# Singularity structures and parameter estimation in mixture models

Long Nguyen

Department of Statistics
University of Michigan

Institute for Mathematical Sciences
National University of Singapore
June 2016

Joint work with Nhat Ho (UM)

- machine learning: “model-free optimization-based algorithms”
    - isn’t it the spirit of empirical likelihood based methods?
    - prediction vs estimation/inference

- Bayesian nonparametrics:
    - how to be Bayesian, yet more empirical by being nonparametric

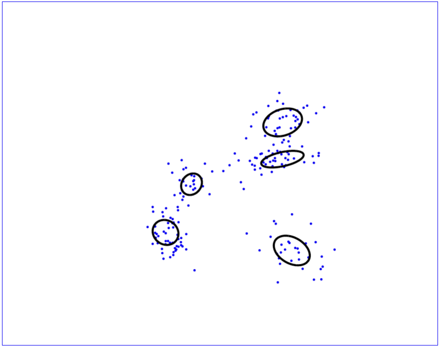
- machine learning: "model-free optimization-based algorithms"
    - isn't it the spirit of empirical likelihood based methods?
    - prediction vs estimation/inference
- Bayesian nonparametrics:
    - how to be Bayesian, yet more empirical by being nonparametric
- empirical likelihood:
    - how to take the inference a bit beyond empirical distributions

- machine learning: "model-free optimization-based algorithms"
  - isn't it the spirit of empirical likelihood based methods?
  - prediction vs estimation/inference
- Bayesian nonparametrics:
  - how to be Bayesian, yet more empirical by being nonparametric
- empirical likelihood:
  - how to take the inference a bit beyond empirical distributions

**modern statistics/ data science**

— data increasingly high-dimensional and complex
— inferential goals increasingly more ambitious
— requiring more sophisticated algorithms and complex statistical modeling

# Mixture modeling



Mixture density

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

e.g.,
$f(x|\eta_i) = \mathrm{Normal}(x|\mu_i, \Sigma_i)$

## Parameter estimation

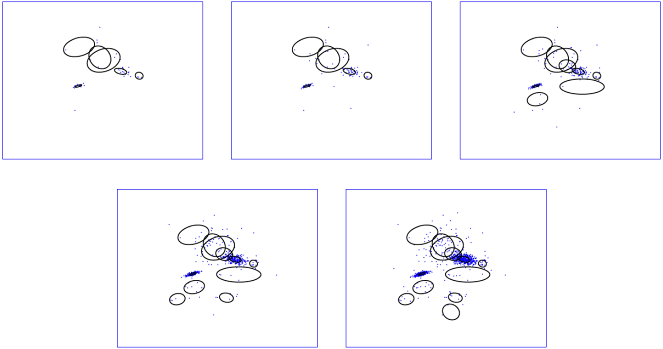
mixing probabilities $\mathbf{p} = (p_1, \ldots, p_k)$?
mixing components $\boldsymbol{\eta} = (\eta_1, \ldots, \eta_k)$?

Text

# Hierarchical models

HM = Mixture of mixture models

Challenge: parameter estimation in latent variable models



[courtesy M. Jordan's slides]

# Estimation in parametric models

Standard methods, e.g., maximum likelihood estimation (via EM) or Bayesian estimation, yield root-$n$ parameter estimation rate

# Estimation in parametric models

Standard methods, e.g., maximum likelihood estimation (via EM) or Bayesian estimation, yield root-$n$ parameter estimation rate

- if Fisher information matrix is *non-singular*

What if we are in a singular situation?

# Fisher singularity

- Cox & Hinkley (1974), Lee & Chesher (1986), Rotnitzky, Cox, Bottai and Robbins (2000), etc: first-order singularity
- Azzalini & Capitanio (1999); Pewsey (2000); DiCiccio & Monti (2004); Hallin & Ley (2012,2014), etc: third order singularities in skewnormal distributions
- Chen (1995); Rousseau & Mengersen (2011); Nguyen (2013): asymptotics for parameter estimation in overfitted mixture models, under strong identifiability conditions
- A full picture of singularity structures in mixture models remain largely unknown (e.g., hitherto there's no asymptotic theory for finite mixtures of location-scale Gaussian mixtures)

Singularity is a common occurence in modern statistics

- high-dimensional and sparse setting
- "complicated" density classes (e.g., gamma, normal, skewnormal), when there are more than one parameter varying
- overfitted/infinite mixture models
- hierarchical models

## Singularities in e-mixtures

# e-mixtures = exact-fitted mixtures

Mixture model indexed by mixing distribution $G = \sum_{i=1}^{k} p_i \delta_{\eta_i}$

$$\left\{ p_G(x) = \sum_{i=1}^{k} p_i f(x|\eta_i) \middle| G \text{ has } k \text{ atoms} \right\},$$

