On Nonparametric Maximum Likelihood Estimation with Heterogeneous Data

ICMS Workshop: Structural Breaks and Shape Constraints

Bodhisattva Sen[1]

Department of Statistics, Columbia University, New York

May 16, 2022

Joint work with Jake Soloff and Aditya Guntuboyina (University of California at Berkeley)

`https://arxiv.org/abs/2109.03466`

[1]Supported by NSF grant DMS-2015376

Basic Model

We observe data $Y_1, \ldots, Y_n \in \mathbb{R}^d$ $(d \geq 1)$ drawn from the model[2]:

$$Y_i = \theta_i + Z_i \qquad \text{with} \quad Z_i \overset{\text{ind}}{\sim} N_d(0, \Sigma_i)$$

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

$$\theta_1, \ldots, \theta_n \overset{\text{iid}}{\sim} G^*, \qquad G^* : \text{unknown distribution on } \mathbb{R}^d$$

- $\Sigma_1, \ldots, \Sigma_n \in \mathbb{R}^{d \times d}$ are known covariance matrices, e.g.,

$$\Sigma_i = I_d \qquad \text{or} \qquad \Sigma_i = \operatorname{diag}(\sigma_{i,1}^2, \ldots, \sigma_{i,d}^2)$$

[2] Here $(\theta_1, \ldots, \theta_n)$ is assumed to be independent of $(Z_1, \ldots, Z_n)$.

Basic Model

We observe data $Y_1, \ldots, Y_n \in \mathbb{R}^d$ ($d \geq 1$) drawn from the model[2]:

$$Y_i = \theta_i + Z_i \qquad \text{with} \quad Z_i \overset{\text{ind}}{\sim} N_d(0, \Sigma_i)$$

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

$$\theta_1, \ldots, \theta_n \overset{\text{iid}}{\sim} G^*, \qquad G^* : \text{unknown distribution on } \mathbb{R}^d$$

- $\Sigma_1, \ldots, \Sigma_n \in \mathbb{R}^{d \times d}$ are known covariance matrices, e.g.,

$$\Sigma_i = I_d \qquad \text{or} \qquad \Sigma_i = \operatorname{diag}(\sigma_{i,1}^2, \ldots, \sigma_{i,d}^2)$$

Examples: Sparsity ($\mathbb{P}_{G^*}\{\theta_1 = 0\}$ large), clustering (G^* discrete), and G^* can have structure (e.g., G^* lower dimensional/manifold structure)

[2] Here $(\theta_1, \ldots, \theta_n)$ is assumed to be independent of $(Z_1, \ldots, Z_n)$.

Basic Model

We observe data $Y_1, \ldots, Y_n \in \mathbb{R}^d$ ($d \geq 1$) drawn from the model[2]:

$$Y_i = \theta_i + Z_i \qquad \text{with} \quad Z_i \overset{\text{ind}}{\sim} N_d(0, \Sigma_i)$$

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

$$\theta_1, \ldots, \theta_n \overset{\text{iid}}{\sim} G^*, \qquad G^* : \text{unknown distribution on } \mathbb{R}^d$$

- $\Sigma_1, \ldots, \Sigma_n \in \mathbb{R}^{d \times d}$ are known covariance matrices, e.g.,

$$\Sigma_i = I_d \qquad \text{or} \qquad \Sigma_i = \operatorname{diag}(\sigma_{i,1}^2, \ldots, \sigma_{i,d}^2)$$

Examples: Sparsity ($\mathbb{P}_{G^*}\{\theta_1 = 0\}$ large), clustering (G^* discrete), and G^* can have structure (e.g., G^* lower dimensional/manifold structure)

Questions: Can we estimate G^* nonparametrically?
Can we denoise the Y_i's and "estimate" the θ_i's?
How to compute the estimator(s)?

[2] Here $(\theta_1, \ldots, \theta_n)$ is assumed to be independent of $(Z_1, \ldots, Z_n)$.

