On the NonParametric Maximum Likelihood Estimator (NPMLE) for Gaussian location mixture densities and applications

BIRS Workshop: New and evolving roles of shrinkage in large-scale prediction and inference (09 April, 2019)

Aditya Guntuboyina

Department of Statistics, University of California, Berkeley

Outline

1. The NPMLE for Gaussian Location Mixture Densities
2. Empirical Bayes Estimation of Normal Means

The above is joint work (https://arxiv.org/pdf/1712.02009.pdf) with Sujayam Saha.

3. Empirical Bayes Testing of Normal Means (ongoing work with Yuting Wei and Cun-hui Zhang)
4. Mixture of Linear Regressions (ongoing work with Hansheng Jiang)

A Basic Model

We observe data $Y_1, \ldots, Y_n \in \mathbb{R}^d$ drawn from the model:

$$Y_i = \theta_i + Z_i \quad \text{with } Z_1, \ldots, Z_n \overset{\text{i.i.d}}{\sim} N_d(0, I_d).$$

$\theta_1, \ldots, \theta_n \in \mathbb{R}^d$ are unknown and we additionally assume:

$$\theta_1, \ldots, \theta_n \overset{\text{i.i.d}}{\sim} G^*$$

for an unknown probability measure G^* on $\mathbb{R}^d$.

Examples: sparsity ($\mathbb{P}_{G^*}\{\theta_1 = 0\}$ large) and clustering (G^* discrete).

The NPMLE presents a very effective estimation strategy in this model.

The NPMLE (Kiefer and Wolfowitz 1956)

Under this model

$$Y_1, \dots, Y_n \overset{\text{i.i.d}}{\sim} f_{G^*} \quad \text{where } f_{G^*}(y) := \int_{\mathbb{R}^d} \phi_d(y-\theta) dG^*(\theta)$$

and ϕ_d is standard Gaussian density.

The NPMLE of G^* is defined by

$$\hat{G}_n \in \underset{G}{\mathrm{argmax}} \sum_{i=1}^{n} \log f_G(Y_i)$$

where the argmax is over all probability measures G on $\mathbb{R}^d$.

We argue that $f_{\hat{G}_n}$ (given by $y \mapsto \int_{\mathbb{R}^d} \phi_d(y-\theta) d\hat{G}_n(\theta)$) is, in general, a very good estimator for f_{G^*}.

Standard references for the NPMLE are the books by Bohning (1999) and Lindsay (1995).

Standard Facts about the NPMLE

$$\hat{G}_n \in \underset{G}{\mathrm{argmax}} \sum_{i=1}^{n} \log f_G(Y_i).$$

This is a convex optimization problem as the objective function is concave in G and the constraint set (all probability measures) is convex.

It is easy to show that $\hat{G}_n$ exists and $f_{\hat{G}_n}(X_1), \dots, f_{\hat{G}_n}(X_n)$ are unique (further uniqueness results are in Lindsay 1983).

Note that G is totally unconstrained here. Enforcing constraints (such as an upper bound on the number of atoms of G) might make the optimization non-convex and/or involve tuning parameters.

Our main point is that the (unconstrained) NPMLE typically adapts to the underlying structure in G^*.

Optimization Details

Even though the NPMLE is given by a convex optimization problem, the optimization is still infinite-dimensional.

Approximate algorithms exist: Bohning (1999), Lindsay (1995), Laird (1978), Lashkari and Golland (2007), Jiang and Zhang (2009), Koenker and Mizera (2015), Feng and Dicker (2015), Dicker and Zhao (2016).

These include Frank-Wolfe methods (VDM, VEM), EM algorithms and direct discretization i.e., fix $\theta_1, \dots, \theta_M \in \mathbb{R}^d$ for large M and solve:

$$\max_{p_1,\dots,p_M} \quad \left\{ \sum_{i=1}^{n} \log \left(\sum_{j=1}^{M} p_j \phi(Y_i - \theta_j) \right) : p_j \geq 0 \text{ and } \sum_{j=1}^{M} p_j = 1 \right\}.$$

Pictures: True Density f_{G^*} (left) and Estimated Density $f_{\hat{G}_n}$ (Right)

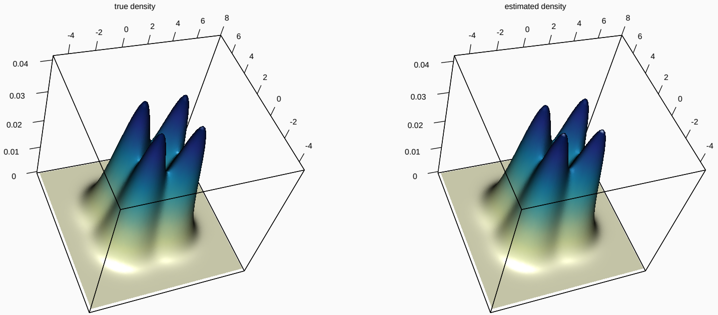


Figure 1: Sample Size is $n = 10000$. G^* is discrete which puts equal mass on the four points $(0,0),(3,0),(0,3),(3,3)$.

