Parameter estimation and interpretability in Bayesian mixture models

Long Nguyen

Department of Statistics
University of Michigan

BNP12 @Oxford, 6/2019

Joint with Aritra Guha (Michigan) and Nhat Ho (Berkeley)

Outline

Interpretability: parameter vs density estimation

Posterior contraction rates of parameters

Impact of model misspecification

Outline

Interpretability: parameter vs density estimation

Posterior contraction rates of parameters

Impact of model misspecification

(Infinite) mixture models

- Good for modeling heterogeneous and complex data
  - black-box modeling device for density estimation
  - clustering and inference about heterogeneity
  - enabling (near) automatic model selection

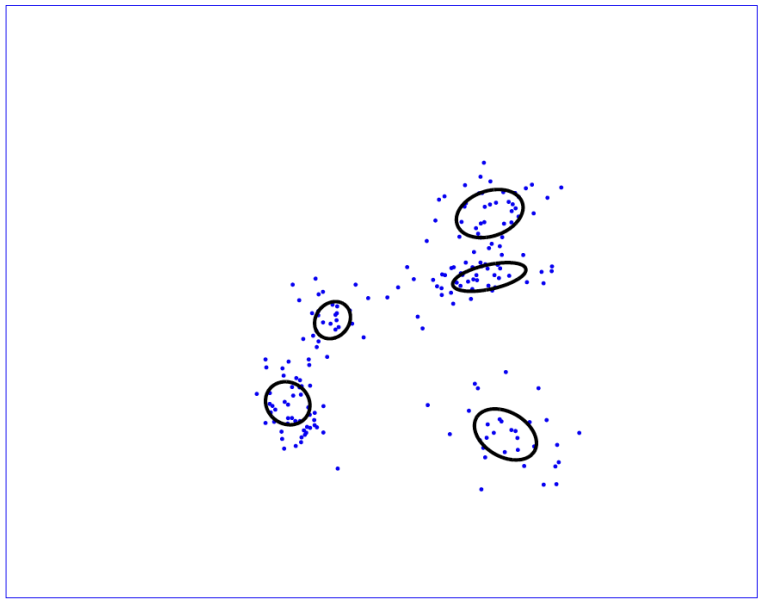
- Questions: quality of parameter estimates and interpretability
  - when number of mixture components unknown?
  - posterior contraction rates for parameters?
  - what is impact of model misspecification?

All mixture models are misspecified...

perhaps some are more interpretable than others

- assume misspecification of either kernel density or mixing measure, what does the posterior contracts to and how fast?
  - broader question is effects of modeling choice on movement of mass from prior to posterior
  - some (misspecified) modeling choice yields faster posterior contraction behavior than others, raising interesting questions regarding interpretability in mixture modeling practice

The BNP way



Mixture model takes the form:

$$p_G(x) = \sum_{i=1}^{k} p_i f(x|\theta_i)$$

Model specification:

$$G = \sum p_i \delta_{\theta_i} \sim \Pi$$

$$X_1, \ldots, X_n | G \overset{iid}{\sim} p_G$$

Posterior inference problems

- density estimation via posterior distribution of p_G?
- parameter estimation: what happens to G a posteriori?

A brief state of affair for DP mixtures

For density estimation:

- the posterior on p_G contracts to true data density at an optimal rate up to a logarithmic factor, under suitable smoothness conditions (Ghosal et al, 1999; Ghosal & van der Vaart, 2001; Lijoi et al, 2005; Tokdar, 2006; Ghosal & van der Vaart, 2007; Walker et al, 2007; Kruijer et al, 2010, Shen et al, 2013; and subsequent works)

For clustering: one needs to be careful

- placing a nonparametric Bayesian prior such as Dirichlet result in inconsistent estimate of number of clusters (Miller & Harrison, 2014)
- if number of clusters is of interest, one can work with overfitted model via a suitable prior or explicit prior on the number of parameters (Rousseau & Mengersen, 2011; Green & Richardson, 1997; Nobile & Fearnside, 2007)

For parameter estimates: what do we mean by saying that the posterior of G contracts to G_0?

Wasserstein distance W_1

If $G = \sum_{i=1}^{k} p_i \delta_{\theta_i}$ and $G_0 = \delta_{\theta_0}$ on some metric space, then

$$W_1(G, G_0) = \sum_{i=1}^{k} p_i \|\theta_0 - \theta_i\|.$$

Wasserstein distance W_1

If $G = \sum_{i=1}^{k} p_i \delta_{\theta_i}$ and $G_0 = \delta_{\theta_0}$ on some metric space, then

$$W_1(G, G_0) = \sum_{i=1}^{k} p_i \|\theta_0 - \theta_i\|.$$

If $G = \sum_{i=1}^{k} \frac{1}{k} \delta_{\theta_i}$, $G' = \sum_{j=1}^{k} \frac{1}{k} \delta_{\theta'_j}$, then

$$W_1(G, G') = \inf_{\pi} \sum_{i=1}^{k} \frac{1}{k} \|\theta_i - \theta'_{\pi(i)}\|,$$

where π ranges over the set of permutations on $(1, \ldots, k)$.

Wasserstein distance W_1

If $G = \sum_{i=1}^{k} p_i \delta_{\theta_i}$ and $G_0 = \delta_{\theta_0}$ on some metric space, then

$$W_1(G, G_0) = \sum_{i=1}^{k} p_i \|\theta_0 - \theta_i\|.$$

If $G = \sum_{i=1}^{k} \frac{1}{k} \delta_{\theta_i}$, $G' = \sum_{j=1}^{k} \frac{1}{k} \delta_{\theta'_j}$, then

$$W_1(G, G') = \inf_{\pi} \sum_{i=1}^{k} \frac{1}{k} \|\theta_i - \theta'_{\pi(i)}\|,$$

where π ranges over the set of permutations on $(1, \ldots, k)$.

Generally, Wasserstein distance is an *optimal transport* distance which quantifies the movement of mass from one probability measure to another

Wasserstein distance W_r, $r \geq 1$

Definition: Let G, G' be two probability measures on $\mathbb{R}^d$. A transport plan, aka, a coupling κ of G, G' is a joint dist on $\mathbb{R}^d \times \mathbb{R}^d$ which induces marginals G, G'.

