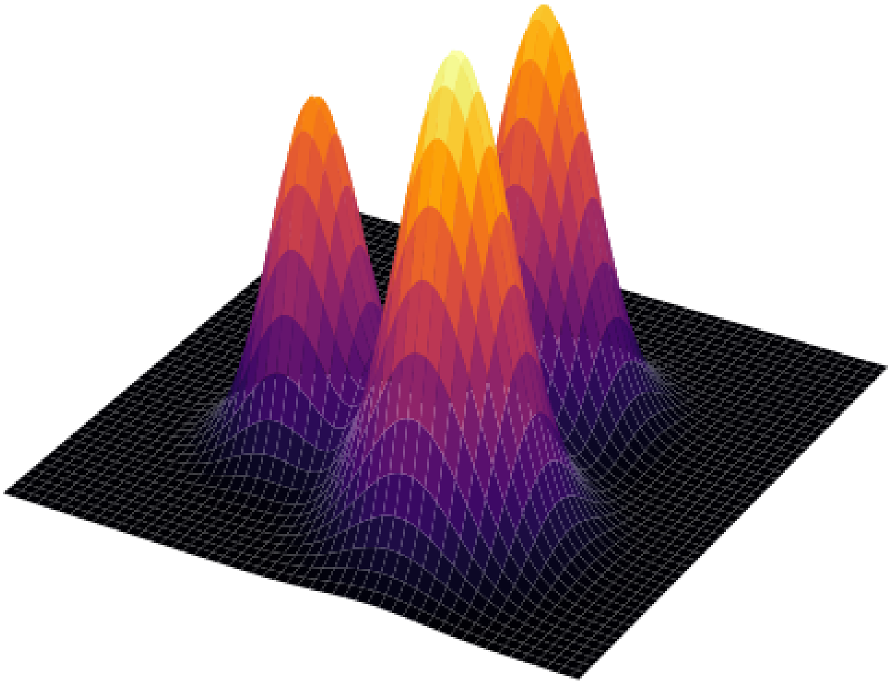
Reproduction and Extension to t-Student distributions of the Article ”Optimal Bayesian Estimation of Gaussian Mixtures with a Growing Number of Components” by Ohn and Lin

Pierre Clayton - Thomas Garnier - Vincent Gimenes

November 2024



Bayesian Statistics - Final Project

Introduction

Finite mixture models, such as Gaussian mixture models, have become fundamental tools in statistical modeling and machine learning, as demonstrated in the study ”Optimal Bayesian Estimation of Gaussian Mixtures with a Growing Number of Components” by Ohn and Lin. Gaussian mixtures offer several advantages, including zero kurtosis, perfect symmetry, linear stability, and the existence of all moments, making them robust frameworks for modeling heterogeneous data. However, these models can be overly restrictive when applied to datasets with heavy tails or significant outliers.

The t-Student distribution provides a compelling alternative due to its heavier tails, enabling it to better accommodate such data. Extending the results of Ohn and Lin to t-Student mixtures enhances the flexibility of finite mixture models for real-world datasets that deviate from Gaussian assumptions. This is particularly valuable in fields like finance, biology, and physics, where heavy-tailed distributions frequently arise. Furthermore, the t-Student distribution is closely related to the Gaussian distribution, converging to a normal distribution as its degrees of freedom approach infinity. This property ensures that the t-Student distribution maintains similarity to the Gaussian while offering additional robustness.

This work explores the theoretical and practical implications of replacing Gaussian components with t-Student components. Specifically, we aim to evaluate whether the Bayesian framework and posterior consistency established for Gaussian mixtures remain valid under this generalization. Beyond its practical applications (see 4 and 5 in the References) , this exploration is also driven by a genuine curiosity to understand how the theoretical properties of Gaussian mixtures translate when extended to t-Student distributions and to uncover any surprising nuances along the way.

Definition for the t-Student Mixture Model

A **finite t-Student mixture model** represents a probability distribution as a convex combination of a finite number of t-Student distributions (you will find a reminder of the t-Student Distribution in Appendix A). This model generalizes the Gaussian mixture model by replacing the normal distribution components with t-Student distributions, which have heavier tails. Formally, a t-Student mixture model can be expressed as:

$$p_{\nu * \mathcal{T}_\lambda}(x) = \int_{\theta \in \Theta} t_\lambda(x - \theta)\, \nu(d\theta),$$

where: $x \in \mathbb{R}$ is the observed data, $\Theta \subseteq \mathbb{R}$ is the parameter space for the component means, $t_\lambda(x - \theta)$ is the probability density function (pdf) of the t-Student distribution with mean θ and degrees of freedom λ, ν is the **mixing distribution**, a probability measure on Θ, $\nu * \mathcal{T}_\lambda$ denotes the convolution of ν with the t-Student distribution.