Pictures: True Density f_{G^*} (left) and Estimated Density $f_{\hat{G}_n}$ (Right)

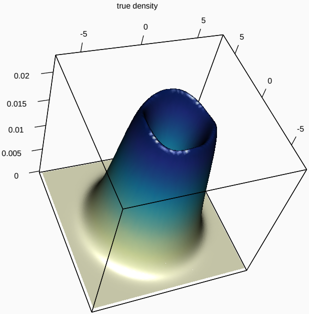


Figure 2: Sample Size is $n = 10000$. G^* is uniformly distributed on a circle of radius 3.

Pictures: True Density f_{G^*} (left) and Estimated Density $f_{\hat{G}_n}$ (Right)

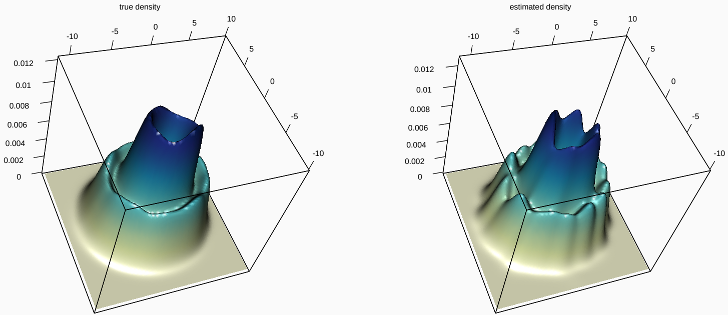


Figure 3: Sample Size is $n = 10000$. G^* is uniformly distributed on two concentric circles of radii 3 and 6.

Our Focus: Accuracy of $f_{\hat{G}_n}$ for f_{G^*}

Suppose $\hat{G}_n$ is an NPMLE. How accurate is $f_{\hat{G}_n}$ as an estimator of f_{G^*} when d is small and n is large?

We study accuracy via risk under squared Hellinger distance:

$$\mathfrak{H}^2(f_{\hat{G}_n}, f_{G^*}) := \int \left(\sqrt{f_{\hat{G}_n}} - \sqrt{f_{G^*}}\right)^2$$

and prove upper bounds for $\mathbb{E}\mathfrak{H}^2(f_{\hat{G}_n}, f_{G^*})$.

We argue that $f_{\hat{G}_n}$ is a very good estimator for f_{G^*} when G^* satisfies natural assumptions (such as being discrete).

Our work is heavily inspired by Zhang (2009) who studied this for $d = 1$ under certain assumptions on G^* (we relax the assumptions on G^* and extend to $d \geq 1$).

Result when G^* has compact support S

Suppose that G^* has compact support S.

Then (for $S^1 := S + B(0,1) = \cup_{x \in S} B(x,1)$),

$$\mathbb{E}\mathfrak{H}^2\left(f_{\hat{G}_n}, f_{G^*}\right) \le C_d \frac{\mathrm{Vol}(S^1)}{n} \left(\sqrt{\log n}\right)^{d+(4-d)_+}$$

for a positive constant C_d depending on d alone.

In words, the risk of the NPMLE is $\lesssim \mathrm{Vol}(S^1)/n$ (ignoring logarithmic factors). Extensions to non-compact support are possible.

Note that the NPMLE is completely tuning-free and does not use any knowledge of the support of G^*.

The results of Zhang (2009) correspond to $d = 1$ and $S = [-M, M]$.

Special case: when G^* is a discrete distribution

Suppose f_{G^*} is a discrete mixture with k components (equivalently, G^* is a discrete probability measure with k atoms). Then

$$\mathbb{E}\mathfrak{H}^2\left(f_{\hat{G}_n}, f_{G^*}\right) \leq C_d \left(\frac{k}{n}\right) \left(\sqrt{\log n}\right)^{d+(4-d)_+}.$$

for a positive constant C_d depending on d alone.

This result shows that the risk of $f_{\hat{G}_n}$ is k/n up to logarithmic factors in n.

This is remarkable because the NPMLE does not a priori know k.

Simple minimax lower bounds show that no estimator can estimate k-component Gaussian mixtures at a uniform rate that is better than k/n.

General Discrete Mixtures

Suppose we observe i.i.d $\tilde{Y}_1, \ldots, \tilde{Y}_n$ from $h^* := \sum_{j=1}^k w_j N_d(\mu_j, \Sigma_j)$.

Let $0 < \sigma_{\min}^2 \leq \min_{1 \leq j \leq k} \lambda_{\min}(\Sigma_j) \leq \max_{1 \leq j \leq k} \lambda_{\max}(\Sigma_j) =: \sigma_{\max}^2$ with $\sigma_{\min}^2$ known.

A natural strategy for estimating h^* would be to compute the NPMLE $f_{\hat{G}_n}$ based on $Y_i := \tilde{Y}_i / \sigma_{\min}$ and then use:

$$\hat{h}_n(y) := \sigma_{\min}^{-d} f_{\hat{G}_n}(\sigma_{\min}^{-1} y) \qquad \text{for } y \in \mathbb{R}^d.$$

This estimator requires knowledge of $\sigma_{\min}$.

Hellinger Accuracy of $\hat{h}_n$

With $\tau := \sqrt{(\sigma^2_{\max}/\sigma^2_{\min}) - 1}$, we have

$$\mathbb{E}\mathfrak{H}^2\left(\hat{h}_n, h^*\right) \leq C_d \left(\frac{k}{n}\right) (\max(1,\tau))^d \left(\sqrt{\log n}\right)^{d+(4-d)_+} .$$

This presents a method for estimating Gaussian mixture densities at nearly the k/n rate in squared Hellinger distance under an a priori lower bound on the eigenvalues of all the component covariance matrices.