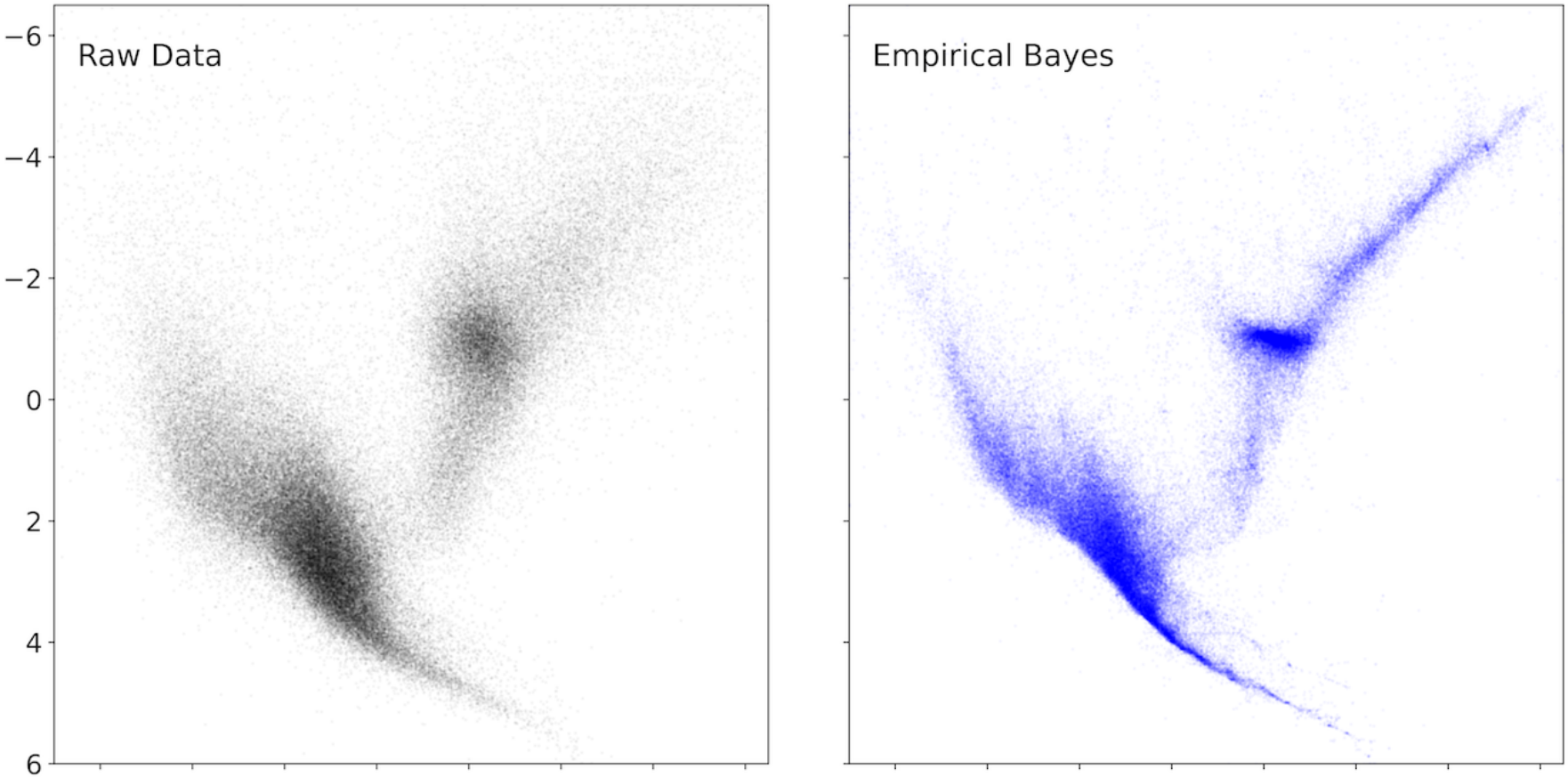


Figure: A sub-sample of $n = 10^5$ TGAS stars and its denoised version in color-magnitude space (see e.g., Anderson et al. (2017)).

The Color Magnitude Diagram (or CMD) is a plot of observational data which shows how a population of stars can be plotted in terms of their brightness (or luminosity) and color (or surface temperature).