## Fisher information matrix

$$I(G) = \mathbb{E}\left\{ \left( \frac{\partial \log p_G(X)}{\partial G} \right) \left( \frac{\partial \log p_G(X)}{\partial G} \right)^T \right\}$$

where $\partial \log p_G / \partial G$ simply denotes partial derivative of score function wrt all parameters $\mathbf{p}, \boldsymbol{\eta}$

Exploiting the representation $p_G = \sum p_i f(x|\eta_i)$, easy to note that $I(G)$ is non-singular iff the collection of

$$\left\{ f(x|\eta_i), \frac{\partial f}{\partial \eta}(x|\eta_i) \middle| i = 1, \ldots, k \right\}$$

are linearly independent functions of $x$

This condition holds if

- $f(x|\eta) = \mathrm{Gaussian}(x|\mu, \sigma^2)$ location-scale Gaussian
- not for Gamma or skewnormal kernels (and many others)

# Gamma mixtures

## Gamma density

$$f(x|a,b) = \frac{b^a}{\Gamma(a)} x^{a-1} e^{-bx}, a > 0, b > 0$$

admits the following pde

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

For Gamma mixture model

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

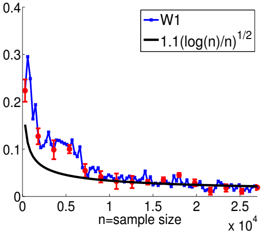
$I(G)$ is a singular matrix if $a_i - a_j = 1$ and $b_i = b_j$ for some pair of components $i, j = 1, \ldots, k$.

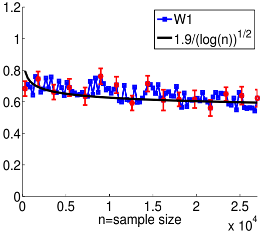
**Theorem** (Ho & Nguyen, 2016)
The minimax rate of estimation for Gamma mixture's parameter is slower than $\frac{1}{n^{1/2r}}$ for any $r \geq 1$, sample size $n$

This is because we can't tell very well when $G$ is singular or not

However if we know that the true $G$ is non-singular, and that $|a_i - a_j| - 1$ and $|b_i - b_j|$ is bounded away from 0, then MLE achieves root-n rate

# Parameter estimation rate for Gamma mixtures




L: MLE restricted to compact set of non-singular $G$: $W_1(\widehat{G}_n, G_0) \asymp n^{-1/2}$.
R: MLE for general set of $G$: $W_1 \approx 1/(\log n)^{1/2}$.

# Skewnormal mixtures

## Skewnormal kernel density

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

where $f(x)$ is the standard normal density and $\Phi(x) = \int f(t)1(t \leq x)\,\mathrm{d}t$.
This generalizes Gaussian densities, which correspond to $m = 0$.

$I(G)$ is singular iif the parameters are real solution of a number of polynomial equations

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

(ii) Type B: $P_2(\boldsymbol{\eta}) = \prod_{1\leq i\neq j\leq k}\left\{(\theta_i-\theta_j)^2 + \left[\sigma_i^2(1+m_j^2) - \sigma_j^2(1+m_i^2)\right]^2\right\}$.

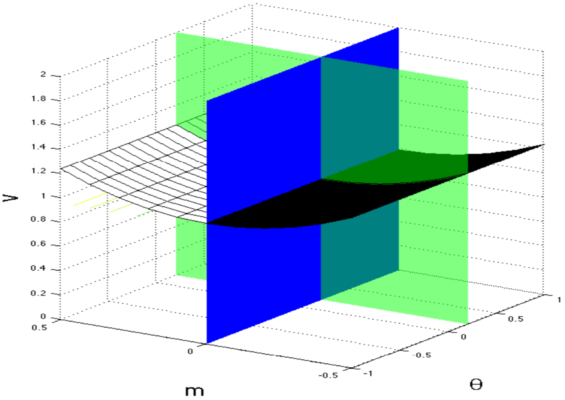



Figure : Illustration of type A and type B singularity.