No other information (such as the number of components) is required.

This result is comparable to the penalized model selection results of Maugis and Michel (2011) for fitting k-component mixtures and subsequently choosing k via model selection.

Comments

1. The constants C_d have bad dependence on d.

2. The logarithmic terms are possibly suboptimal.

3. Additional results and proof techniques can be found in the paper (https://arxiv.org/pdf/1712.02009.pdf).

Application 1: Empirical Bayes Estimation of Normal Means

Empirical Bayes Estimation of Normal Means

Consider the problem of estimating $\theta_1, \ldots, \theta_n$ in squared error loss from data $Y_1, \ldots, Y_n \in \mathbb{R}^d$ drawn from the model:

$$Y_i = \theta_i + Z_i \quad \text{with } Z_1, \ldots, Z_n \overset{\text{i.i.d}}{\sim} N_d(0, I_d).$$

A simple Bayesian approach to this problem starts by the model

$$\theta_1, \ldots, \theta_n \overset{\text{i.i.d}}{\sim} G^*$$

for a prior G^* on $\mathbb{R}^d$ and estimates each θ_i by the posterior mean

$$\hat{\theta}_i^* := \mathbb{E}(\theta_i | Y_i)$$

These estimates $\hat{\theta}_1^*, \ldots, \hat{\theta}_n^*$ crucially depend on G^* and can be approximated via the NPMLE.

NPMLE for estimating $\hat{\theta}_i^*, i = 1, \ldots, n$

To see how the NPMLE can estimate $\hat{\theta}_i^*$, note that by Tweedie's formula,

$$\hat{\theta}_i^* = Y_i + \frac{\nabla f_{G^*}(Y_i)}{f_{G^*}(Y_i)}.$$

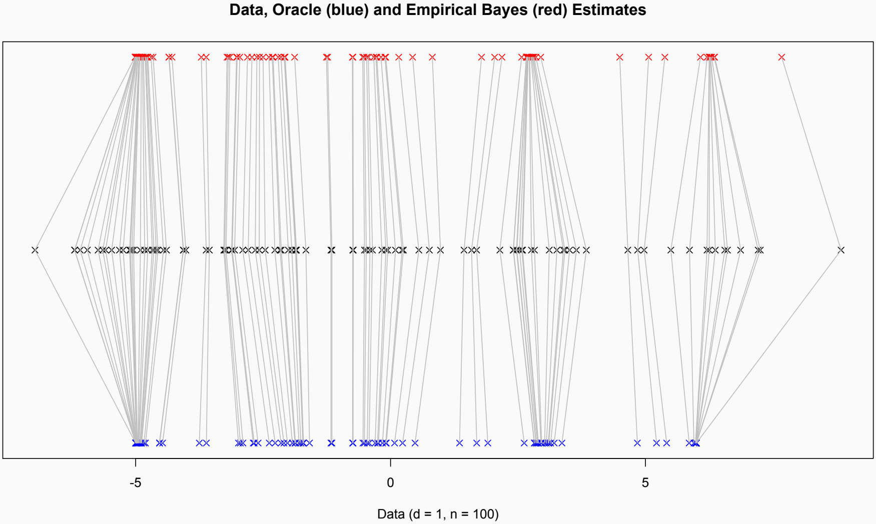
and this suggests the following estimator for $\hat{\theta}_i^*$ (where $\hat{G}_n$ is NPMLE):

$$\hat{\theta}_i := Y_i + \frac{\nabla f_{\hat{G}_n}(Y_i)}{f_{\hat{G}_n}(Y_i)}$$

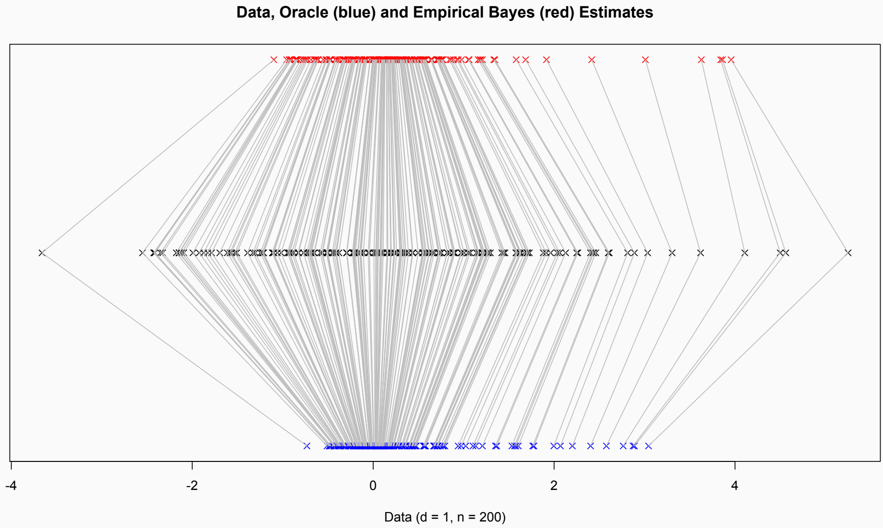
This is the General Maximum Likelihood Empirical Bayes (GMLEB) estimator of Jiang and Zhang (2009) who studied it in $d = 1$ for sparse normal mean estimation.