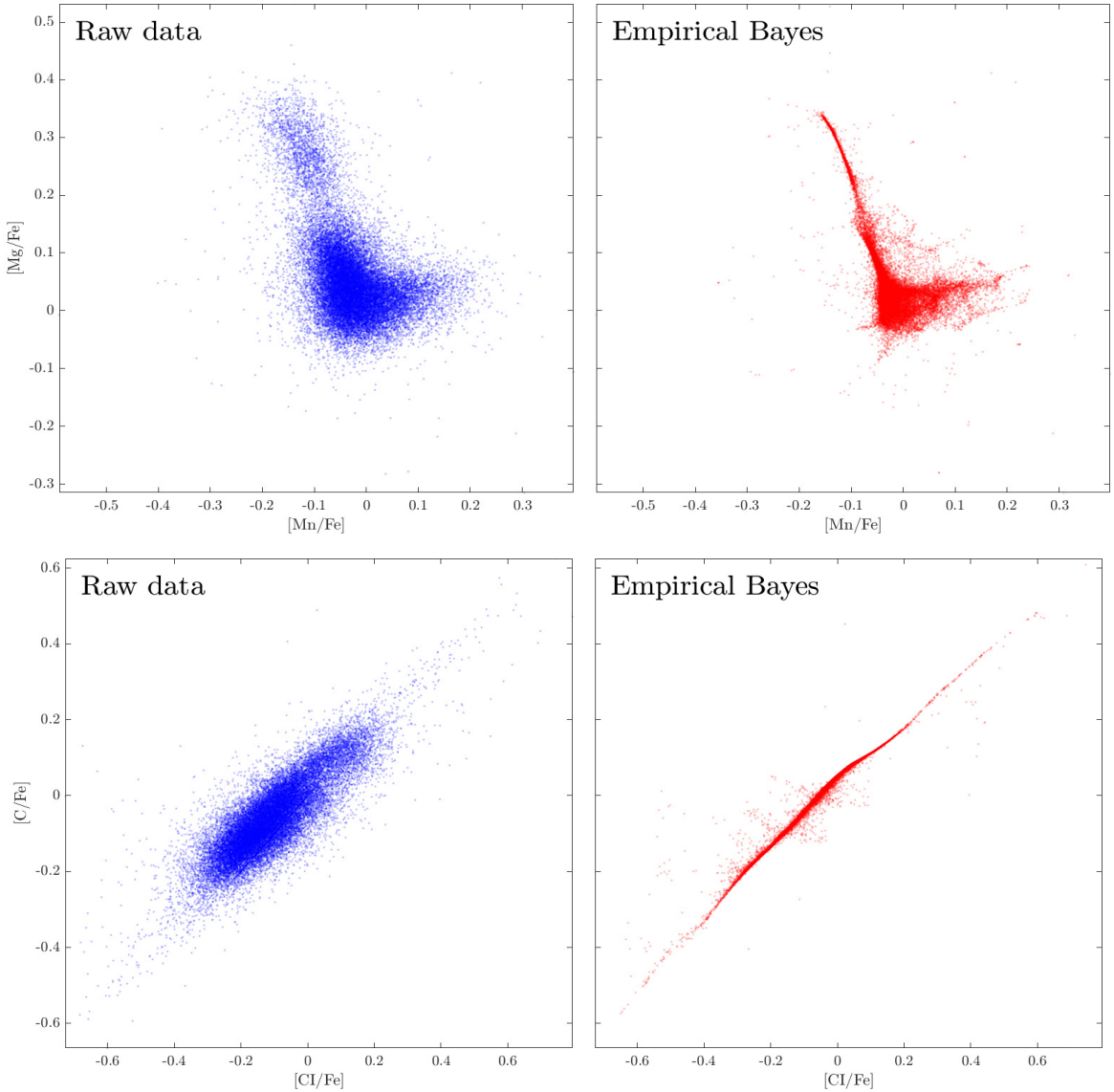


Figure: The chemical abundance of $n \approx 3 \times 10^4$ red clump stars and its denoised version in standardized [Mg/Fe]-[Mn/Fe] plane (top) and [C/Fe]-[CI/Fe] plane (bottom); ongoing work with Yangjing Zhang & Ying Cui (Uni. Minnesota); also see Ratcliffe et al. (2020).