When ν is a discrete measure with finite support, the mixture model becomes finite:

$$\nu = \sum_{j=1}^{k} w_j \delta_{\theta_j},$$

so that the mixture density simplifies to:

$$p_{\nu * \mathcal{T}_\lambda}(x) = \sum_{j=1}^{k} w_j t_\lambda(x - \theta_j),$$

where $k \in \mathbb{N}$ is the number of components; $w = (w_1, \dots, w_k)$ are the **mixing weights**, satisfying $w_j \geq 0$ and $\sum_{j=1}^{k} w_j = 1$; $\theta = (\theta_1, \dots, \theta_k) \in \Theta^k$ are the **component means**; δ_{θ_j} is the Dirac delta measure at θ_j.

Assume we observe $X_1, X_2, \dots, X_n$ independent and identically distributed (i.i.d.) samples from the mixture distribution $p_{\nu * \mathcal{T}_\lambda}$.

Theorical implications

Our report must be short. Thus, we will summarize in a nutshell the main ideas of the article. We strongly recommend the reader to have a look at the article before reading the following parts.

Context

The article by Ohn and Lin focuses on Bayesian estimation for finite Gaussian mixture models, particularly in scenarios where the number of components grows with the sample size. This approach addresses the need to model increasing heterogeneity in data as sample sizes expand. The authors propose a novel framework that introduces a sample size dependent prior, designed to handle challenges posed by the increasing number of components while ensuring optimal posterior contraction rates for the mixing distribution under the Wasserstein distance. They establish posterior consistency for the number of components, demonstrating that this consistency holds even when the true number of components diverges with the sample size.

The authors also show that separation conditions for the mixture components lead to both adaptive and minimax-optimal convergence rates of the posterior distribution. Additionally, their framework is extended to accommodate scenarios where the number of components grows rapidly with the sample size. Simulation studies and real-world applications validate their theoretical findings, demonstrating the practical utility of their Bayesian approach for Gaussian mixture models.

Now, we are going to investigate whether, theoretically, the results would work for a t-Student mixture model and, if not, why.

The tail is too fat

At the section 2.6 from the article, the authors proposed an extension to their main results to general mixture models. Nevertheless, such models must verify the assumption (F) which is:

The family of distribution functions $\{F(\cdot, \theta) : \theta \in \Theta\}$ on $\mathbb{R}$ satisfies the following conditions:

(F1) Θ is a compact subset of $\mathbb{R}$ with nonempty interior.

(F2) There exists a constant $c_1 > 0$ such that:

$$\|f(\cdot, \theta_1) - f(\cdot, \theta_2)\|_\infty \leq c_1 |\theta_1 - \theta_2|,$$

for any $\theta_1, \theta_2 \in \Theta$. Moreover, there are constants $c_2 > 0$ and $r \in (0, 1]$ such that:

$$\int p_{\nu_1 * F}(x) \left(\frac{p_{\nu_1 * F}(x)}{p_{\nu_2 * F}(x)} \right)^r \lambda(dx) \leq c_2,$$

for any $\nu_1, \nu_2 \in \mathcal{M}(\Theta)$.

(F3) For any $k \in \mathbb{N}$, there exists an estimator $\hat{M}_k$ of the moment $m_k(\nu)$, based on the sample $X_1, \ldots, X_n \overset{iid}{\sim} P_{\nu * F}$, such that:

$$\mathbb{P}^{(n)}_{\nu * F} \hat{M}_k = m_k(\nu),$$

and:

$$\mathbb{P}^{(n)}_{\nu * F} \left(\hat{M}_k - m_k(\nu) \right)^2 \lesssim \frac{1}{n} (c_3 + \sqrt{k})^{2k},$$

for any $\nu \in \mathcal{M}(\Theta)$, and for some constant $c_3 > 0$.

The (F3) is quite strong and our t-Student model thereby does not check this assumption, since the number of finite moment for such a model is finite and fixed to the number of freedom minus one. By checking the proof of the theorem 2.2 for example, we can see that the assumption (F3) allows to state the Lemma A.2 (in appendix A) which establishes an exponential tail bound on the estimators of moments. Therefore, we cannot prove that the main theorems work for our t-student mixture model espacially due to its heavy tail.

Nevertheless, since t-student distribution tends to Gaussian distribution when the degree of freedom tends to infinity, the assumptions and therefore the theorems should be checked asymptotically with the degree of freedom. It could be very interesting to make the same simulations as the ones for Gaussian mixture models to see how a t-student mixture model behaves when the degree of freedom increases.

The Appendix B

In the article, still at the part "2.6 Extension to general mixture model", at the remark 3, we inquired that like us, a reader was dispointed that the assumption (F3) was too strong for some models like a Cauchy model:

"**Remark 3**: As a reviewer pointed out, Assumption (F3) is somewhat strong and a number of mixture models do not satisfy it. For example, although the Cauchy location mixture model with $f(x;\theta) = \frac{1}{\pi(1+(x-\theta)^2)}$ is strongly identifiable (by [2][Theorem 3]) and can be analyzed under a different theoretical framework given in **Appendix B**, it does not satisfy Assumption (F3) since the Cauchy distribution does not have finite moments of order greater than or equal to 1."