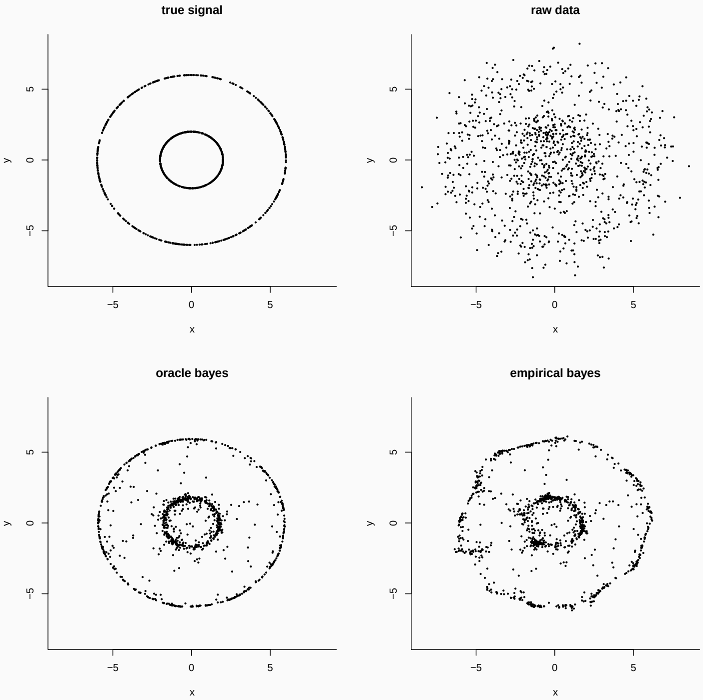
This estimator is tuning-free and provides excellent shrinkage (although we assumed known noise level) .

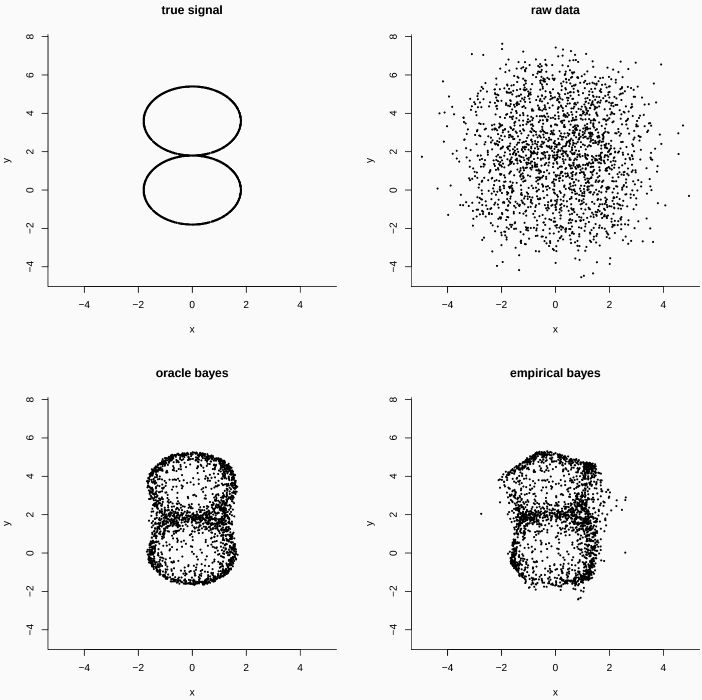
Shrinkage Illustration ($G^* = \textbf{Unif}\{-5, -2, 0, 3, 6\}$)

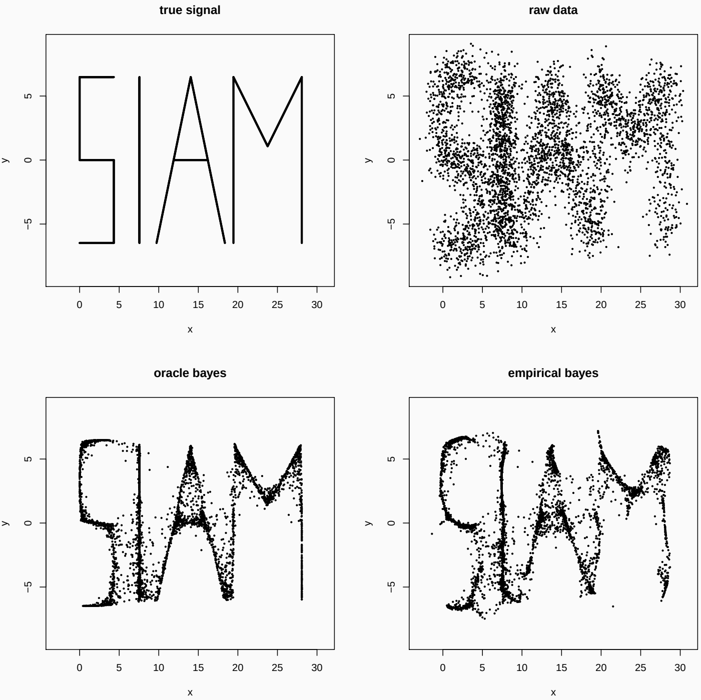


Shrinkage Illustration ($G^* = 0.9N(0, 0.25) + 0.1N(2.5, 0.25)$)