Viewing $\|\theta - \theta'\|^r$ is the cost of moving an unit mass from θ to θ', then the optimal transportation cost defines W_r:

$$W_r(G, G') := \left[\inf_{\kappa} \int \|\theta - \theta'\|^r d\kappa(\theta, \theta') \right]^{1/r}.$$

Wasserstein distance W_r, $r \geq 1$

Definition: Let G, G' be two probability measures on $\mathbb{R}^d$. A transport plan, aka, a coupling κ of G, G' is a joint dist on $\mathbb{R}^d \times \mathbb{R}^d$ which induces marginals G, G'.

Viewing $\|\theta - \theta'\|^r$ is the cost of moving an unit mass from θ to θ', then the optimal transportation cost defines W_r:

$$W_r(G, G') := \left[\inf_{\kappa} \int \|\theta - \theta'\|^r d\kappa(\theta, \theta') \right]^{1/r}.$$

Also useful as distance between Bayesian hierarchies (Nguyen, 2016).

Interpretation: if $W_r(G_n, G_0) \asymp \omega_n = o(1)$, and G_0 has finite number of support points

- there are atoms of G_n converging to that of G_0 at rate ω_n
- there are also redundant atoms of G_n vanishing at rate $\omega_n^r \ll \omega_n$

Posterior contraction for DP location mixtures

(Nguyen, 2013): Given n-iid sample from p_{G_0}, obtain $\Pi(G|X_1, \ldots, X_n)$
Rates: depending on the smoothness β of kernel density f

- supersmooth kernel f, rate is $W_2(G, G_0) \lesssim (\log n)^{-1/\beta}$
  - Gaussian kernel: $\beta = 2$.
- ordinary smooth kernel f, rate is $W_2(G, G_0) \lesssim n^{-\gamma}$ for any $\gamma < 2/((d+2)(4+(2\beta+1)d))$
  - Laplace kernel: $\beta = 2$.
- for Laplace, improved bounds for W_1 (Gao & van der Vaart (2016); Donnet et al (2018))

Posterior contraction for DP location mixtures

(Nguyen, 2013): Given n-iid sample from p_{G_0}, obtain $\Pi(G|X_1, \ldots, X_n)$
Rates: depending on the smoothness β of kernel density f

- supersmooth kernel f, rate is $W_2(G, G_0) \lesssim (\log n)^{-1/\beta}$
  - Gaussian kernel: $\beta = 2$.
- ordinary smooth kernel f, rate is $W_2(G, G_0) \lesssim n^{-\gamma}$ for any $\gamma < 2/((d+2)(4+(2\beta+1)d))$
  - Laplace kernel: $\beta = 2$.
- for Laplace, improved bounds for W_1 (Gao & van der Vaart (2016); Donnet et al (2018))

Post-processing sample G (Guha, Ho & N, 2019)

- suppose $W_2(G, G_0) = o_p(\omega_n)$, G is a posterior sample
- sequentially and probabilistically merge all atoms that are within distance ω_n
- truncate/merge clusters with total weight less than a threshold to obtain $\tilde{G}$
- resulting $|\tilde{G}|$ gives consistent estimate of $|G_0|$

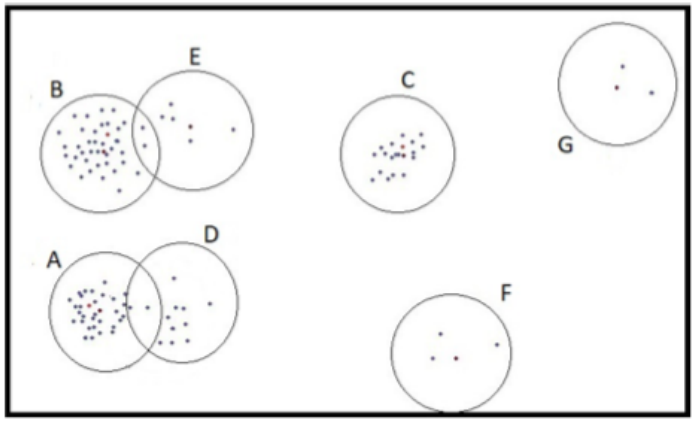


Figure 1: Initial distribution G.

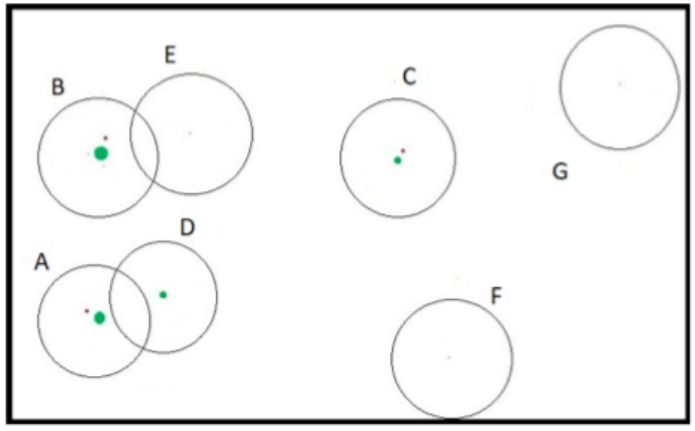


Figure 2: After first stage-"merge".

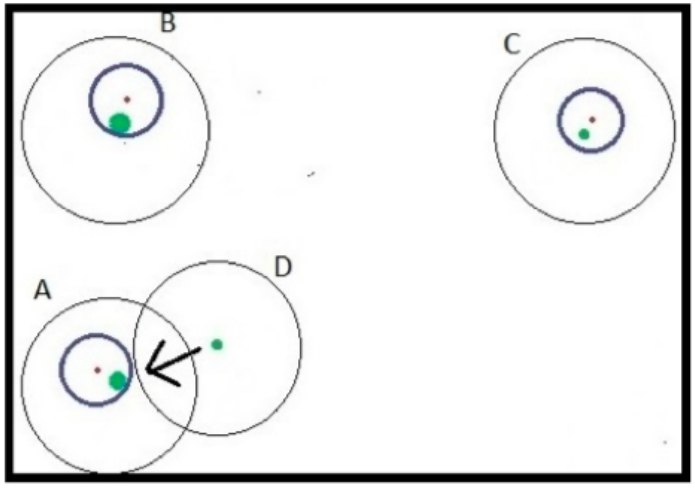


Figure 3: After second stage-"truncation".