As you should know, a Cauchy distribution is actually a t-student distribution where the degree of freedom is equal to 1.

Unlike the main section, where the number of components k grows with the sample size, the framework in appendix B of the article assumes a fixed number of components. Let's see if our model is well suited for this framework. Here is an assumption to be hold:

Assumption $\mathbf{F}^*(q)$

The family $\{F(\cdot;\theta) : \theta \in \Theta\}$ of distribution functions on $\mathbb{R}$ satisfies Assumptions (F1) and (F2) as well as the following conditions:

- **($\mathbf{F}^*$1)** For any $x \in \mathbb{R}$, $F(x;\theta)$ is q-differentiable with respect to θ.
- **($\mathbf{F}^*$2)** $\{F(\cdot;\theta) : \theta \in \Theta\}$ is q-strongly identifiable (see definition in Annexe).

We can show that the t-Student distribution model check all of these assumptions. The computation details are in Appendix D. Accordingly, the t-Student model verifies the theorem that shows the posterior contraction rate for the strongly identifiable mixtures under this setup:

Theorem B.1

Assume that the family $\{F(\cdot,\theta) : \theta \in \Theta\}$ of distribution functions on $\mathbb{R}$ satisfies Assumption $F^*(q)$ with $q = 2k^*$. Then with the prior distribution Π satisfying Assumption P

(see Appendix C for details), we have:

$$\sup_{\nu^* \in \mathcal{M}_{k^*}(\Theta)} P^{(n)}_{\nu^* * F} \left[\Pi_F \left(W_1(\nu, \nu^*) \geq M \left(\frac{\log n}{n} \right)^{\frac{1}{4k^*-2}} \middle| X_{1:n} \right) \right] = o(1),$$

for some constant $M > 0$ depending only on k^*.

Another interesting remark and discussion

After exposed the theorem B.1, the authors wrote an interesting remark:

"**Remark 6:** As we mentioned before, although the Gaussian mixture model considered in Section 2 satisfies Assumption F with $q = \infty$, we cannot immediately derive Theorem 2.2 from Theorem B.1 since the latter theorem assumes a fixed number of components. We believe, however, that even if the number of components grows, the result of Theorem B.1 still holds with the same convergence rate as (B.1), up to a constant depending on k^*, provided that Assumption F is met with $q = \infty$. We need to establish a uniform version of Theorem 6.3 of over the number of components, which is a key technical tool for the proof. This could be an objective of future work."

If we have well understood this remark, the authors would believe that a model q-strongly identifiable with $q = \infty$ (which is actually the case for the t-student model), the theorems exposed in the framework where there are a growing number of components would hold, with the same convergence rate ! This mean that the t-student mixture model finally could work as well as the Gaussian mixture model does. That would be a wonderful result and we tried to see whether or not this result seems to hold empirically.

Simulation Study

We adopt the same methodology as in the original paper, applying it to a mixing Student's tdistribution rather than a Gaussian mixture model. The results are presented in Figure 1, and they revealed that the variability in the performance of the different methods is particularly influenced by two main factors: the number of parameters and the degrees of freedom of the Student's t-distribution. In configurations where the atoms were widely separated (case 4), we found that the Bayesian methods provided more consistent results, even though their performance was slightly worse compared to the other configurations (cases 1, 2, and 3). This phenomenon could be due to the increased difficulty in capturing the data when the atoms are farther apart, which makes model fitting more complex. In scenarios with low degrees of freedom (= 5, 20), we observed that the Wasserstein distance between the predicted distributions and the true data distributions was significantly larger for the DMM method compared to the Bayesian and MAP approaches. This finding reinforces the idea that DMM does not handle the increased variability typical of low degrees of freedom distributions, a characteristic inherent to Student's t-distributions. It confirms that DMM is particularly sensitive to non-Gaussian distributions and the presence of heavy tails in the data. On the other hand, the MAP and Bayesian methods performed similarly, although we noted greater variability in performance for lower degrees of freedom and in case 4, where the components were more separated. This variability could be due to the flexibility of Bayesian models, which, while adaptive, may be sensitive to the distribution of the data, especially when it involves complex structures, such as widely spaced atoms. Moreover, for higher degrees of freedom (= 100, 1000), the results from the Bayesian methods increasingly resembled those obtained with normal distributions, as expected, since a Student's tdistribution with a large number of degrees of freedom converges towards a normal distribution. Scenario 4, revealed suboptimal performance for all methods. This observation highlights the challenges encountered when the data is dispersed and the number of components becomes too high to be effectively modeled, especially by non-Bayesian methods. We also observe that the Student's t-distribution results approaches the ones in the original paper, with a normal distribution, as the degrees of freedom grow. However, in scenarios with low degrees of freedom, where Student's t-distributions are still quite different from normal ones, the Bayesian method appears to be better equipped to handle this complexity, unlike DMM, which loses performance as the distribution becomes heavier-tailed.