true signal
raw data
oracle bayes
empirical bayes
x
y
−5
0
5


true signal
raw data
oracle bayes
empirical bayes
x
y
−4
−2
0
2
4
6
8



true signal
raw data
oracle bayes
empirical bayes
x
y
0
5
10
15
20
25
30
−5
0
5

Accuracy of $\hat{\theta}_i$ for estimating $\hat{\theta}_i^*$

Accuracy can be measured via $\mathbb{E}T(\hat{\theta}, \hat{\theta}^*)$ where

$$T(\hat{\theta}, \hat{\theta}^*) := \frac{1}{n}\sum_{i=1}^{n} \|\hat{\theta}_i - \hat{\theta}_i^*\|^2$$

and the expectation is over $Y_1, \dots, Y_n \overset{\text{i.i.d}}{\sim} f_{G^*}$.

We proved $\mathbb{E}T(\hat{\theta}, \hat{\theta}^*) \to 0$ as $n \to \infty$ under some assumptions on G^*.

Previous studies of estimating $\hat{\theta}^*$ have focused on $d = 1$ and imposed sparsity assumptions on G^* (Johnstone and Silverman 2004, Brown and Greenshtein 2007, Jiang and Zhang 2009, Donoho and Reeves 2013).

We study $d \geq 1$ under more general assumptions on G^* (results and proofs inspired by Jiang and Zhang 2009) and our loss function $T(\hat{\theta}, \hat{\theta}^*)$ is different from the error measures in these other papers.

Accuracy Result when G^* has compact support S

If G^* has compact support S (with $S^1 := S + B(0,1) = \cup_{x \in S} B(x,1)$)

$$\mathbb{E} T(\hat{\theta}, \hat{\theta}^*) \le C_d \frac{\mathrm{Vol}(S^1)}{n} \left(\sqrt{\log n}\right)^{d + (4-d)_+} (\log n)^{\max(d,3)}$$

for a constant C_d. This rate is $\mathrm{Vol}(S^1)/n$ up to log factors (note that $\hat{\theta}$ uses no knowledge of S).

Special case is clustering where G^* is supported on a set of size k:

$$\mathbb{E} T(\hat{\theta}, \hat{\theta}^*) \le C_d \left(\frac{k}{n}\right) \left(\sqrt{\log n}\right)^{d + (4-d)_+} (\log n)^{\max(d,3)}.$$

so we get the k/n rate (up to log factors) without prior knowledge of k.

Such results do not appear to exist for other clustering algorithms based on convex optimization such as convex clustering (Chen et al 2015, Tan and Witten 2015, Radchenko and Mukherjee 2014 etc.).

Application 2: Empirical Bayes Hypothesis Testing of Normal Means

(on-going work with Yuting Wei and Cun-Hui Zhang)

Hypothesis Testing of Normal Means

Consider data $Y_1, \ldots, Y_n \in \mathbb{R}$ drawn from the model:

$$Y_i = \theta_i + Z_i \quad \text{with } Z_1, \ldots, Z_n \overset{\text{i.i.d}}{\sim} N(0, 1).$$

with unknown $\theta_1, \ldots, \theta_n$. We want to test the n hypotheses

$$H_{0i} : \theta_i \leq 0 \quad \text{against} \quad H_{1i} : \theta_i > 0 \qquad \text{for } i = 1, \ldots, n.$$

under the natural loss function:

$$L(\theta, a) := \frac{1}{n} \sum_{i=1}^{n} \{(-\theta_i)_+ I\{a_i = 1\} + (\theta_i)_+ I\{a_i = 0\}\}$$

The risk of $\delta(Y_1, \ldots, Y_n) = (\delta_1(Y_1, \ldots, Y_n), \ldots, \delta_n(Y_1, \ldots, Y_n))$ is

$$R(\theta, \delta) = \mathbb{E}_\theta L(\theta, \delta(Y_1, \ldots, Y_n)).$$

This problem has been studied by Karunamuni (1996), Liang (2000) and Li and Gupta (2006).

Bayesian Approach

Assume additionally

$$\theta_1, \ldots, \theta_n \overset{\text{i.i.d}}{\sim} G^*.$$

and observe then that the best test is

$$\delta_i^*(Y_1, \ldots, Y_n) = \delta_i^*(Y_i) = I\left\{\mathbb{E}(\theta_i | Y_i) > 0\right\} = I\left\{Y_i + \frac{f'_{G^*}(Y_i)}{f_{G^*}(Y_i)} > 0\right\}.$$

This test crucially depends on G^* but only through f_{G^*}.

It is thus natural to approximate this Bayes test by replacing f_{G^*} by $f_{\hat{G}_n}$.