- Singularities of skewnormal mixtures lie in affine varieties
- If we know that the parameters are bounded away from these algebraic sets, then method such as MLE continues to produce root-n rate

  Otherwise it is worse, especially if the true model is near singularity

  - in practice, want to test if the true parameters are a singular point
  - in theory, we want to know the actual rate of estimation for singular points, necessitating the need to look into deep structure of singularities

- Singularities of skewnormal mixtures lie in affine varieties
- If we know that the parameters are bounded away from these algebraic sets, then method such as MLE continues to produce root-n rate

  Otherwise it is worse, especially if the true model is near singularity

  - in practice, want to test if the true parameters are a singular point
  - in theory, we want to know the actual rate of estimation for singular points, necessitating the need to look into deep structure of singularities
- We'll show that the singularity structure is very rich and go beyond the singularities of Fisher information
  - introduce singularity levels, which provide multi-level partitions of parameter space
- There is a consequence: the "more" singular the parameter values, the worse the MLE and minimax rates of estimation, $n^{-1/2}$ to $n^{-1/4}$ to $n^{-1/8}$, ad infinitum

# General theory

# General theory

behavior of likelihood function in the neighborhood of model parameters

# From parameter space to space of mixing measure $G$

The map $(\mathbf{p}, \boldsymbol{\eta}) \mapsto G(\mathbf{p}, \boldsymbol{\eta}) = \sum_i p_i \delta_{\eta_i}$ is many-to-one, e.g.

$$\frac{1}{2}\delta_0 + \frac{1}{2}\delta_1 = \frac{1}{2}\delta_0 + \frac{1}{3}\delta_1 + \frac{1}{6}\delta_1$$

Say $(\mathbf{p}, \boldsymbol{\eta})$ and $(\mathbf{p'}, \boldsymbol{\eta'})$ are equivalent if the corresponding mixing measures are equal, e.g.,

$$[\mathbf{p}, \boldsymbol{\theta}] = [(1/2, 1/2), (0, 1)] \equiv [(1/2, 1/3, 1/6), (0, 1, 1)]$$

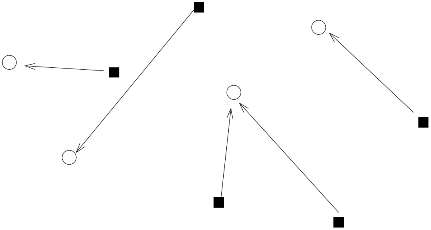
$$G = \sum p_i \delta_{\boldsymbol{\eta}_i} \;\mapsto\; p_G(\cdot) = \sum p_i f(\cdot|\boldsymbol{\eta}_i)$$

Wasserstein space of measures/ optimal transport distance

## Optimal transportation problem (Monge-Kantorovich)

how to transport good products from a collection of producers to a collection of consumers located in a common space

how to move the mass from one distribution to another?



squares: locations of producers; circles: locations of consumers

Optimal transport/Wasserstein distance is the optimal cost of transportation of mass
from — "production distribution" — to — "consumption distribution"

# Wasserstein distance

Let $G, G'$ be two prob. measures on $\mathbb{R}^d$, $r \geq 1$.
A coupling $\kappa$ of $G, G'$ is a joint dist on $\mathbb{R}^d \times \mathbb{R}^d$ which induces marginals $G, G'$.
$\kappa$ also called a "transportation plan".

The $r$-th order Wasserstein distance, denoted by $W_r$, is given by

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

$$G = \sum p_i \delta_{\boldsymbol{\eta}_i} \;\mapsto\; p_G(\cdot) = \sum p_i f(\cdot|\boldsymbol{\eta}_i)$$

# Behavior of likelihood in a Wasserstein neighborhood

As $G \to G_0$ in Wasserstein metric, apply Taylor expansion up to the $r$-th order:

$$p_G(x) - p_{G_0}(x) = \sum_{i=1}^{k_0} \sum_{j=1}^{s_i} p_{ij} \sum_{|\kappa|=1}^{r} \frac{(\Delta\eta_{ij})^{\kappa}}{\kappa!} \frac{\partial^{|\kappa|} f}{\partial \eta^{\kappa}}(x|\eta_i^0) + \sum_{i=1}^{k_0} \Delta p_{i.} f(x|\eta_i^0) + R_r(x),$$

where $R_r(x)$ is the Taylor remainder and $R_r(x)/W_r^r(G, G_0) \to 0$.

We have seen examples where the partial derivatives up to the first order are not linearly independent (for gamma kernel and skewnormal kernel)

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

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

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

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

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

$$\frac{\partial^2 f(x|\eta)}{\partial \theta^2} - 2\frac{\partial f(x|\eta\ )}{\partial v} + \frac{m^3 + m}{v}\frac{\partial f(x|\eta)}{\partial m} = 0.$$