Outline

1. The NPMLE for Heterogenous Gaussian Location Mixtures
2. Empirical Bayes Estimation of Normal Means
3. Deconvolution

The NPMLE (Kiefer and Wolfowitz, 1956)

- **Model**: $Y_i = \theta_i + Z_i \quad \text{with } \theta_i \overset{\text{iid}}{\sim} G^*,\ Z_i \overset{\text{ind}}{\sim} N_d(0, \Sigma_i),\ i = 1, \ldots, n$

- Under this model $Y_1, \ldots, Y_n$ are independent and marginally

$$Y_i \sim f_{G^*, \Sigma_i} \qquad \text{with} \qquad f_{G^*, \Sigma_i}(y) := \int_{\mathbb{R}^d} \phi_{\Sigma_i}(y - \theta) dG^*(\theta)$$

and $\phi_{\Sigma_i}(\cdot)$ is the density of $N_d(0, \Sigma_i)$

The NPMLE (Kiefer and Wolfowitz, 1956)

- **Model**: $Y_i = \theta_i + Z_i \quad \text{with } \theta_i \overset{\text{iid}}{\sim} G^*,\ Z_i \overset{\text{ind}}{\sim} N_d(0, \Sigma_i),\ i = 1, \ldots, n$

- Under this model $Y_1, \ldots, Y_n$ are independent and marginally

$$Y_i \sim f_{G^*, \Sigma_i} \qquad \text{with} \qquad f_{G^*, \Sigma_i}(y) := \int_{\mathbb{R}^d} \phi_{\Sigma_i}(y - \theta) dG^*(\theta)$$

  and $\phi_{\Sigma_i}(\cdot)$ is the density of $N_d(0, \Sigma_i)$

- The nonparametric maximum likelihood estimator (NPMLE) of G^*:

$$\hat{G}_n \in \operatorname*{argmax}_G \sum_{i=1}^{n} \log f_{G, \Sigma_i}(Y_i)$$

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

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

$$\hat{G}_n \in \operatorname*{argmax}_G \sum_{i=1}^{n} \log f_{G,\Sigma_i}(Y_i) \qquad (\star)$$

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

- $f_{\hat{G}_n,\Sigma_1}(Y_1), \ldots, f_{\hat{G}_n,\Sigma_n}(Y_n)$ are unique
- There exists a discrete solution $\hat{G}_n$ with at most n atoms; usually $\#$ atoms of $\hat{G}_n \ll n$

$$\hat{G}_n \in \operatorname*{argmax}_G \sum_{i=1}^{n} \log f_{G,\Sigma_i}(Y_i) \qquad (\star)$$

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

- $f_{\hat{G}_n,\Sigma_1}(Y_1), \ldots, f_{\hat{G}_n,\Sigma_n}(Y_n)$ are unique
- There exists a discrete solution $\hat{G}_n$ with at most n atoms; usually # atoms of $\hat{G}_n \ll n$
- When $d = 1$ and $\Sigma_i = \sigma^2$, it is known that $(\star)$ has a unique solution $\hat{G}_n$ for all observations $X_1, \ldots, X_n$ (Lindsay and Roeder, 1993)
- However, uniqueness of $\hat{G}_n$ may not hold when $d > 1$
- G is unconstrained here. Enforcing constraints (e.g., fixing # atoms of G) makes problem non-convex & involve tuning parameters

$$\hat{G}_n \in \operatorname*{argmax}_G \sum_{i=1}^{n} \log f_{G,\Sigma_i}(Y_i) \qquad \text{with } f_{G,\Sigma_i}(\cdot) := \int \phi_{\Sigma_i}(\cdot - \theta) dG(\theta)$$

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

- Approximate algorithms: Laird (1978), Bohning (1999), Lindsay (1995), Lashkari & Golland (2007), Jiang & Zhang (2009), Koenker & Mizera (2014), Feng & Dicker (2015), Dicker & Zhao (2016)

- These include Frank-Wolfe methods (VDM, VEM[3]), EM algorithms

[3] Vertex Direction Method (VDM, Lindsay, 1983), Vertex Exchange M. (VEM, Böhning, 1985)

$$\hat{G}_n \in \operatorname*{argmax}_G \sum_{i=1}^n \log f_{G,\Sigma_i}(Y_i) \qquad \text{with } f_{G,\Sigma_i}(\cdot) := \int \phi_{\Sigma_i}(\cdot - \theta) dG(\theta)$$

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

- Approximate algorithms: Laird (1978), Bohning (1999), Lindsay (1995), Lashkari & Golland (2007), Jiang & Zhang (2009), Koenker & Mizera (2014), Feng & Dicker (2015), Dicker & Zhao (2016)

- These include Frank-Wolfe methods (VDM, VEM[3]), EM algorithms

- **Direct discretization**: Fix $a_1, \ldots, a_m \in \mathbb{R}^d$ for large m and solve:

$$\max_{w_1,\ldots,w_m} \left\{ \sum_{i=1}^n \log\Big(\sum_{j=1}^m w_j \, \phi_{\Sigma_i}(Y_i - a_j) \Big) : w_j \geq 0 \text{ and } \sum_{j=1}^m w_j = 1 \right\}$$

- **Question**: How to choose $a_1, \ldots, a_m$?