Conclusion

In conclusion, this work extends the methodology presented in Optimal Bayesian Estimation of Gaussian Mixtures with Growing Number of Components by replacing Gaussian components with a mixture of Student's t-distributions, which possess the advantage of heavy tails. While we were able to replicate the Bayesian framework and its posterior consistency for Gaussian mixtures, the heavy tails of the Student's t-distribution introduce additional complexities that prevent us from fully proving the validity of the main theorems, particularly due to the violation of condition F3. Despite this theoretical limitation, our empirical analysis provides valuable insights into the performance of the Bayesian approach when applied to Student's t-distributions. The results highlight that the performance variability of different methods is primarily influenced by the number of components and the degrees of freedom of the distribution. Bayesian methods performed consistently, particularly in configurations with widely separated components, though with slightly reduced performance compared to other setups. DMM, on the other hand, struggled with the increased variability of low degrees of freedom, reinforcing its sensitivity to heavy-tailed distributions

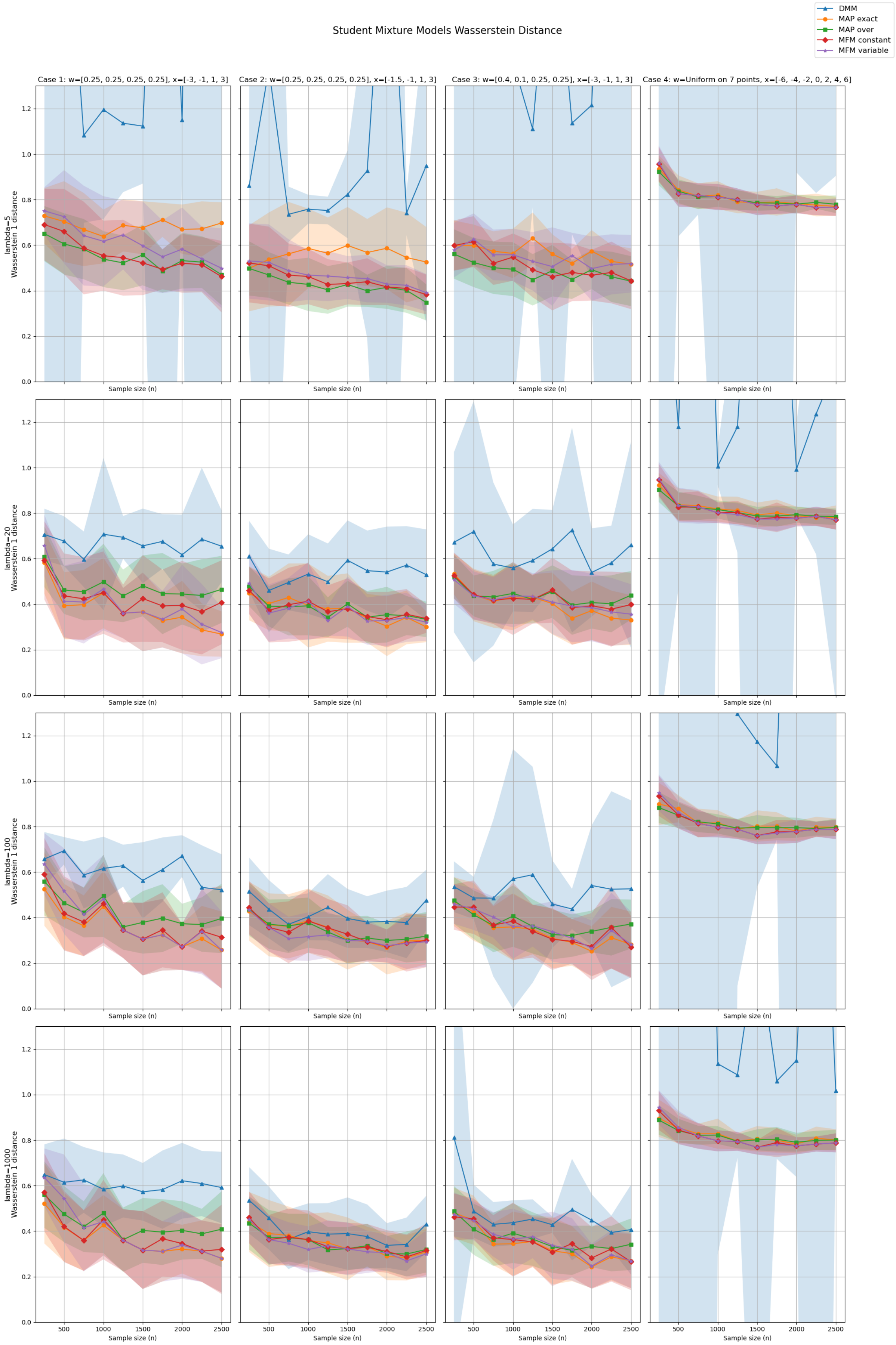


Figure 1: Simulation result for Student t-distribution

References

[1] Ohn, I. Lin, L. *Optimal Bayesian Estimation of Gaussian Mixtures with Growing Number of Components.*