## $r$-canonical form

Fix $r \geq 1$. For some sequence of $G_n \in \mathcal{G}$, such that $W_r(G_n, G_0) \to 0$,

$$\frac{p_{G_n}(x) - p_{G_0}(x)}{W_r^r(G_n, G_0)} = \sum_{l=1}^{L_r} \left( \frac{\xi_l(G_0, \Delta G_n)}{W_r^r(G, G_0)} \right) H_l(x|G_0) + o(1),$$

where

- $H_l(x|G_0)$ are *linearly independent* functions
- coefficients $\xi_l(G_0, \Delta G_n)/W_r^r(G_n, G_0)$ are ratio of two semi-polynomials of the parameter perturbation of $G_n$ around $G_0$

$H_l$ may be obtained by reducing partial derivatives to linearly independent ones
For Gamma kernel $f$,

$$\frac{\partial f(x|\eta_j^0)}{\partial m} = -\sum_{j=1}^{k} \frac{\alpha_{1j}}{\alpha_{4k}} f(x|\eta_j^0) + \frac{\alpha_{2j}}{\alpha_{4k}} \frac{\partial f(x|\eta_j^0)}{\partial \theta} + \frac{\alpha_{3j}}{\alpha_{4k}} \frac{\partial f(x|\eta_j^0)}{\partial v} - \sum_{j=1}^{k-1} \frac{\alpha_{4j}}{\alpha_{4k}} \frac{\partial f(x|\eta_j^0)}{\partial m}$$

For Gaussian kernel $f(x|\eta) = f(x|\theta, v)$ all partial derivatives wrt both $\theta$ and $v$ can be eliminated via the following reduction: for any $\kappa_1, \kappa_2 \in \mathbb{N}$, for any $j = 1, \ldots, k_0$,

$$\frac{\partial^{\kappa_1+\kappa_2} f(x|\eta_j^0)}{\partial \theta^{\kappa_1} v^{\kappa_2}} = \frac{1}{2^{\kappa_2}} \frac{\partial^{\kappa_1+2\kappa_2} f(x|\eta_j^0)}{\partial \theta^{\kappa_1+2\kappa_2}}.$$

Thus, this reduction is valid for all parameter values $(\mathbf{p}, \boldsymbol{\eta})$, and $r$-canonical forms for all orders.

For skewnormal kernel $f(x|\eta) = f(x|\theta, v, m)$ for any $j = 1, \ldots, k_0$, any $\eta = \eta_j^0 = (\theta_j^0, v_j^0, m_j^0)$,

$$\frac{\partial^2 f(x|\eta)}{\partial \theta^2} = 2\frac{\partial f(x|\eta)}{\partial v} - \frac{m^3 + m}{v}\frac{\partial f(x|\eta)}{\partial m}.$$

For higher order partial derivatives:

$$\begin{aligned}
\frac{\partial^3 f}{\partial \theta^3} &= 2\frac{\partial^2 f}{\partial \theta \partial v} - \frac{m^3 + m}{v}\frac{\partial^2 f}{\partial \theta \partial m}, \\
\frac{\partial^3 f}{\partial \theta^2 \partial v} &= 2\frac{\partial^2 f}{\partial v^2} + \frac{m^3 + m}{v^2}\frac{\partial f}{\partial m} - \frac{m^3 + m}{v}\frac{\partial^2 f}{\partial v \partial m}, \\
\frac{\partial^3 f}{\partial \theta^2 \partial m} &= 2\frac{\partial^2 f}{\partial v \partial m} - \frac{3m^2 + 1}{v}\frac{\partial f}{\partial m} - \frac{m^3 + m}{v}\frac{\partial^2 f}{\partial m^2}.
\end{aligned}$$

# $r$-singularity

Let $\mathcal{G}$ be a space of mixing measure.

## Definition

For each finite $r \geq 1$, say $G_0$ is $r$-singular relative to $\mathcal{G}$ if $G_0$ admits a $r$-canonical form for some sequence of $G \in \mathcal{G}$, according to which $W_r(G, G_0) \to 0$, coefficients $\xi_l^{(r)}(G)/W_r^r(G, G_0) \to 0$ for all $l = 1, \ldots, L_r$.

# $r$-singularity

Let $\mathcal{G}$ be a space of mixing measure.

**Definition**

For each finite $r \geq 1$, say $G_0$ is $r$-singular relative to $\mathcal{G}$ if $G_0$ admits a $r$-canonical form for some sequence of $G \in \mathcal{G}$, according to which $W_r(G, G_0) \to 0$, coefficients $\xi_l^{(r)}(G)/W_r^r(G, G_0) \to 0$ for all $l = 1, \ldots, L_r$.

**Lemma**

*(a) The notion of r-singularity is independent of the specific r-form. That is, the existence of the sequence G in the definition holds for all r-canonical forms once it holds for at least one of them.*

# $r$-singularity

Let $\mathcal{G}$ be a space of mixing measure.