[3] Vertex Direction Method (VDM, Lindsay, 1983), Vertex Exchange M. (VEM, Böhning, 1985)

$$\hat{G}_n \in \operatorname*{argmax}_G \sum_{i=1}^{n} \log f_{G,\Sigma_i}(Y_i) \qquad (\star)$$

- Every solution to $(\star)$ is discrete with a finite number of atoms, supported on the ridgeline manifold (Ray & Lindsay, 2005)

$$\mathcal{M} := \left\{ x^*(\alpha) : \alpha = (\alpha_1, \ldots, \alpha_n),\ \alpha_i \geq 0,\ \sum_{i=1}^{n} \alpha_i = 1 \right\}$$

$$\text{where } x^*(\alpha) := \left(\sum_{i=1}^{n} \alpha_i \Sigma_i^{-1} \right)^{-1} \sum_{i=1}^{n} \alpha_i \Sigma_i^{-1} Y_i$$

- (Homoscedastic) If $\Sigma_i = \Sigma$ for all i, or if $\Sigma_i = c_i \Sigma$, the ridgeline manifold $\mathcal{M}$ is the convex hull of the data $\operatorname{conv}(\{Y_1, \ldots, Y_n\})$

- (Diagonal Covariances) If Σ_i is a diagonal matrix for every i, $\mathcal{M}$ is contained in the axis-aligned minimum bounding box of the data:

$$\Pi_{j=1}^{d} \left[\min_{i=1,\ldots,n} Y_{ij}, \max_{i=1,\ldots,n} Y_{ij} \right]$$