Empirical Bayes Test

We are thus led to the Empirical Bayes Test:

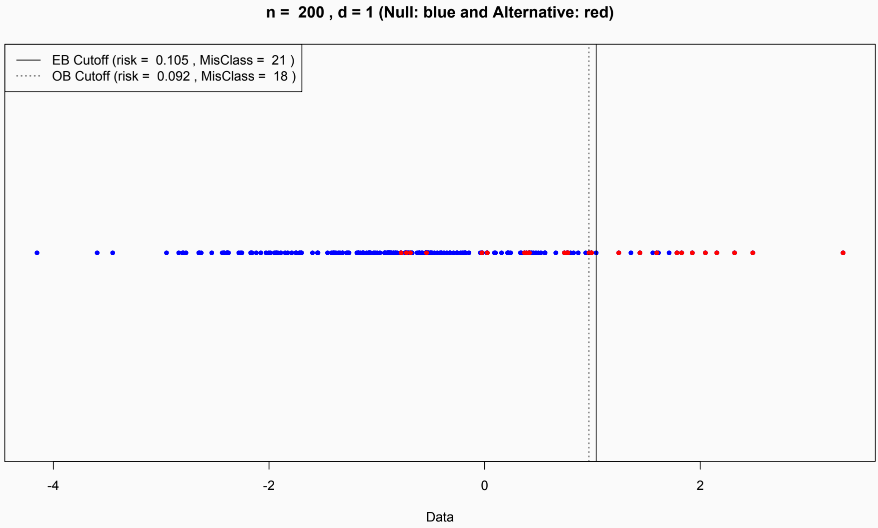
$$\hat{\delta}_i(Y_1, \ldots, Y_n) = I\left\{ Y_i + \frac{f'_{\hat{G}_n}(Y_i)}{f_{\hat{G}_n}(Y_i)} > 0 \right\}$$

for $i = 1, \ldots, n$.

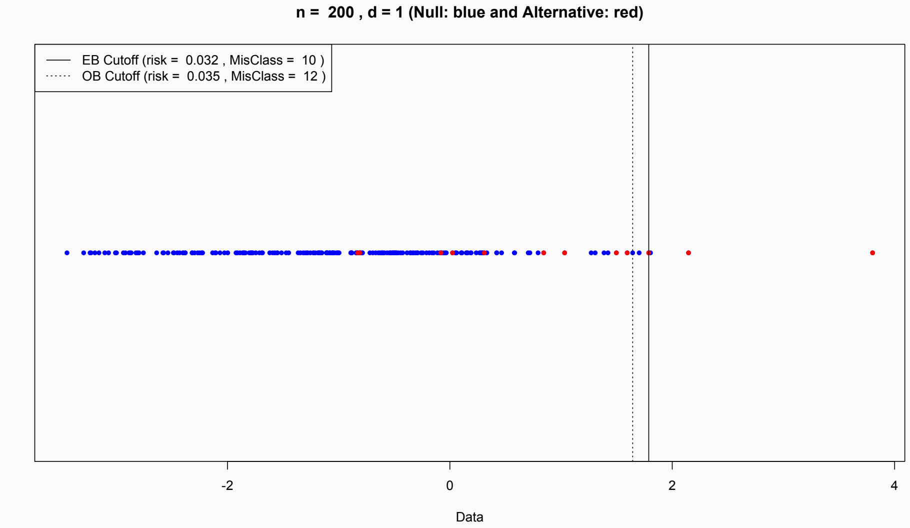
This test approximates the Bayes test δ^* very well in simulations.

Existing tests for this problem are based on kernel approaches (and involve certain tuning parameters).

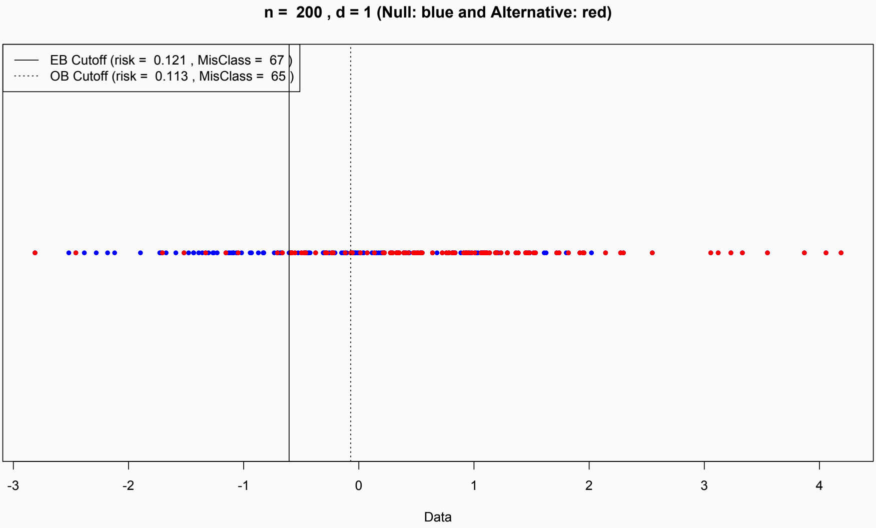
EB Testing One ($G^* = 0.9N(-1, 0.25) + 0.1N(1.5, 0.25)$)



EB Testing Two ($G^* = 0.98N(-1, 0.25) + 0.02N(1.5, 0.25)$)



EB Testing Three ($G^* = 0.9N(0, 0.25) + 0.1N(2.5, 0.25)$)



Theoretical Results (in progress)

We are currently studying the regret of the Empirical Bayes test where

$$\text{Regret} := \text{Bayes risk of } \hat{\delta}_n - \text{Bayes risk of } \delta^*$$

Liang (2000) constructed a test whose regret is bounded by $(\log n)^{1.5}/n$ under certain boundedness assumptions on G^* and Li and Gupta (2006) proved that $(\log n)^{1.5}/n$ is minimax optimal under those assumptions.