**Definition**

For each finite $r \geq 1$, say $G_0$ is $r$-singular relative to $\mathcal{G}$ if $G_0$ admits a $r$-canonical form for some sequence of $G \in \mathcal{G}$, according to which $W_r(G, G_0) \to 0$, coefficients $\xi_l^{(r)}(G)/W_r^r(G, G_0) \to 0$ for all $l = 1, \ldots, L_r$.

**Lemma**

*(a) The notion of $r$-singularity is independent of the specific $r$-form. That is, the existence of the sequence $G$ in the definition holds for all $r$-canonical forms once it holds for at least one of them.*
*(b) If $G_0$ is $r$-singular for some $r > 1$, then $G_0$ is $(r-1)$-singular.*

## Definition

The singularity level of $G_0$ relative to ambient space $\mathcal{G}$, denoted by $\ell(G_0|\mathcal{G})$, is

- 0, if $G_0$ is not $r$-singular for any $r \geq 1$;
- $\infty$, if $G_0$ is $r$-singular for all $r \geq 1$;
- otherwise, the largest natural number $r \geq 1$ for which $G_0$ is $r$-singular.

## Theorem

If $\ell(G_0|\mathcal{G}) = r$, then for any $G \in \mathcal{G}$ subject to a compactness condition

$$V(p_G, p_{G_0}) \gtrsim W_{r+1}^{r+1}(G, G_0).$$

So, singularity level $r$ implies that MLE has rate

$$n^{-\frac{1}{2(r+1)}}$$

Under additional regularity condition, this is also a local minimax lower bound for estimating $G_0$

# Role of ambient spaces

If $\mathcal{G}_1 \subset \mathcal{G}_2$ then

$$\ell(G_0|\mathcal{G}_1) \leq \ell(G_0|\mathcal{G}_2)$$

For location-scale Gaussian e-mixtures

$$\ell(G_0|\mathcal{E}_{k_0}) = 0.$$

# Role of ambient spaces

If $\mathcal{G}_1 \subset \mathcal{G}_2$ then

$$\ell(G_0|\mathcal{G}_1) \leq \ell(G_0|\mathcal{G}_2)$$

For location-scale Gaussian e-mixtures

$$\ell(G_0|\mathcal{E}_{k_0}) = 0.$$

For any o-mixtures: $G_0$ has $k_0$ support points, but we consider the space $\mathcal{O}_k$ of all measures with at most $k > k_0$ support points. Then,

$$\ell(G_0|\mathcal{O}_k) \geq 1.$$

In fact, for location-scale Gaussian o-mixtures, we can show

$$\ell(G_0|\mathcal{O}_k) \geq 3.$$

# Singularities in o-mixtures

$G_0$ has $k_0$ support points,
$\mathcal{G} := \mathcal{O}_k$, are space of mixing measures having at most $k > k_0$ support points

# When do we have $\ell(G_0|\mathcal{O}_k) = 1$?

**Definition** [Chen, 1995; Nguyen, 2013]

$G$ is non-singular in a o-mixture model with at most $k$ components if

$$\left\{ f(x|\theta_i), \frac{\partial f}{\partial \theta}(x|\theta_i), \frac{\partial^2 f}{\partial \theta^2}(x|\theta_i), |i = 1, \ldots, k \right\}$$

are linearly independent functions of $x$

**Theorem** [Nguyen, 2013; Chen, 1995]

Under non-singularity of $G_0$, and compactness conditions on parameter space, there holds

$$V(p_G, p_{G_0}) \gtrsim W_2^2(G, G_0).$$

This is a corollary of our theory, due to the fact that one can establish

$$\ell(G_0|\mathcal{O}_k) = 1.$$

Since MLE or Bayes estimators yield root-$n$ *density estimation* rate, it implies that

$$W_2(\hat{G}_n, G_0) = O_p(n^{-1/4})$$

where $G_0$ denotes true mixing distribution, $G_n$ estimate from an $n$-iid sample

Since MLE or Bayes estimators yield root-$n$ *density estimation* rate, it implies that

$$W_2(\hat{G}_n, G_0) = O_p(n^{-1/4})$$

where $G_0$ denotes true mixing distribution, $G_n$ estimate from an $n$-iid sample

This result is applicable to location Gaussian o-mixture, scale Gaussian o-mixture, but not applicable to

- location-scale Gaussian o-mixture
- skewnormal o-mixture

# Location-scale Gaussian o-mixture

- given $n$-iid sample from a location-scale Gaussian mixture with mixing measure $G_0$ (which has $k_0$ components)
- fit the data with a mixture model with $k > k_0$ components
- $\ell(G_0|\mathcal{O}_k)$ is determined by $(k - k_0)$ — specifically, by the following system of polynomial equations:

# Location-scale Gaussian o-mixture

- given $n$-iid sample from a location-scale Gaussian mixture with mixing measure $G_0$ (which has $k_0$ components)
- fit the data with a mixture model with $k > k_0$ components
- $\ell(G_0|\mathcal{O}_k)$ is determined by $(k - k_0)$ — specifically, by the following system of polynomial equations:

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

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

# Location-scale Gaussian o-mixture

- given $n$-iid sample from a location-scale Gaussian mixture with mixing measure $G_0$ (which has $k_0$ components)
- fit the data with a mixture model with $k > k_0$ components
- $\ell(G_0|\mathcal{O}_k)$ is determined by $(k - k_0)$ — specifically, by the following system of polynomial equations:

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

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

- let $\bar{r} \geq 1$ *minimum* value of $r \geq 1$ such that the above equations do *not* have any non-trivial real-valued solution. A solution is considered non-trivial if all $c_j$s are non-zeros, and at least one of the $a_j$s is non-zero.

## Theorem [Singularity in localion-scale Gaussian o-mixtures]

$\ell(G_0|\mathcal{O}_k) = \bar{r} - 1.$

## Corrolary [Location-scale Gaussian finite mixtures]

convergence rate of mixing measure $G$ by either MLE or Bayesian estimation is $(\log n/n)^{1/2\bar{r}}$, under both $W_{\bar{r}}$ and $W_1$.

[Ho & Nguyen (2016)]

# More on system of $r$ polynomial equations (1):

let us consider the case $k = k_0 + 1$, and let $r = 3$, then we have

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

a non-trivial solution exists, by choosing $c_2 = c_1 \neq 0$, $b_1 = b_2 = -1/2$, and $a_1 = 1, a_2 = -1$.
Hence, $\overline{r} \geq 4$.

# More on system of $r$ polynomial equations (1):

let us consider the case $k = k_0 + 1$, and let $r = 3$, then we have

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

a non-trivial solution exists, by choosing $c_2 = c_1 \neq 0$, $b_1 = b_2 = -1/2$, and $a_1 = 1, a_2 = -1$.
Hence, $\bar{r} \geq 4$.

For $r = 4$, the system consists of the three equations above, plus

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

This system has no non-trivial solution. So, for $k = k_0 + 1$, we have $\bar{r} = 4$.

Using the Groebner bases method, we can figure out the zeros of a system of real polynomial equations. So,

(1) [Overfitting by one] if $k - k_0 = 1$, then $\ell(G_0|\mathcal{O}_k) = 3$. So, MLE/Bayes estimation rate is $n^{-1/8}$.

(2) [Overfitting by two] if $k - k_0 = 2$, then $\ell(G_0|\mathcal{O}_k) = 5$, so the rate is $n^{-1/12}$.

(3) [Overfitting by three] if $k - k_0 = 3$, then $\ell(G_0|\mathcal{O}_k) \geq 6$. So, the rate is not better than $n^{-1/14}$.

Using the Groebner bases method, we can figure out the zeros of a system of real polynomial equations. So,

(1) [Overfitting by one] if $k - k_0 = 1$, then $\ell(G_0|\mathcal{O}_k) = 3$. So, MLE/Bayes estimation rate is $n^{-1/8}$.

(2) [Overfitting by two] if $k - k_0 = 2$, then $\ell(G_0|\mathcal{O}_k) = 5$, so the rate is $n^{-1/12}$.

(3) [Overfitting by three] if $k - k_0 = 3$, then $\ell(G_0|\mathcal{O}_k) \geq 6$. So, the rate is not better than $n^{-1/14}$.

### Lessons for Gaussian location-scale mixtures

- do not overfit

Using the Groebner bases method, we can figure out the zeros of a system of real polynomial equations. So,

(1) [Overfitting by one] if $k - k_0 = 1$, then $\ell(G_0|\mathcal{O}_k) = 3$. So, MLE/Bayes estimation rate is $n^{-1/8}$.

(2) [Overfitting by two] if $k - k_0 = 2$, then $\ell(G_0|\mathcal{O}_k) = 5$, so the rate is $n^{-1/12}$.

(3) [Overfitting by three] if $k - k_0 = 3$, then $\ell(G_0|\mathcal{O}_k) \geq 6$. So, the rate is not better than $n^{-1/14}$.