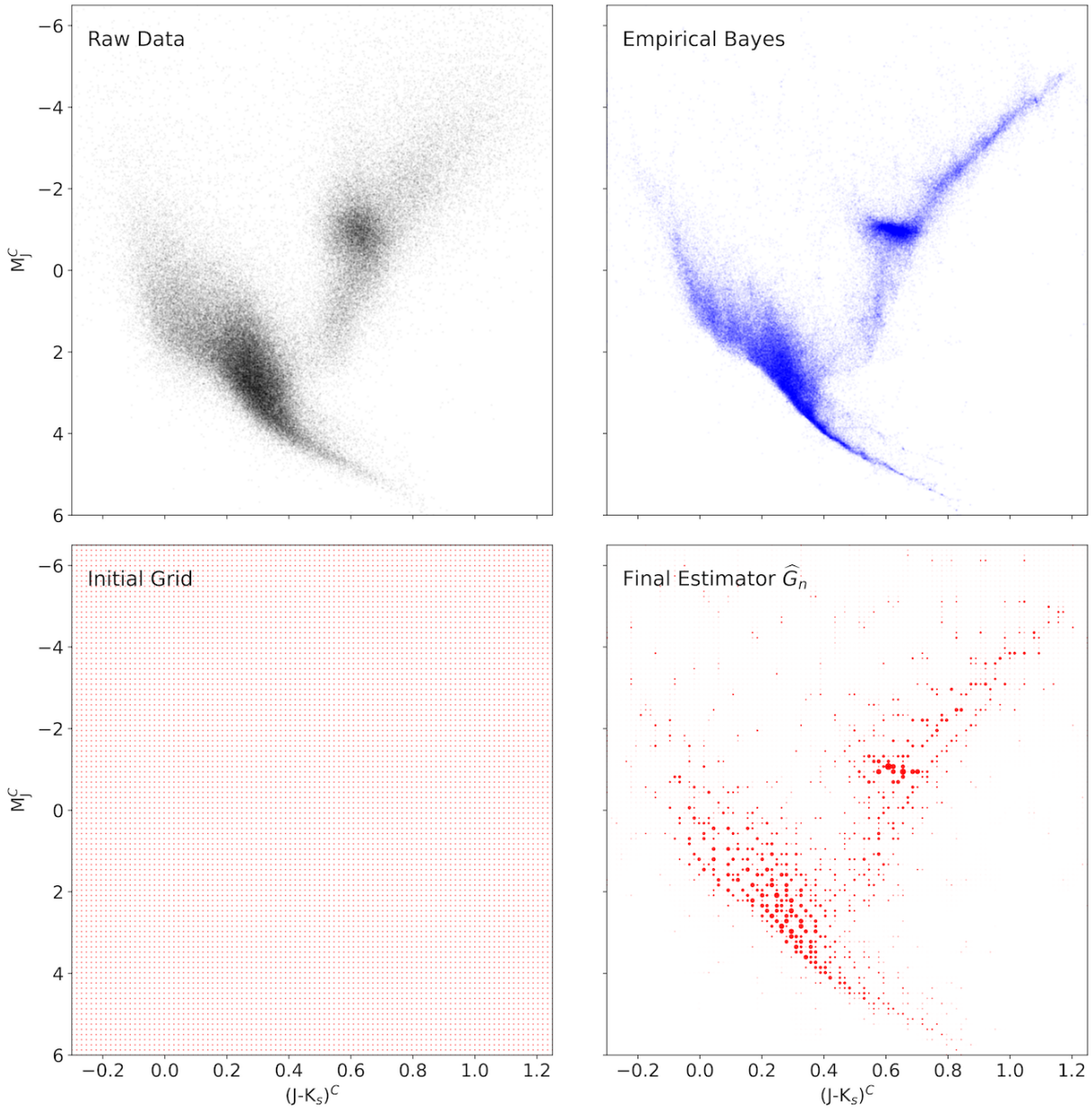


Figure: A sub-sample of $n = 10^5$ TGAS stars and its denoised version in color-magnitude space (see e.g., Anderson et al. (2017)).

True f_{G^*} (left) and estimated $f_{\hat{G}_n}$ (right) when $\Sigma_i \equiv I_2$

When $\Sigma_i \equiv I_2$, then $Y_1, \ldots, Y_n$ are iid $f_{G^*}(\cdot) \equiv \int \phi_{I_2}(\cdot - \theta) dG^*(\theta)$.