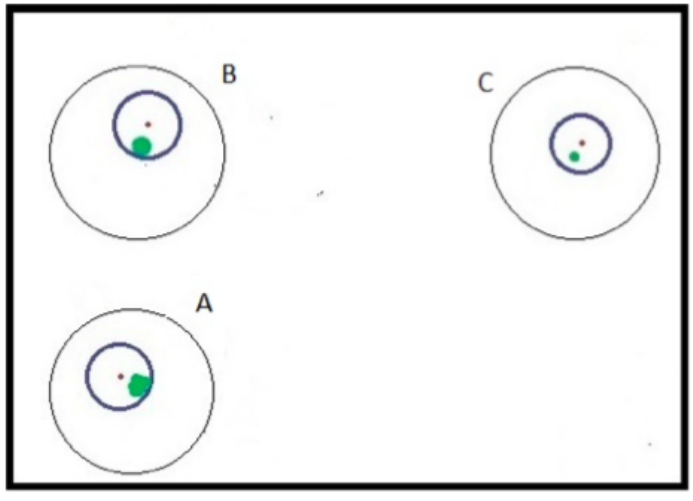


Figure 4: After second stage-"merge".

Merge-Truncate-Merge (MTM) algorithm

INPUT: posterior sample $G = \sum_i p_i \delta_{\theta_i}$, rate ω_n, constant $c > 0$.

Stage 1: Merge procedure:

- Reorder atoms $\{\theta_1, \theta_2, \dots\}$ by sampling with replacement with weights $\{p_1, p_2, \dots\}$. And then let $\tau_1, \tau_2, \dots$ be the new indices, set $\mathcal{E} = \{\tau_j\}_j$ as existing set of atoms.
- Sequentially for each index $\tau_j \in \mathcal{E}$, if there exists an index $\tau_i < \tau_j$ such that $\|\theta_{\tau_i} - \theta_{\tau_j}\| \leq \omega_n$, then:

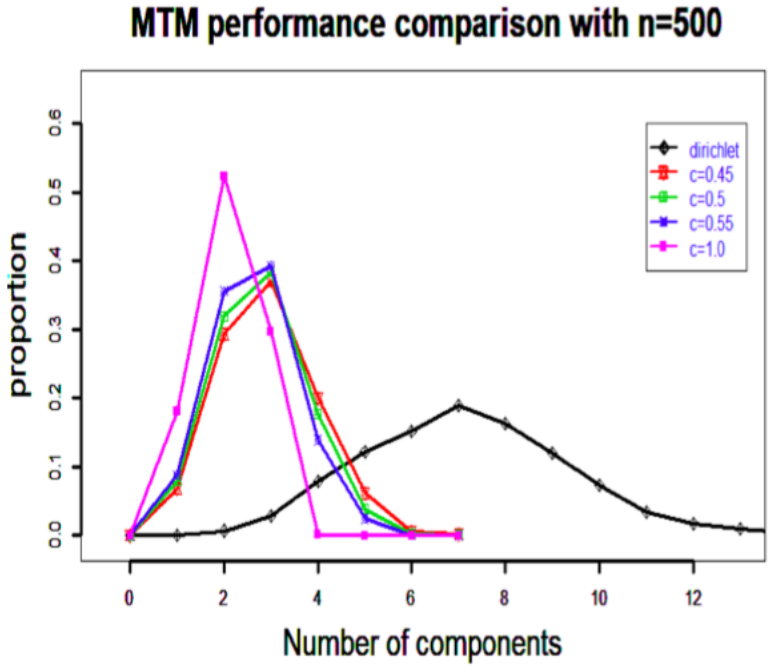
  update $p_{\tau_i} = p_{\tau_i} + p_{\tau_j}$, and remove τ_j from $\mathcal{E}$.
- Collect $G' = \sum_{j:\ \tau_j \in \mathcal{E}} p_{\tau_j} \delta_{\theta_{\tau_j}} := \sum_{i=1}^{k} q_i \delta_{\phi_i}$ so that $q_1 \geq q_2 \geq \dots$.

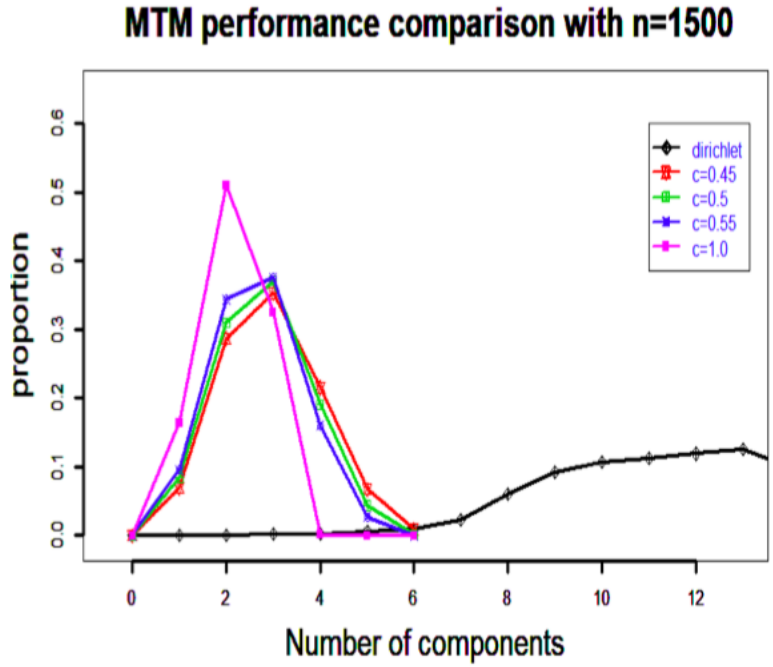
Stage 2: Truncate-Merge procedure:

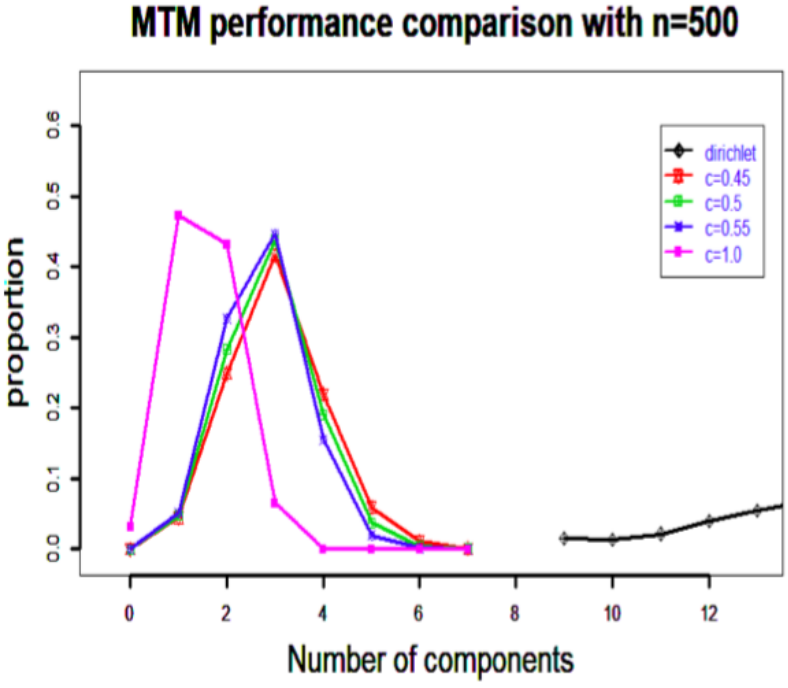
- Set $\mathcal{A} = \{i : q_i > (c\omega_n)^r\}$, $\mathcal{N} = \{i : q_i \leq (c\omega_n)^r\}$.
- For each index $i \in \mathcal{A}$, if there is $j \in \mathcal{A}$ such that $j < i$ and $q_i \|\phi_i - \phi_j\|^r \leq (c\omega_n)^r$, then remove i from $\mathcal{A}$ and add it to $\mathcal{N}$.
- For each $i \in \mathcal{N}$, find atom ϕ_j among $j \in \mathcal{A}$ that is nearest to ϕ_i

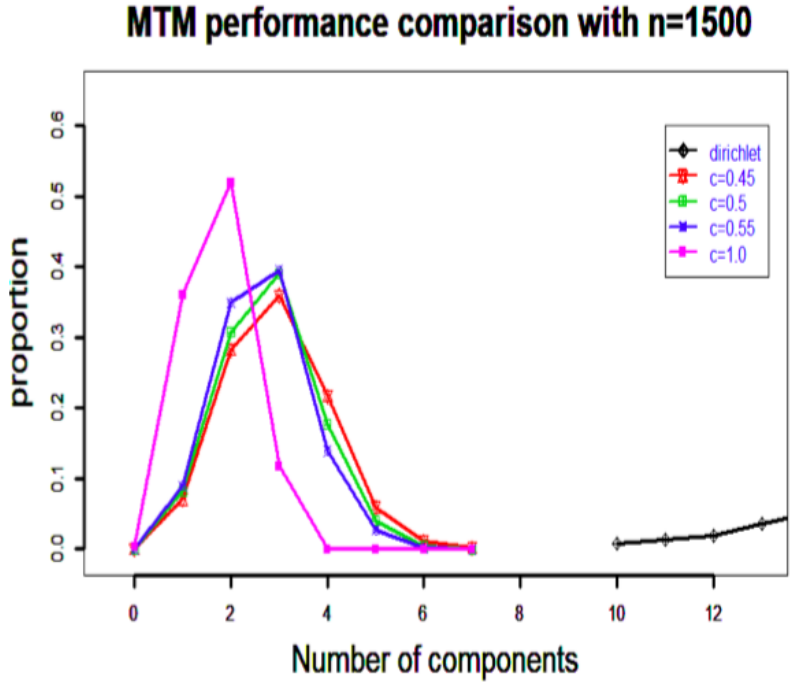
  update $q_j = q_j + q_i$.

OUTPUT: $\tilde{G} = \sum_{j \in \mathcal{A}} q_j \delta_{\phi_j}$ and $\tilde{k} = |\mathcal{A}|$.


MTM performance comparison with n=500
proportion
Number of components
dirichlet
c=0.45
c=0.5
c=0.55
c=1.0


MTM performance comparison with n=1500
proportion
Number of components
dirichlet
c=0.45
c=0.5
c=0.55
c=1.0


MTM performance comparison with n=500
proportion
Number of components
dirichlet
c=0.45
c=0.5
c=0.55
c=1.0


MTM performance comparison with n=1500
proportion
Number of components
dirichlet
c=0.45
c=0.5
c=0.55
c=1.0

MTM post-processing leads to consistent estimate of k.

Let G be a posterior sample from posterior distribution of any Bayesian procedure, namely, $\Pi(\cdot|X_1, \ldots, X_n)$ according to which the upper bound holds for all $\delta > 0$.

$$\Pi\left(G \in \overline{\mathcal{G}}(\Theta) : W_r(G, G_0) \leq \delta\omega_n \middle| X_1, \ldots, X_n \right) \xrightarrow{p_{G_0}} 1 \tag{1}$$

Theorem

Given posterior sample G for which (1) holds. Let $\tilde{G}$ and $\tilde{k}$ be the outcome of Algorithm MTM applied to G for an arbitrary constant $c > 0$. Then the following hold as $n \to \infty$.

(a) *For all $\delta > 0$,* $\Pi\left(G \in \overline{\mathcal{G}}(\Theta) : W_r(\tilde{G}, G_0) \leq \delta\omega_n \middle| X_1, \ldots, X_n \right) \xrightarrow{p_{G_0}} 1$

(b) $\boxed{\Pi(\tilde{k} = k_0 | X_1, \ldots, X_n) \xrightarrow{p_{G_0}} 1}$.

quick summary

- distribution of $\tilde{G}$ is a meaningful Bayesian quantity, obtained via *push-forward map* MTM $\mathcal{T}$ applied to the posterior of G:

$$\text{law}(\tilde{G}) = \Pi(G|X_1, \ldots, X_n)\#\mathcal{T}$$

- nonetheless, posterior contraction behavior for components parameters of $\tilde{G}$ remains slow
  - tiny mass of many redundant atoms for G is the culprit
  - this is the price we pay for being nonparametric: what works well for density estimation may not work as well for parameter estimation performance (and interpretability)
- what can we do to improve parameter estimation behavior?

Outline

Interpretability: parameter vs density estimation

Posterior contraction rates of parameters

Impact of model misspecification

Order of identifiability

recall that a mixture density is a linear combination of such kernels f

- classical 0-identifiability: if kernel density functions $\{f(x|\theta)|\theta \in \mathbb{R}^d\}$ are linearly independent
- 1-identifiability: the kernel density f, its 1st order derivative of f wrt parameter θ are linearly independent
- 2-identifiability: linear independence up to 2nd derivatives
- ∞-identifiability: ...

