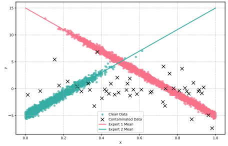


(b) Over-fitted $\hat{G}_n^o$

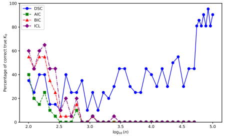


(c) Merged $\hat{G}_n^m$

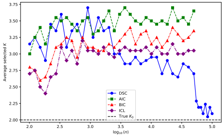
Exp 2: Robustness to Laplace Noise Contamination



(a) Contaminated sample



(b) Correct Selections (%)

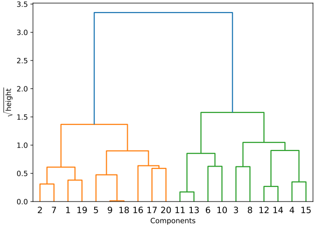


(c) Avg. Components

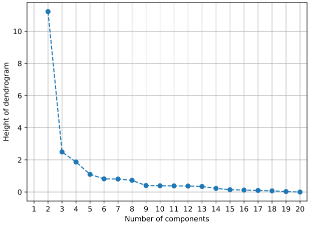
DSC demonstrates superior robustness compared to AIC, BIC, and ICL.

Real Data Illustration

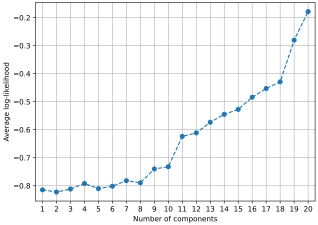
Maize drought-responsive traits dataset: DSC identifies meaningful hierarchical structure without needing to sweep through K.



(a) Dendrogram



(b) Level Heights



(c) Avg. Log-likelihood

Summary

Table: Summary of convergence rates among different settings.

Setting	**Loss**	$p_{G_0}(y \mid \mathrm{x})$	$\exp(\omega_{0k}^0)$	$\boldsymbol{\omega}_{1k}^0,\ b_k^0$	$\mathrm{a}_k^0,\ \sigma_k^0$
Exact-fit	$\mathcal{V}_{DE}$	$\mathcal{O}((\log N/N)^{1/2})$	$\mathcal{O}((\log N/N)^{1/2})$	$\mathcal{O}((\log N/N)^{1/2})$	$\mathcal{O}((\log N/N)^{1/2})$
Over-fit	$\mathcal{V}_{DO}$	$\mathcal{O}((\log N/N)^{1/2})$	$\mathcal{O}((\log N/N)^{1/2})$	$\mathcal{O}((\log N/N)^{1/2\bar{r}(\|\mathcal{A}_k\|)})$	$\mathcal{O}((\log N/N)^{1/\bar{r}(\|\mathcal{A}_k\|)})$
Merged	$\mathcal{V}_{DFRA}$	$\mathcal{O}((\log N/N)^{1/2})$	$\mathcal{O}((\log N/N)^{1/2})$	$\mathcal{O}((\log N/N)^{1/2})$	$\mathcal{O}((\log N/N)^{1/2})$

Thank You!