[2] Chen, Jiahua. *Optimal Rate of Convergence for Finite Mixture Models.*

[3] Heinrich, P., Kahn, J. (2018). *Strong Identifiability and Optimal Minimax Rates for Finite Mixture Estimation. The Annals of Statistics*, 46, 2844–2870.

[4] Gerogiannis, D., Nikou, C., Likas, A. *The Mixtures of Student's t-Distributions as a Robust Framework for Rigid Registration.*

[5] Revillon, G., Mohammad-Djafari, A., and Enderli, C. *A Generalized Multivariate Student-t Mixture Model for Bayesian Classification and Clustering of Radar Waveforms.*

A Reminder on the t-Student Distribution

The t-Student distribution is parameterized by a location parameter μ, a scale parameter σ, and degrees of freedom $\lambda > 0$, which control the heaviness of the tails. Its probability density function (PDF) is given by:

$$f(x;\mu,\sigma,\lambda) = \frac{\Gamma\left(\frac{\lambda+1}{2}\right)}{\sqrt{\lambda\pi}\sigma\,\Gamma\left(\frac{\lambda}{2}\right)}\left(1+\frac{1}{\lambda}\left(\frac{x-\mu}{\sigma}\right)^2\right)^{-\frac{\lambda+1}{2}},$$

where $\Gamma(\cdot)$ is the Gamma function; μ represents the location (mean for $\lambda > 1$); σ is the scale (standard deviation for $\lambda > 2$); λ is the degrees of freedom, which control the tail behavior.

In the following, we will assume that σ is equal to 1 for sake of simplicity.

Moments of the t-Student Distribution

For a t-Student distribution with degrees of freedom $\lambda > 0$:

- The **mean** exists if $\lambda > 1$ and is given by:

$$\mathbb{E}[X] = \mu.$$

- The **variance** exists if $\lambda > 2$ and is:

$$\mathrm{Var}(X) = \frac{\lambda}{\lambda - 2}$$

- Higher moments exist if $\lambda > k$, where k is the order of the moment.
- The **kurtosis** (for $\lambda > 4$) is:
$$\kappa = \frac{6}{\lambda - 4}.$$

We notice that unlike the gaussian distribution, the t student distribution has a finite number of finite moment (exactly $\lambda - 1$). Moreover, the kurtosis (the size of the tail of the distribution) decreases with λ.

B Definitions and notations

Definition q-strongly identifiable

A family of distribution functions $\{F(\cdot;\theta) : \theta \in \Theta\}$ for $\Theta \subset \mathbb{R}$, is said to be q-strongly identifiable if for any finite subset B of Θ,

$$\sup_{x \in \mathbb{R}} \left| \sum_{j=0}^{q} \sum_{\theta' \in B} a_{j,\theta'} \frac{\partial^j f}{\partial \theta^j}(x;\theta') \right| = 0 \quad \Longrightarrow \quad \max_{j \in \{0,1,\ldots,q\}} \max_{\theta' \in B} |a_{j,\theta'}| = 0.$$

We say that a mixture distribution $\nu * F$ is q-strongly identifiable if $\{F(\cdot;\theta) : \theta \in \Theta\}$ is q-strongly identifiable.

Definitions

- **Mixing Distribution Space:** Let $\mathcal{M}(\Theta)$ denote the set of all probability measures on Θ. Specifically, when $\Theta = [-L, L]$ for some $L > 0$, we write $\mathcal{M}([-L, L])$.
- **Finite Mixtures:** Define $\mathcal{M}_k \subset \mathcal{M}([-L, L])$ as the subset of mixing distributions that are discrete with at most k atoms.
- ***p*-Wasserstein distance**: between two probability measures μ and ν on a metric space $(\mathbb{R}, d)$ is defined as:
$$W_p(\mu, \nu) = \left(\inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}\times\mathbb{R}} d(x,y)^p \, d\gamma(x,y) \right)^{1/p},$$
where:
  - $\Gamma(\mu, \nu)$ is the set of all couplings of μ and ν, i.e., the set of joint distributions on $\mathbb{R} \times \mathbb{R}$ with marginals μ and ν.
  - $d(x, y)$ is the metric on $\mathbb{R}$, often the absolute difference $d(x, y) = |x - y|$.