Order of identifiability

recall that a mixture density is a linear combination of such kernels f

- classical 0-identifiability: if kernel density functions $\{f(x|\theta)|\theta \in \mathbb{R}^d\}$ are linearly independent
- 1-identifiability: the kernel density f, its 1st order derivative of f wrt parameter θ are linearly independent
- 2-identifiability: linear independence up to 2nd derivatives
- ∞-identifiability: ...

> Most single-parameter families of kernel densities (e.g., exponential families of distributions) are identifiable in arbitrary order

Parameter estimation rates under 2-strong identifiability (Chen, 1995; Rousseau & Mengersen, 2011; Nguyen, 2013, Ho & Nguyen, 2016); minimax optimal rates under (very) strong identifiability conditions (Heinrich and Kahn, 2018)

How common is weak identifiability? Very.

For Gaussian kernel density $f(x|\mu, v)$,

$$\frac{\partial^2 f}{\partial \mu^2} = 2\frac{\partial f}{\partial v}.$$

How common is weak identifiability? Very.

For Gaussian kernel density $f(x|\mu, v)$,

$$\frac{\partial^2 f}{\partial \mu^2} = 2\frac{\partial f}{\partial v}.$$

For gamma kernel density $f(x|a, b)$,

$$\frac{\partial f}{\partial b}(x|a, b) = \frac{a}{b} f(x|a, b) - \frac{a}{b} f(x|a + 1, b).$$

How common is weak identifiability? Very.

For Gaussian kernel density $f(x|\mu, v)$,

$$\frac{\partial^2 f}{\partial \mu^2} = 2\frac{\partial f}{\partial v}.$$

For gamma kernel density $f(x|a, b)$,

$$\frac{\partial f}{\partial b}(x|a, b) = \frac{a}{b} f(x|a, b) - \frac{a}{b} f(x|a+1, b).$$

For skewnormal kernel density $f(x|\theta, v, m)$,

$$\frac{\partial^2 f}{\partial \theta^2} - 2\frac{\partial f}{\partial v} + \frac{m^3 + m}{v}\frac{\partial f}{\partial m} = 0.$$

How common is weak identifiability? Very.

For Gaussian kernel density $f(x|\mu, v)$,

$$\frac{\partial^2 f}{\partial \mu^2} = 2\frac{\partial f}{\partial v}.$$

For gamma kernel density $f(x|a, b)$,

$$\frac{\partial f}{\partial b}(x|a, b) = \frac{a}{b} f(x|a, b) - \frac{a}{b} f(x|a+1, b).$$

For skewnormal kernel density $f(x|\theta, v, m)$,

$$\frac{\partial^2 f}{\partial \theta^2} - 2\frac{\partial f}{\partial v} + \frac{m^3 + m}{v}\frac{\partial f}{\partial m} = 0.$$

$$2m\frac{\partial f}{\partial m} + (m^2 + 1)\frac{\partial^2 f}{\partial m^2} + 2vm\frac{\partial^2 f}{\partial v \partial m} = 0.$$

Overfitted mixtures under 2-identifiability condition

Nguyen (2013): e.g., location Gaussian mixtures, scale Gaussian mixtures, compact parameter space

- Data are n-iid sample from a k_0-mixture, where $k_0 < k$
- Placing any "standard" prior on mixing measures with k atoms $G = \sum_{i=1}^{k} p_i \delta_{\theta_i}$
- Then, the posterior for G contracts to G_0 at rate $O_p((\log n/n)^{1/4})$ under W_2

Consequences:

- redundant weights vanish at rate $O_p(\log n/n)^{1/2}$
- component parameters contract a posteriori at rate $O_p(\log n/n)^{1/4}$

Overfitted mixtures with weakly identifiable kernels

Ho & Nguyen (2016): overfitted location-scale Gaussian mixtures

- Same overfitted setting as before

> Posterior contraction of G to G_0 under Wasserstein metric occurs at rate is affected by how much overfiting

- overfitting by one: $k - k_0 = 1$, then $W_4(G, G_0) = O_p((\log n/n)^{1/8})$ which implies redundant weights vanish at $n^{-1/2}$ rate, but component parameters converge at $n^{-1/8}$ rate
- overfitting by two: $k - k_0 = 2$, then $W_6(G, G_0) = O_p((\log n/n)^{1/12})$ which implies component parameters converge at $n^{-1/12}$ rate
- generally, for $k > k_0$, precise rate r relates to dimension of some *real affine varieties* (solving a system of polynomial equations), so that

$$W_r(G_0, G) \lesssim V(p_{G_0}, p_G)^{1/r}$$

Polynomial eqns derived from Gaussian kernel's PDE

For $r \geq 1$ there are r equations for $3(k - k_0 + 1)$ vars $(c_j, a_j, b_j)_{j=1}^{k-k_0+1}$

$$\sum_{j=1}^{k-k_0+1} \sum_{n_1+2n_2=\alpha} \frac{c_j^2 a_j^{n_1} b_j^{n_2}}{n_1! n_2!} = 0 \quad \text{for each } \alpha = 1, \ldots, r$$

Polynomial eqns derived from Gaussian kernel's PDE

For $r \geq 1$ there are r equations for $3(k - k_0 + 1)$ vars $(c_j, a_j, b_j)_{j=1}^{k-k_0+1}$

$$\boxed{\sum_{j=1}^{k-k_0+1} \sum_{n_1+2n_2=\alpha} \frac{c_j^2 a_j^{n_1} b_j^{n_2}}{n_1! n_2!} = 0 \quad \text{for each } \alpha = 1, \ldots, r}$$

Example: if $k = k_0 + 1$, and let $r = 3$, then

$$\begin{aligned}
&c_1^2 a_1 + c_2^2 a_2 = 0, \\
&\frac{1}{2}(c_1^2 a_1^2 + c_2^2 a_2^2) + c_1^2 b_1 + c_2^2 b_2 = 0, \\
&\frac{1}{3!}(c_1^2 a_1^3 + c_2^2 a_2^3) + c_1^2 a_1 b_1 + c_2^2 a_2 b_2 = 0.
\end{aligned}$$

This has a non-trivial solution, but not if we add another equation corresponding to $r = 4$ to above:

$$\frac{1}{4!}(c_1^2 a_1^4 + c_2^2 a_2^4) + \frac{1}{2!}(c_1^2 a_1^2 b_1 + c_2^2 a_2^2 b_2) + \frac{1}{2!}(c_1^2 b_1^2 + c_2^2 b_2^2) = 0.$$

Hence, we arrive at rate $n^{-1/(2r)} = n^{-1/8}$ when $k = k_0 + 1$.