We are currently working on proving a version of this result for the Empirical Bayes test as well as exploring its behavior under other assumptions on G^*.

Mixture of Linear Regressions (on-going work with Hansheng Jiang)

The Mixture of Homoscedastic Linear Regressions Model

We observe data $(x_1, Y_1), \ldots, (x_n, Y_n)$ with $x_i \in \mathbb{R}^p$ and $Y_i \in \mathbb{R}$ from:

$$Y_i = x_i^T \beta_i + Z_i \qquad \text{with } Z_1, \ldots, Z_n \overset{\text{i.i.d}}{\sim} N(0,1)$$

and assume that $x_1, \ldots, x_n$ are deterministic (fixed-design setting).

$\beta_1, \ldots, \beta_n \in \mathbb{R}^p$ are unknown and we additionally assume:

$$\beta_1, \ldots, \beta_n \overset{\text{i.i.d}}{\sim} G^*.$$

for an unknown probability measure G^* on $\mathbb{R}^p$.

This is the Mixture of Homoscedastic Linear Regressions Model (see e.g., Jordan and Jacobs 1994, Faria and Soromenho 2010; many recent papers dealing with the high-dimensional setting).

NPMLE in this model

Under this model, $Y_1, \ldots, Y_n$ are independent with $Y_i \sim f_{G^*}^{x_i}$ where

$$f_{G^*}^{x_i}(y) := \int \phi\left(y - x_i^T \beta\right) dG^*(\beta)$$

The NPMLE of G^* is then defined by

$$\hat{G}_n := \underset{G}{\mathrm{argmax}} \left(\sum_{i=1}^{n} \log f_G^{x_i}(y_i) \right)$$

where the argmax is over all probability measures on $\mathbb{R}^p$.

As before, this is a convex optimization problem as the objective function is concave in G and the constraint set is convex.

$\hat{G}_n$ always exists and the n values $f_{\hat{G}_n}^{x_1}(y_1), \ldots, f_{\hat{G}_n}^{x_n}(y_n)$ are unique.

The optimization can be approximately solved by Frank-Wolfe.

Frank-Wolfe for NPMLE in MLR

$\mathfrak{f}_G := (f_G^{x_1}(y_1), \dots, f_G^{x_n}(y_n))$ and $\mathfrak{f}_\beta := (\phi(y_1 - x_1^T\beta), \dots, \phi(y_n - x_n^T\beta))$.

We need to solve $\mathrm{argmax}_G\, L(\mathfrak{f}_G)$ where $L(\mathfrak{f}) := \sum_{i=1}^n \log \mathfrak{f}(i)$.

Initialize: $\beta^{(0)}$ by linear regression and $G^{(0)} := \delta_{\beta^{(0)}}$.

Approximate $L(\mathfrak{f}_G) \approx L(\mathfrak{f}_{G^{(0)}}) + \langle \mathfrak{f}_G - \mathfrak{f}_{G_0}, \nabla L(\mathfrak{f}_{G^{(0)}})\rangle$ and solve

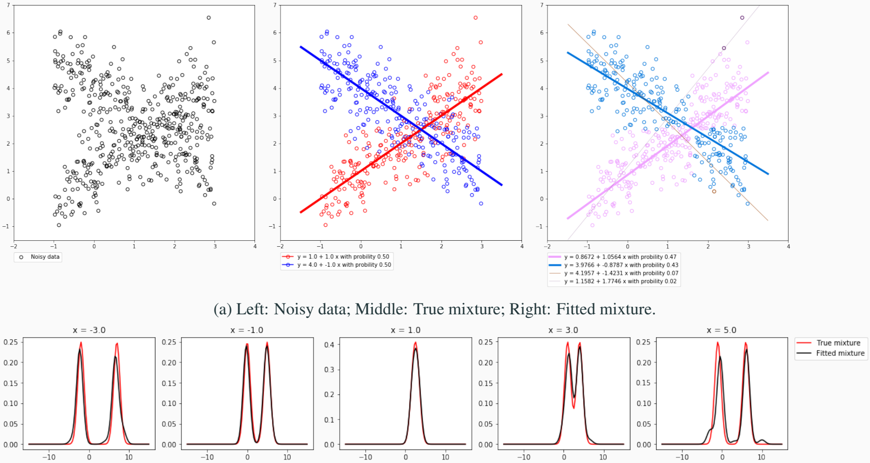
$$\beta^{(1)} = \underset{G}{\mathrm{argmax}}\, \langle \mathfrak{f}_G, \nabla L(\mathfrak{f}_{G^{(0)}})\rangle = \underset{\beta}{\mathrm{argmax}}\, \langle \mathfrak{f}_\beta, \nabla L(\mathfrak{f}_{G^{(0)}})\rangle.$$

Take $G^{(1)} = \pi_0^{(1)}\beta^{(0)} + \pi_1^{(1)}\beta^{(1)}$ where

$$(\pi_0^{(1)}, \pi_1^{(1)}) := \underset{\pi_0, \pi_1 \geq 0 : \pi_0 + \pi_1 = 1}{\mathrm{argmax}} L(\pi_0 \mathfrak{f}_{\beta^{(0)}} + \pi_1 \mathfrak{f}_{\beta^{(1)}}).$$

Continue with $G^{(2)}, G^{(3)}, \dots$.