For $p = 1$, the Wasserstein distance simplifies to:

$$W_1(\mu, \nu) = \inf_{\gamma \in \Gamma(\mu,\nu)} \int_{\mathbb{R}\times\mathbb{R}} |x - y| \, d\gamma(x, y).$$

Notations

- $X_{1:n} := (X_1, \ldots, X_n)$

C Assumption P on the prior

Recall that k_n is the known upper bound on the true number of components. The prior distribution Π satisfies the following conditions:

- **(P1)** The prior distribution on the number of components k is sample size dependent. There exist constants $c_1 > 0$ and a sufficiently large constant $A > 0$ such that for any sample size $n \in \mathbb{N}$ and any $k^\circ \in \mathbb{N}$,

  $$\frac{\Pi(k = k^\circ + 1)}{\Pi(k = k^\circ)} \leq c_1 e^{-Ak_n \log n}.$$

  Additionally, there exist constants $c_2 > 0$ and $c_3 > 0$ such that for any $n \in \mathbb{N}$ and $k^\dagger \in [k_n]$,

  $$\Pi(k = k^\dagger) \geq c_2 e^{-(c_3 k_n \log n) k^\dagger}.$$

- **(P2)** For any $k \in \mathbb{N}$ and any $(w_1^0, \ldots, w_k^0) \in \Delta_k$, there exist positive constants c_4 and c_5 such that for any $\eta \in (0, 1/k)$,

  $$\Pi\left(\sum_{j=1}^{k} |w_j - w_j^0| \leq \eta \,\middle|\, k\right) \geq c_4 \eta^{c_5 k}.$$

- **(P3)** For any $k \in \mathbb{N}$ and any $\theta_0 \in [-L, L]^k$, there exist positive constants c_6 and c_7 such that for any $\eta > 0$,

  $$\Pi\left(\max_{1\leq j\leq k} |\theta_j - \theta_j^0| \leq \eta \,\middle|\, k\right) \geq c_6 \eta^{c_7 k}.$$

In **(P1)**, the prior distribution heavily penalizes mixture models with a large number of components and assumes that this penalization becomes more severe as the sample size grows. This ensures that the posterior distribution does not overestimate the number of components.

D Verification of Assumptions (F1), (F2), (F^{*1}), and (F^{*2}) for the t-Student Mixture Model

We consider the univariate t-**Student mixture model** on a compact parameter set $\Theta \subset \mathbb{R}$. For each $\theta \in \Theta$, we define

$$F(x;\theta) \;=\; T_\nu(x-\theta),$$

where T_ν is the univariate t-Student distribution with degrees of freedom $\nu > 0$, *location* θ, and (for simplicity) unit scale. The corresponding density function is

$$f(x;\theta) \;=\; C_\nu \left(1 + \tfrac{(x-\theta)^2}{\nu}\right)^{-\frac{\nu+1}{2}}, \quad C_\nu \;=\; \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\,\pi}\,\Gamma\left(\frac{\nu}{2}\right)}.$$

Below we verify that this family $\{F(\cdot;\theta) : \theta \in \Theta\}$ satisfies the assumptions **(F1)**, **(F2)**, **(F*1)**, and **(F*2)** stated in the main text.

(F1) Compactness of the parameter space

- **(F1)** requires that Θ is a compact subset of $\mathbb{R}$ with nonempty interior.

This is satisfied by design: choose any closed interval $\Theta = [-L, L]$ with $L > 0$. Clearly, Θ is compact and its interior $(-L, L)$ is nonempty.

(F2) Lipschitz condition & ratio-integrability condition

(i) Lipschitz continuity in θ. We need to show there exists a constant $c_1 > 0$ such that for all $\theta_1, \theta_2 \in \Theta$,

$$\|\, f(\cdot,\theta_1) \;-\; f(\cdot,\theta_2)\,\|_\infty \;\le\; c_1\,|\theta_1-\theta_2|.$$

Consider

$$f(x,\theta) \;=\; C_\nu \left(1 + \tfrac{(x-\theta)^2}{\nu}\right)^{-\frac{\nu+1}{2}}.$$

By the mean-value theorem,

$$\left|f(x,\theta_1) - f(x,\theta_2)\right| \;\le\; \max_{\xi\in[\theta_1,\theta_2]}\left|\tfrac{\partial}{\partial\theta} f(x,\xi)\right|\,|\theta_1-\theta_2|.$$

A direct computation shows

$$\frac{\partial}{\partial\theta}\left[(1 + \tfrac{(x-\theta)^2}{\nu})^{-\frac{\nu+1}{2}}\right] = -\frac{\nu+1}{\nu}\;\frac{(x-\theta)}{\left(1+\frac{(x-\theta)^2}{\nu}\right)^{\frac{\nu+3}{2}}}.$$