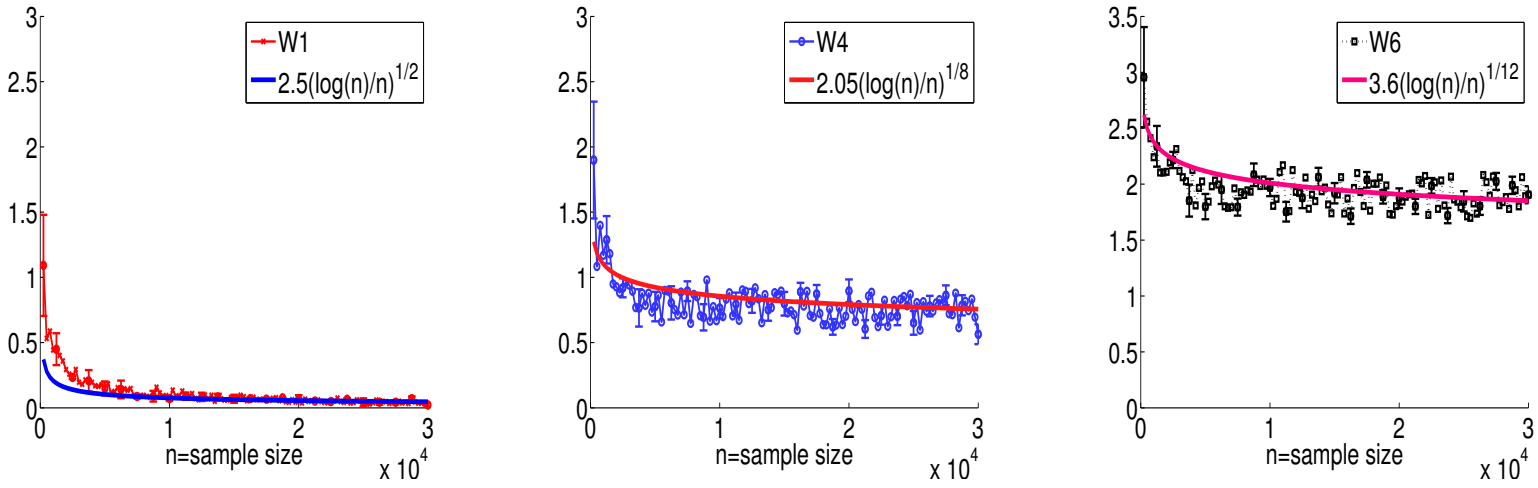


Figure: Posterior contraction rates for location-covariance mixtures of Gaussians.

L to R:
(1) Exact-fitted: $W_1 \asymp O_p(n^{-1/2})$.
(2) Over-fitted by one: $W_4 \asymp O_p(n^{-1/8})$.
(3) Over-fitted by two: $W_6 \asymp O_p(n^{-1/12})$.

Inhomogeneity of ~~data~~ parameter space

Ho & Nguyen (2016, 2018):
most mixture models used in practice can be highly non-regular due to

- the PDE governing the density kernels (Gaussian, Gamma, skewnormal, etc)
- mixture distribution = convex combination of these kernels

Thus, Fisher information matrix degenerates in certain affine varieties forming subsets in parameter space; some are "more" singular than others. So, a standard prior oblivious of these singularities implies that

- parameters of different types may possess different rates of posterior contraction (e.g., scale parameter contracts faster than location for Gaussian components)
- even parameters of the same type may carry distinct rates of estimation (e.g., shape parameters associated with different skewnormal mixture components)

Open question How can we design suitable prior distribution or reparameterization to combat this inhomogeneity of parameter space?

Skewnormal mixtures

Mixture density $p_G(x) = \sum_{j=1}^{k} p_j f(x|\theta_j, \sigma_j, m_j)$, where the skewnormal kernel takes form

$$f(x|\theta, \sigma, m) := \frac{2}{\sigma}\phi\left(\frac{x-\theta}{\sigma}\right)\Phi(m(x-\theta)/\sigma),$$

where $\phi(x)$ is the standard normal density and

$\Phi(x) = \int \phi(t)1(t \le x)\,\mathrm{d}t.$

This generalizes Gaussian densities, which correspond to $m = 0$.

Exact-fitted mixtures: $I(G)$ is singular iif the parameters are real solution of a number of polynomial equations

(i) Type A: $P_1(\eta) = \prod_{j=1}^{k} m_j$.

(ii) Type B:

$$P_2(\eta) = \prod_{1 \le i \ne j \le k}\left\{(\theta_i - \theta_j)^2 + \left[\sigma_i^2(1+m_j^2) - \sigma_j^2(1+m_i^2)\right]^2\right\}.$$