Example One

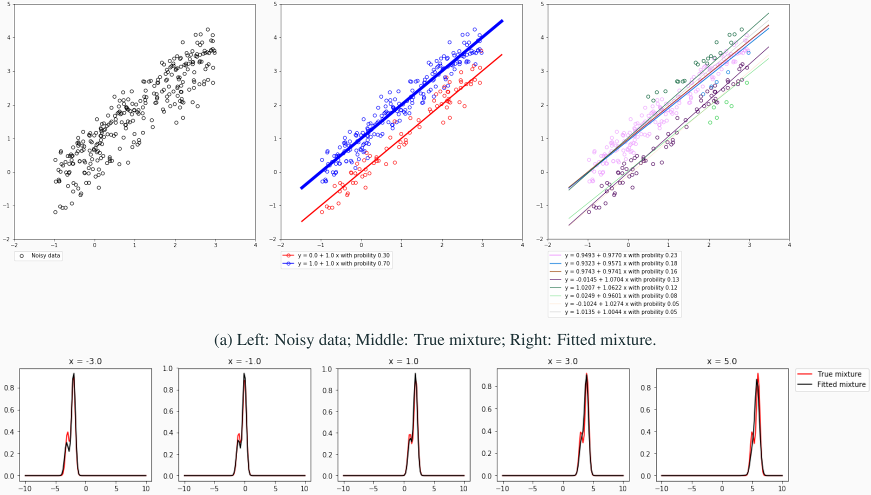


(a) Left: Noisy data; Middle: True mixture; Right: Fitted mixture.

(b) Probability density functions at different x's

Figure 1: Two-component concurrent mixture

Example Two

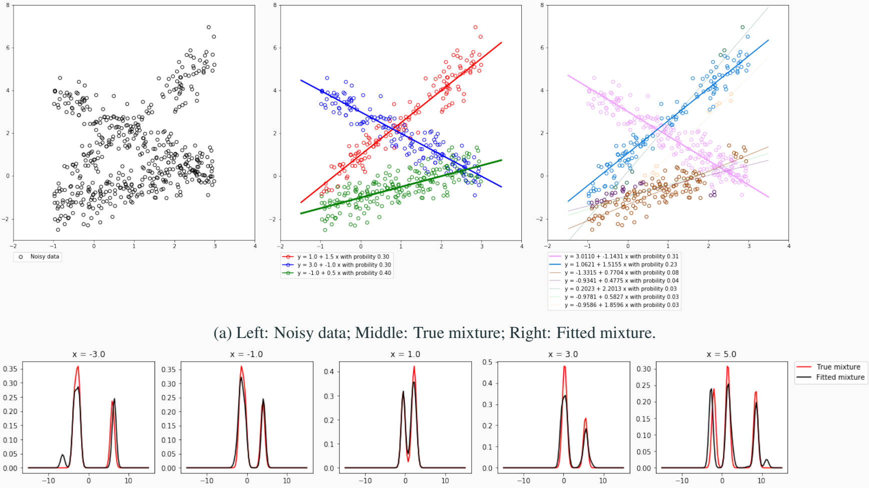


(a) Left: Noisy data; Middle: True mixture; Right: Fitted mixture.

(b) True and fitted probability density functions at different x's

Figure 2: Two-component parallel mixture

Example Three



(a) Left: Noisy data; Middle: True mixture; Right: Fitted mixture.

(b) True and fitted probability density functions at different x's

Figure 3: Three-component mixture

Loss Function

$f^{x}_{\hat{G}_n}$ seems to be a good estimator for $f^{x}_{G^*}$ for many x's.

This motivated us to consider the loss function

$$L(\hat{G}_n, G^*) := \frac{1}{n}\sum_{i=1}^{n}\left(\sqrt{f^{x_i}_{\hat{G}_n}(y)} - \sqrt{f^{x_i}_{G^*}(y)}\right)^2 dy$$

We assumed a fixed design setting and worked with this loss function.

Risk Bound on $L(\hat{G}_n, G^*)$

Suppose G^* has compact support $S \subseteq \mathbb{R}^p$ and let

$$S_0 := \{x_1, \dots, x_n\} \quad \text{and} \quad S_{<>} := \cup_{i=1}^{n} \{\langle x_i, \beta\rangle : \beta \in S\}.$$

Then, for a positive constant C_p depending on p alone,

$$\mathbb{E}L(\hat{G}_n, G^*) \leq C_p \mathrm{Vol}(S_0^1) \frac{\mathrm{Vol}(S_{<>}^1)}{n} (\sqrt{\log n})^{p+5}.$$

In the special case when G^* is discrete with k atoms supported on $\beta_1^*, \dots, \beta_k^*$,

$$\mathbb{E}L(\hat{G}_n, G^*) \leq C_p \mathrm{Vol}(S_0^1) \frac{k}{n} (\sqrt{\log n})^{p+5} \left(2 + \max_{1 \leq j \leq k} \|\beta_j^*\| \max_{i \neq j} \|x_i - x_j\|\right)$$

Summary

The NPMLE is a very good density estimator for Gaussian location mixtures.

We proved some Hellinger accuracy results for the NPMLE for $d \geq 1$.

The NPMLE is naturally applicable for empirical Bayes estimation and testing of normal means.

We have also defined and studying a version of the NPMLE for the mixture of homoscedastic linear regressions model.

THANK YOU.