Multiplying by C_ν, we see that this derivative is uniformly bounded in x and θ, because although $|x-\theta|$ can be large, the polynomial decay $\bigl(1+\frac{(x-\theta)^2}{\nu}\bigr)^{-\frac{\nu+3}{2}}$ suppresses the tails. Hence there exists

$$c_1 \;=\; C_\nu \sup_{x,\theta}\Bigl|\frac{\partial}{\partial\theta}\left[(1+\tfrac{(x-\theta)^2}{\nu})^{-\frac{\nu+1}{2}}\right]\Bigr| \;<\; \infty.$$

Therefore

$$\|\, f(\cdot,\theta_1) \;-\; f(\cdot,\theta_2)\,\|_\infty \;\le\; c_1\,|\theta_1-\theta_2|.$$

Thus the Lipschitz condition holds.

(ii) Mixture ratio-integrability condition. Here we prove that there exist a constant $c_2>0$ and an exponent $r\in(0,1]$ such that

$$\int_{\mathbb{R}}\Bigl[p_{\nu_1*F}(x)\Bigl(\tfrac{p_{\nu_1*F}(x)}{p_{\nu_2*F}(x)}\Bigr)^r\Bigr]\,dx \quad\le\quad c_2,$$

for every pair of probability measures $\nu_1,\nu_2\in\mathcal{M}(\Theta)$, where

$$p_{\nu*F}(x) \;=\; \int_\Theta f(x,\theta)\,\nu(d\theta).$$

Step 1: Strict positivity and tail behavior

Since the t-Student density $f(x,\theta)$ is strictly positive for every real x and every $\theta\in\Theta$, the mixture density

$$p_{\nu*F}(x) \;=\; \int f(x,\theta)\,\nu(d\theta)$$

is also strictly positive for all $x\in\mathbb{R}$.

Moreover, for large $|x|$, we know that the t-Student density $\bigl(1+(x-\theta)^2/\nu\bigr)^{-(\nu+1)/2}$ decays polynomially in $|x|$. Because $\theta\in[-L,L]$ and L is finite, there are constants $a,A>0$ (independent of θ) such that, for sufficiently large $|x|$,

$$a\,|x|^{-(\nu+1)} \;\le\; f(x,\theta) \;\le\; A\,|x|^{-(\nu+1)}.$$

Hence each mixture $p_{\nu*F}(x)$ inherits a similar polynomial decay. In particular, no mixture density goes to zero or infinity faster than some fixed polynomial rate as $|x|\to\infty$.

Step 2: Uniform boundedness of the mixture-density ratio

We now argue that the ratio

$$\frac{p_{\nu_1*F}(x)}{p_{\nu_2*F}(x)}$$

remains bounded above by some global constant, uniformly in $x \in \mathbb{R}$. Once we establish this uniform bound, the desired integral bound will follow easily.

Claim: *There is a finite constant $R \geq 1$ such that for all $x \in \mathbb{R}$,*

$$0 \ < \ \frac{p_{\nu_1 * F}(x)}{p_{\nu_2 * F}(x)} \ \leq \ R.$$

Proof of the Claim. Define

$$m(x) \ = \ \min_{\theta \in \Theta} f(x, \theta), \quad M(x) \ = \ \max_{\theta \in \Theta} f(x, \theta).$$

Since Θ is compact and $\theta \mapsto f(x, \theta)$ is continuous, both $m(x)$ and $M(x)$ are well-defined and finite for each x. Moreover, $m(x) > 0$ because t-Student densities are strictly positive.

For any $\nu_1, \nu_2 \in \mathcal{M}(\Theta)$, the mixture densities can be bounded as

$$p_{\nu_1 * F}(x) \ = \ \int f(x, \theta)\, \nu_1(d\theta) \ \leq \ \int M(x)\, \nu_1(d\theta) \ = \ M(x).$$

Similarly,

$$p_{\nu_2 * F}(x) \ = \ \int f(x, \theta)\, \nu_2(d\theta) \ \geq \ \int m(x)\, \nu_2(d\theta) \ = \ m(x).$$

Hence

$$\frac{p_{\nu_1 * F}(x)}{p_{\nu_2 * F}(x)} \ \leq \ \frac{M(x)}{m(x)}.$$

Thus the problem reduces to showing $\sup_x \frac{M(x)}{m(x)}$ is finite.

- *For bounded* x (e.g. $|x| \leq x_0$): the function $(x, \theta) \mapsto f(x, \theta)$ is continuous on the compact set $\{|x| \leq x_0\} \times \Theta$. Hence $m(x)$ and $M(x)$ vary continuously on a compact domain and remain away from 0 and ∞. Therefore $M(x)/m(x)$ attains some finite maximum on $\{|x| \leq x_0\}$.