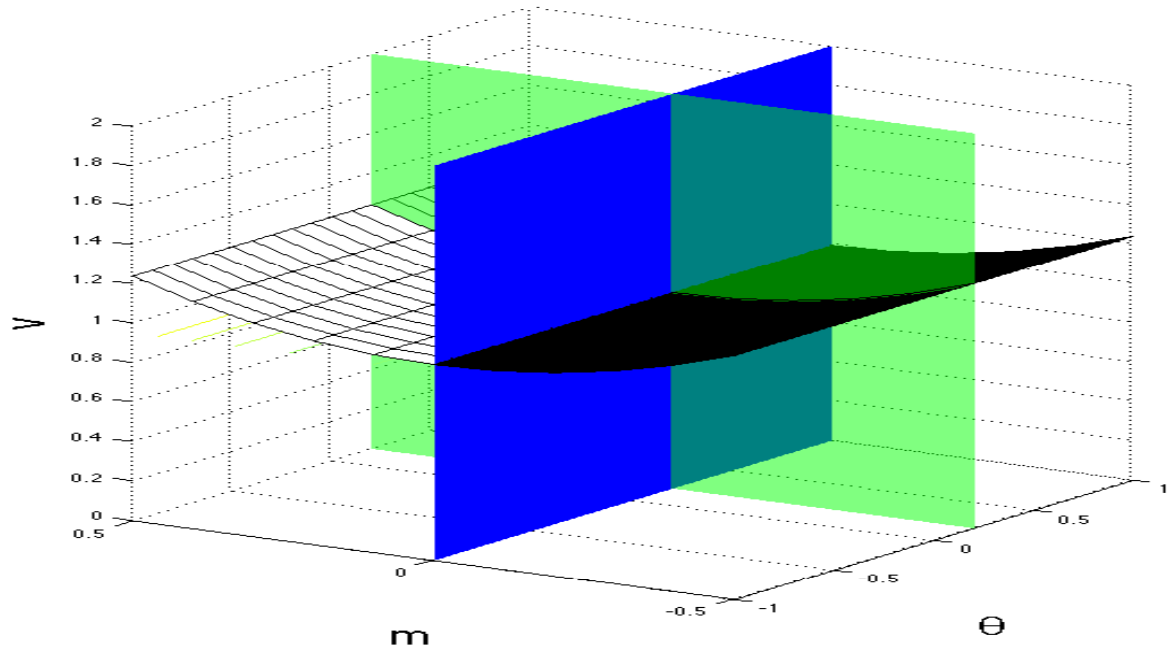


Figure: Illustration of type A and type B singularity (for parameter triplet $(m_1, \theta_1, v_1 = \sigma_1^2)$ for component 1, given a specific value of parameter triplet for component 2

Note: for overfitted mixtures, true model is always singular

Cleverer priors for finite mixtures

A number of ways in handling unknown number of components

(1) overfitted mixtures, with suitable Dirichlet prior on mixing weights (Rousseau & Mengersen, 2011) —> rates $n^{-1/2+\epsilon}$ for any $\epsilon > 0$ under 2-identifiability

(2) overfitted mixtures with "repulsive" mixture components (Petralia et al, 2012; Xie & Xu, 2017)

(3) mixture of finite mixtures: placing a prior on the number of mixing components (Nobile, 1994; Richardson & Green, 1997; Miller & Harrison, 2014)

It can be shown that approach (3) yields adaptive posterior contraction rate for parameters, $(\log n/n)^{1/2}$, under minimal identifiability condition, i.e., 1-identifiability (Guha, Ho & N, 2019)

Cleverer priors for finite mixtures

A number of ways in handling unknown number of components

(1) overfitted mixtures, with suitable Dirichlet prior on mixing weights (Rousseau & Mengersen, 2011) —> rates $n^{-1/2+\epsilon}$ for any $\epsilon > 0$ under 2-identifiability

(2) overfitted mixtures with "repulsive" mixture components (Petralia et al, 2012; Xie & Xu, 2017)

(3) mixture of finite mixtures: placing a prior on the number of mixing components (Nobile, 1994; Richardson & Green, 1997; Miller & Harrison, 2014)

It can be shown that approach (3) yields adaptive posterior contraction rate for parameters, $(\log n/n)^{1/2}$, under minimal identifiability condition, i.e., 1-identifiability (Guha, Ho & N, 2019)

None of this is applicable under misspecification

Outline

All mixture models are misspecified...

Assume the data $X_1, \ldots, X_n$ generated from "true" density p_{G_0, f_0}.

Fitting this with a mixture model $p_{G,f} = \int f(x|\theta) G(\mathrm{d}\theta)$:

- either f is misspecified
- or G is misspecified,
- or both

By Bayes' rule, one obtains $\Pi(G|X_1, \ldots, X_n)$

Question: what becomes of this posterior distribution as $n \to \infty$?

- under some identifiability condition G tends to some G_*
- if $f \neq f_0$, the best we can hope for is $G \to G_*$, but $G_* \neq G_0$
- what does this mean?

... but some may be more interpretable than others

- we know what is bad: if f is wildly different from f_0, then G_* must probably be wildly different from G_0
- mathematically, it is of interest to find the relationship between G_* and G_0
  - standard but challenging question of approximation theory in math. analysis
- from an epistemological standpoint, we can never know the truth about G_0 and f_0, yet we will continue to fit the data with a certain choice of kernel f (Gaussian, Laplace, skewnormal, gamma, etc, etc)
  - it is meaningful to find the posterior contraction behavior to G_*, subject to our choice of kernel f (this says more about the efficiency of the data in moving mass from a prior to a posterior)
  - it is not entirely clear should one prefer a f that results in fast posterior contraction rate to G_*, or a slower rate?

Condition regarding the "truth": this specifies how misspecified are f and G as they are related to f_0, G_0:

(M) The support of G_0, namely, $\mathrm{supp}(G_0)$ is a bounded subset of $\mathbb{R}^d$. Moreover, there are some constants $C_0, C_1, \alpha > 0$ such that for any $R > 0$,

$$\sup_{x \in \mathbb{R}^d, \theta \in \Theta, \theta_0 \in \mathrm{supp}(G_0)} \frac{f(x|\theta)}{f_0(x|\theta_0)} 1_{\|x\|_2 \leq R} \leq C_1 \exp(C_0 R^\alpha).$$

Condition on prior:

(P) A standard prior on G: either a mixture of discrete measures with finite support, or Dirichlet process on bounded subset Θ of $\mathbb{R}^d$ that may or may not contain the support of G_0

Misspecified density

Kleijn & van der Vaart (2006); White (1982): the posterior of p_G contracts to the minimizer of the Kullback-Leibler (KL) distance from the true population density, p_{G_0,f_0}, to a density function residing in the support of the induced prior on p_G, provided that the minimizer exists (in fact, existence of the minimizer implies uniqueness due to convexity of the space).

The KL minimizer can be expressed as a mixture density p_{G_*}, where G_* is a probability measure on Θ. We may write

$$G_* \in \arg\min_{G \in \mathcal{P}(\Theta)} K(p_{G_0,f_0}, p_G).$$

Now, under a standard identifiability condition of f, the uniqueness of p_{G_*} implies that of G_*.

We are interested in the posterior contraction behavior of G toward G_*.