## Lessons for Gaussian location-scale mixtures

- do not overfit
- if you must, be conservative in allowing extra mixing components

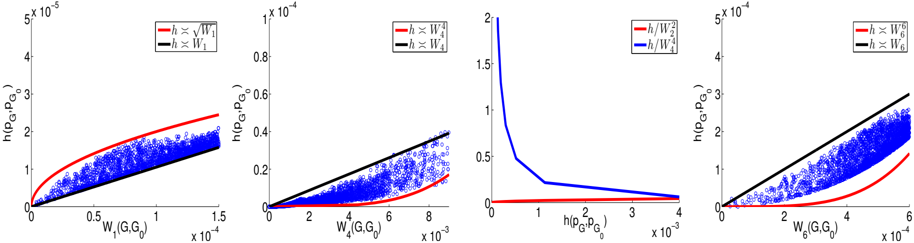



Figure : Location-scale Gaussian mixtures. From left to right: (1) Exact-fitted setting; (2) Over-fitted by one component; (3) Over-fitted by one component; (4) Over-fitted by two components.

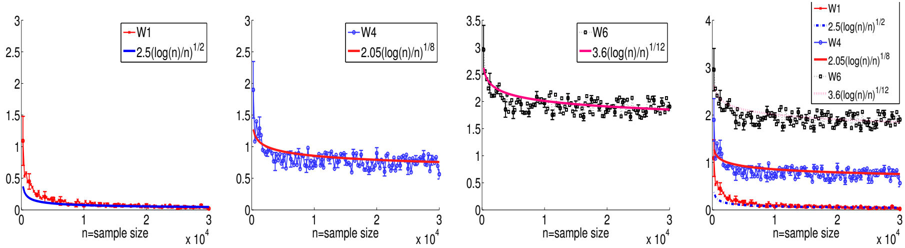



Figure : MLE rates for location-covariance mixtures of Gaussians. L to R: (1) Exact-fitted: $W_1 \asymp n^{-1/2}$. (2) Over-fitted by one: $W_4 \asymp n^{-1/8}$. (3) Over-fitted by two: $W_6 \asymp n^{-1/12}$.

# The direct link to algebraic geometry

Recall $r$-canonical form

$$\frac{p_{G_n}(x) - p_{G_0}(x)}{W_r^r(G_n, G_0)} = \sum_{l=1}^{L_r} \left( \frac{\xi_l(G_0, \Delta G_n)}{W_r^r(G, G_0)} \right) H_l(x) + o(1),$$

where

- coefficients $\xi_l(G_0, \Delta G_n)/W_r^r(G_n, G_0)$ are the ratio of two semi-polynomials of the parameter perturbation of $G_n$ around $G_0$
- as $G_n \to G_0$, the collection of these ratios tend to a system of real polynomial equations

# Singularities in e-mixtures of skewnormals

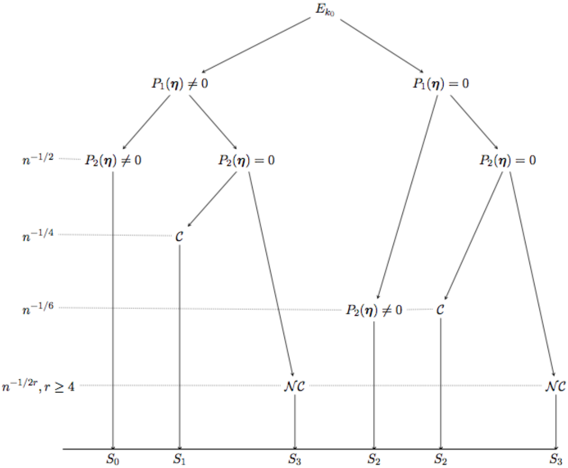
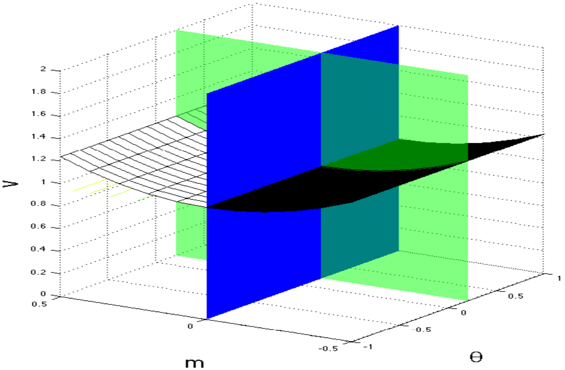



Figure 2: The singularity levels of $G_0 \in \mathcal{E}_{k_0}$. "$\mathcal{C}$" stands for conformant and "$\mathcal{NC}$" stands for nonconformant. The leaf is arranged horizontally according to its level of singularity and vertically according to its influence on the corresponding settings of $G_0$.