- *For large* $|x|$ (i.e. $|x| > x_0$): each $f(x, \theta)$ decays polynomially at the same rate in $|x|$ (since $\theta \in [-L, L]$ just shifts x by a bounded amount). Concretely, for sufficiently large $|x|$,

  $$\bigl(1 + \tfrac{(x-\theta)^2}{\nu}\bigr)^{-\frac{\nu+1}{2}} \ \approx \ \kappa\, |x|^{-(\nu+1)}$$

  for some positive κ. Because all $\theta \in \Theta$ lie within $\pm L$, there exist uniform constants $\underline{c}$ and $\overline{c}$ such that

  $$\underline{c}\, |x|^{-(\nu+1)} \ \leq \ f(x, \theta) \ \leq \ \overline{c}\, |x|^{-(\nu+1)},$$

  whenever $|x| > x_0$. Consequently,

  $$m(x) \ \geq \ \underline{c}\, |x|^{-(\nu+1)}, \quad M(x) \ \leq \ \overline{c}\, |x|^{-(\nu+1)}.$$

Thus

$$\frac{M(x)}{m(x)} \;\le\; \frac{\overline{c}\,|x|^{-(\nu+1)}}{\underline{c}\,|x|^{-(\nu+1)}} \;=\; \frac{\overline{c}}{\underline{c}},$$

a constant. Therefore $M(x)/m(x)$ remains bounded as $|x| \to \infty$.

Since $M(x)/m(x)$ is bounded for $|x| \le x_0$ and also for $|x| > x_0$, it follows that $\sup_{x\in\mathbb{R}} \frac{M(x)}{m(x)} < \infty$. Hence the ratio $\sup_x \frac{p_{\nu_1 * F}(x)}{p_{\nu_2 * F}(x)}$ is also finite. Denote this supremum by $R \ge 1$. This completes the proof of the claim.

Step 3: Completing the integral bound

We now have, for each $x \in \mathbb{R}$,

$$0 \;<\; \left(\tfrac{p_{\nu_1 * F}(x)}{p_{\nu_2 * F}(x)}\right)^r \;\le\; R^r.$$

Thus

$$p_{\nu_1 * F}(x) \left(\frac{p_{\nu_1 * F}(x)}{p_{\nu_2 * F}(x)}\right)^r \;\le\; R^r\, p_{\nu_1 * F}(x).$$

Integrating over $\mathbb{R}$ in x gives

$$\int_{\mathbb{R}} p_{\nu_1 * F}(x) \left(\frac{p_{\nu_1 * F}(x)}{p_{\nu_2 * F}(x)}\right)^r dx \;\le\; R^r \int_{\mathbb{R}} p_{\nu_1 * F}(x)\, dx.$$

But $\int p_{\nu_1 * F}(x)\, dx = 1$, since $p_{\nu_1 * F}$ is a probability density. Therefore

$$\int_{\mathbb{R}} p_{\nu_1 * F}(x) \left(\frac{p_{\nu_1 * F}(x)}{p_{\nu_2 * F}(x)}\right)^r dx \;\le\; R^r.$$

Hence the desired uniform integrability condition holds with $c_2 = R^r$. This completes the proof that there is a finite constant c_2 (depending only on Θ, ν_1, ν_2, and ν) such that

$$\int p_{\nu_1 * F}(x) \left(\frac{p_{\nu_1 * F}(x)}{p_{\nu_2 * F}(x)}\right)^r dx \;\le\; c_2.$$

($\mathbf{F}^*1$) q-differentiability

- **($\mathbf{F}^*1$)**: For each $x \in \mathbb{R}$, $F(x;\theta)$ is q-times differentiable in θ.

The kernel

$$\left(1 + \tfrac{(x-\theta)^2}{\nu}\right)^{-\frac{\nu+1}{2}}$$

is an analytic function of θ for each fixed x. Thus it is infinitely differentiable in θ. Hence for any finite q, the model is q-differentiable.

($\mathbf{F}^*2$) q-strong identifiability

- **($\mathbf{F}^*2$)**: The family $\{f(\cdot;\theta) : \theta \in \Theta\}$ is q-strongly identifiable.

Strong identifiability of the t-Student location family is well-known (and follows from general location-family arguments or from results in, e.g., Heinrich and Kahn, 2018, Theorem 3). Informally, different location parameters θ induce linearly independent "patterns" of partial derivatives with respect to θ, ensuring that no non-trivial linear combination of $\{\partial^j f / \partial\theta^j\}$ can vanish identically. Since the t-kernel is smooth in θ to arbitrarily high order, it is in fact *infinitely* ($q = \infty$) strongly identifiable.

Conclusion

Since all four conditions **(F1)**, **(F2)**, **(F***1**)** and **(F***2**)** are met by the t-Student family with a compact parameter space Θ, the mixture of univariate t-Student distributions satisfies the assumptions of Section 2.6 in Ohn and Lin (2022). This holds for any fixed degrees of freedom $\nu > 0$ and fixed scale (e.g. scale $= 1$).