Two choices of kernel for location mixtures

- Gaussian kernel with some fixed $d \times d$ covariance matrix Σ

$$f(x|\theta) = |2\pi\Sigma|^{-1/2} \exp\left\{ -(x-\theta)^\top \Sigma^{-1}(x-\theta)/2 \right\}.$$

- Laplace kernel

$$f(x|\theta) = \frac{2}{\lambda(2\pi)^{d/2}} \frac{K_{(d/2)-1}\left(\sqrt{2/\lambda}\sqrt{(x-\theta)^\top \Sigma^{-1}(x-\theta)}\right)}{\left(\sqrt{\lambda/2}\sqrt{(x-\theta)^\top \Sigma^{-1}(x-\theta)}\right)^{(d/2)-1}},$$

where Σ and $\lambda > 0$ are respectively fixed covariance matrix and scale parameter such that $|\Sigma| = 1$. Here, K_v is a Bessel function of the second kind of order v.

Recall the condition on misspecification:

(M) The support of G_0, namely, $\mathrm{supp}(G_0)$ is a bounded subset of $\mathbb{R}^d$. Moreover, there are some constants $C_0, C_1, \alpha > 0$ such that for any $R > 0$,

$$\sup_{x \in \mathbb{R}^d, \theta \in \Theta, \theta_0 \in \mathrm{supp}(G_0)} \frac{f(x|\theta)}{f_0(x|\theta_0)} 1_{\|x\|_2 \leq R} \leq C_1 \exp(C_0 R^{\alpha}).$$

Theorem (misspecified Gaussian location mixture)

Let f be the Gaussian density kernel and (M) holds. Then, under (P) as n tends to infinity,

$$\Pi\left(G \in \overline{\mathcal{G}}(\Theta) : W_2(G, G_*) \lesssim \left(\frac{\log \log n}{\log n} \right)^{1/2} \middle| X_1, \ldots, X_n \right) \to 1$$

in p_{G_0, f_0}-probability.

(Guha, Ho & N, 2019)

Theorem (misspecified Laplace location mixture)

Let f be the Laplace density kernel for fixed Σ and λ such that $|\Sigma| = 1$ and (M) holds. Then, under (P), as n tends to infinity,

$$\Pi\left(G \in \overline{\mathcal{G}}(\Theta) : W_2(G, G_*) \lesssim \exp\left\{ -\frac{m\tau(\alpha)}{2} \left(\frac{\log n - \log\log n}{2(d+2)} \right)^{1/\alpha} \right\} \middle| X_1, \ldots, X_n \right) \to 1$$

in p_{G_0, f_0}-probability for any positive constant $m < 4/(4+5d)$.
Here, constant $\tau(\alpha)$ takes the form, with $\lambda_{\min}, \lambda_{\max}$ being the minimum and maximum eigenvalues of Σ:

$$\tau(\alpha) := \frac{\sqrt{2/(\lambda\lambda_{\max})}}{\left(\sqrt{2/(\lambda\lambda_{\min})} + \sqrt{2/(\lambda\lambda_{\max})} + C_0 \right)^{1/\alpha}}.$$

For simplicity, if $\alpha = 1$, the posterior contraction bound takes the polynomial rate

$$n^{-\frac{m\tau(1)}{4(d+2)}}$$

Some remarks

(i) Posterior contraction bound for Gaussian location mixtures remains logarithmic rate $(\log\log n/\log n)^{1/2}$ regardless of whether f_0 is misspecified or not. Should we use Gaussian kernel?

Some remarks

(i) Posterior contraction bound for Gaussian location mixtures remains logarithmic rate $(\log\log n/\log n)^{1/2}$ regardless of whether f_0 is misspecified or not. Should we use Gaussian kernel?

- No: this is too slow a movement of mass from prior to posterior
- Yes: when you are misspecified, it's good to be conservative

Some remarks

(i) Posterior contraction bound for Gaussian location mixtures remains logarithmic rate $(\log\log n/\log n)^{1/2}$ regardless of whether f_0 is misspecified or not. Should we use Gaussian kernel?

- No: this is too slow a movement of mass from prior to posterior
- Yes: when you are misspecified, it's good to be conservative

(ii) For Laplace mixtures, the posterior contraction bounds remain $n^{-\gamma'}$, provided $\alpha \geq 1$. Due to misspecification, there is a loss of a constant factor in the exponent γ'.

- consider the scenario where the true kernel f_0 happens to be a Gaussian kernel, but sample size n is small. Should we intentionally misspecify by selecting f to be Laplace instead?

Some remarks

(i) Posterior contraction bound for Gaussian location mixtures remains logarithmic rate $(\log\log n/\log n)^{1/2}$ regardless of whether f_0 is misspecified or not. Should we use Gaussian kernel?

- No: this is too slow a movement of mass from prior to posterior
- Yes: when you are misspecified, it's good to be conservative

(ii) For Laplace mixtures, the posterior contraction bounds remain $n^{-\gamma'}$, provided $\alpha \geq 1$. Due to misspecification, there is a loss of a constant factor in the exponent γ'.

- consider the scenario where the true kernel f_0 happens to be a Gaussian kernel, but sample size n is small. Should we intentionally misspecify by selecting f to be Laplace instead?
- ultimate answer may lie in tension between bias (how far is G_* to G_0) vs the contracting variance (convergence to G_*)

$$W_2(G_0, G_*) >><< W_2(G_n, G_*)$$

Summary

- no size fits all: what's good for density estimation tends to perform poorly from parameter estimation perspective
- parametric vs nonparametric prior specification
- some form of misspecification is more acceptable than others for the sake of interpretability
- perhaps one should also be interested in selective parameter inference (e.g., dominant components or outlying ones)?

Selected References

- J. Miller and M. Harrison. Mixture models with a prior on the number of components. JASA, 2017.
- J. Rousseau and K. Mengersen. Asymptotic behaviour of the posterior distribution of overfitted mixture models. JRSSB, 2011.
- B. Kleijn and A. van der Vaart. Misspecification in infinite-dimensional Bayesian statistics. Annals of Statistics, 2006.
- X. Nguyen. Convergence of latent mixing measures in finite and infinite mixture models. Annals of Statistics, 2013.
- X. Nguyen. Borrowing strength in hierarchical Bayes: posterior concentration of the Dirichlet base measure. Bernoulli, 2016.
- N. Ho and X. Nguyen. Convergence rates of parameter estimation for some weakly identifiable finite mixtures. Annals of Statistics, 2016.
- N. Ho and X. Nguyen. Singularity structures and impacts on parameter estimation in finite mixtures of distributions. arXiv:1609.02655.
- A. Guha, N. Ho and X. Nguyen. Posterior contraction of parameters and interpretability in Bayesian mixture modeling. arXiv:1901.05078.