$P_1(\boldsymbol{\eta}) = \prod_{j=1}^{k} m_j.$

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

# Partition of subset NC, without Gaussian components

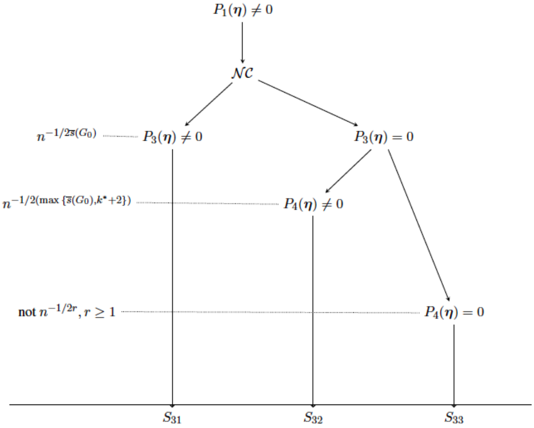



Figure 3: The level of singularity structure of $G_0$ relating to the nonconformant without symmetry e-mixtures setting. Here, "$\mathcal{NC}$" stands for nonconformant.The leaf is arranged horizontally according to its level of singularity and vertically according to its influence on the corresponding setting. $k^*$ is the maximum length of nonconformant homologous sets with C.1 singularity in $G_0$ respectively. Finally, the term $\overline{s}(G_0)$ is defined as in (27).

# Partition of subset NC, with some Gaussian components

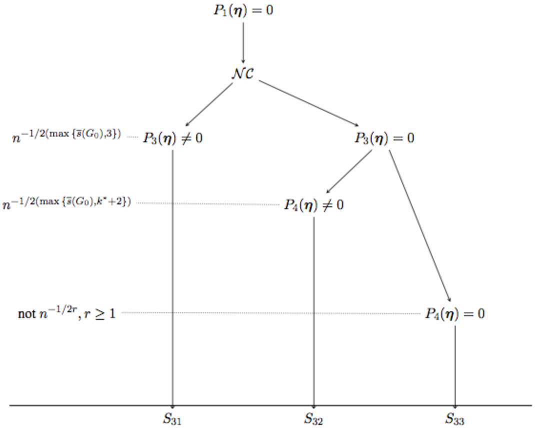



Figure 4: The level of singularity structure of $G_0$ relating to the nonconformant with symmetry e-mixtures setting. Here, "$\mathcal{NC}$" stands for nonconformant.The leaf is arranged horizontally according to its level of singularity and vertically according to its influence on the corresponding setting. $k^*$ is the maximum length of nonconformant homologous sets with C.1 singularity in $G_0$ respectively. The term $\overline{s}(G_0)$ is defined as in (27).

# Singularities in o-mixtures of skewnormal mixtures

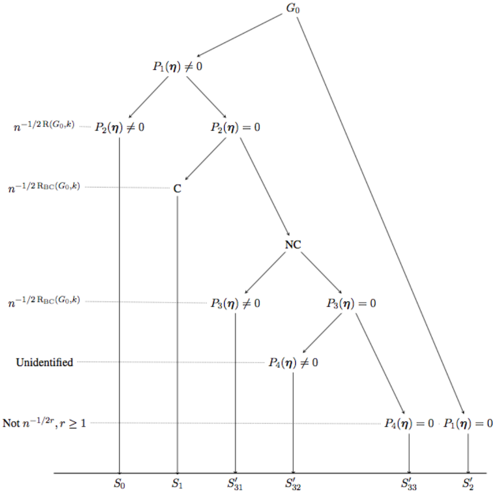



Figure 6: The singularity level of $G_0 \in \mathcal{E}_{k_0}$ under the o-mixtures setting. "C" stands for conformant and "NC" stands for nonconformant. The leaf is arranged horizontally according to its level of singularity and vertically according to its influence on the corresponding setting of $G_0$. The term $\mathrm{R}(G_0, k)$ is defined as in (43) and $\mathrm{R_{BC}}(G_0, k)$ is defined as in (54). Note that, for the setting $P_2(\boldsymbol{\eta}) = 0$, we assume $G_0$ has no generic components for the simplicity of presentation.

# Summary

- Singularities are common in mixture models, including finite mixtures
- Beyond the singular points of Fisher information
- They are organized into levels, which subdivide the parameter space into multi-level partitions, each of which allow different minimax and MLE convergence rate
- Now that we know what the singularities are (mostly), how to go about improving the estimation algorithm both statistically and computationally?

For details, see

- *Convergence of latent mixing measures in finite and infinite mixture models.* (Ann. Statist., 2013)
- *Convergence rates of parameter estimation in some weakly identifiable models* (with N. Ho, Ann. Statist., 2016)
- *Singularity structures and parameter estimation in finite mixture models* (manuscript to be submitted)