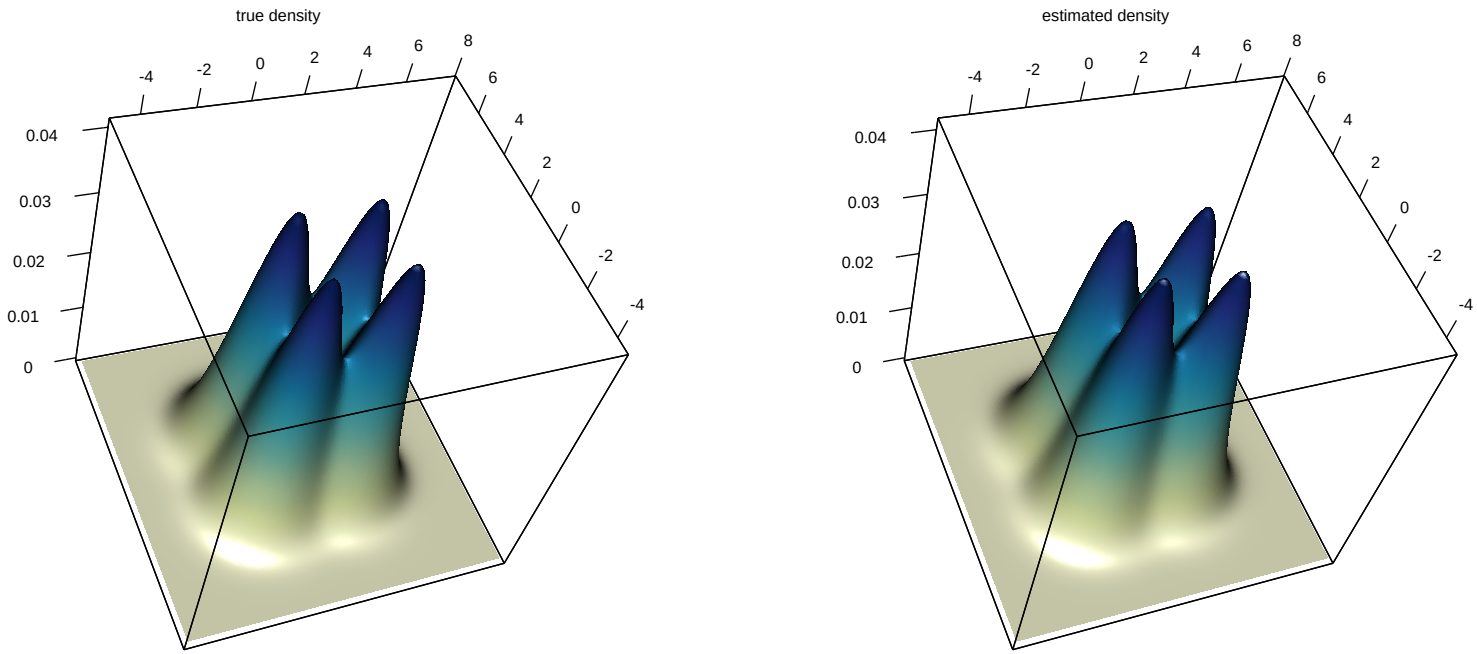


Figure: Here sample size is $n = 10^4$. G^* is discrete which puts equal mass on the four points $(0,0), (3,0), (0,3), (3,3)$.

True f_{G^*} (left) and estimated $f_{\hat{G}_n}$ (right) when $\Sigma_i \equiv I_2$

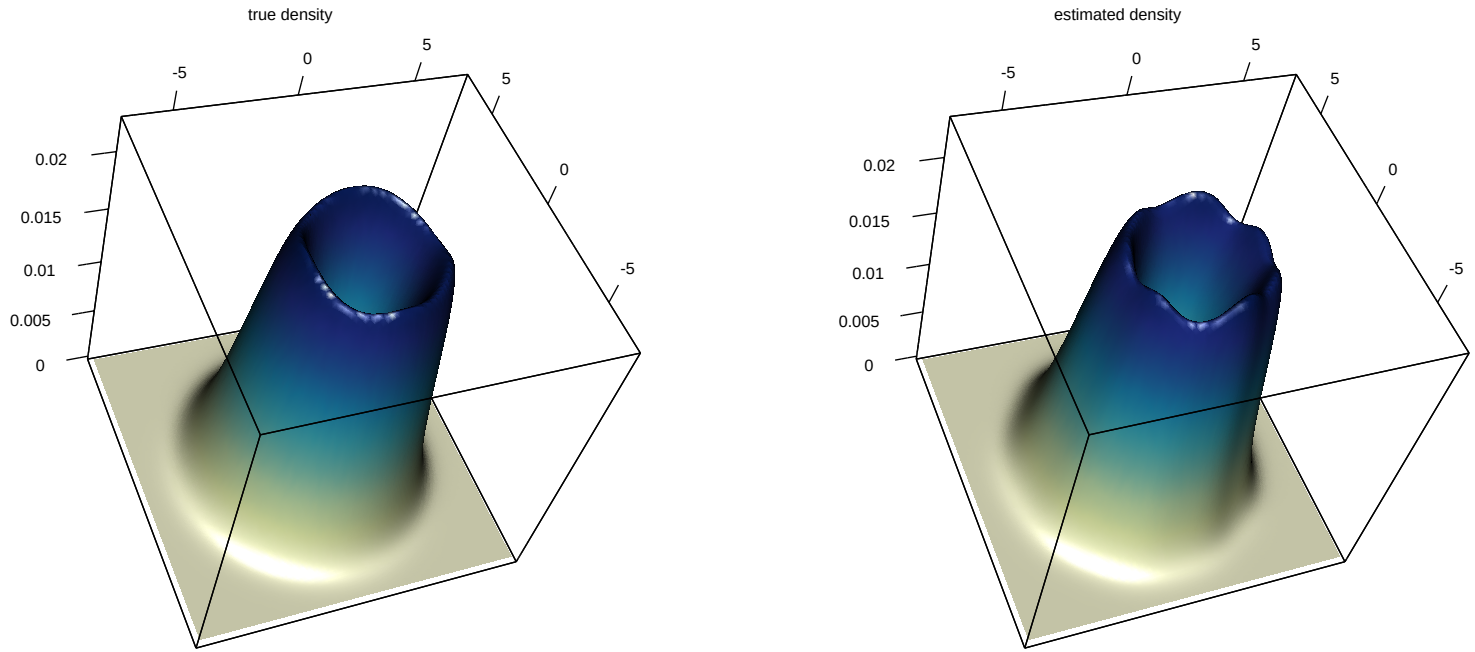


Figure: Here $n = 10^4$. G^* is uniformly distributed on a circle of radius 3.

True f_{G^*} (left) and estimated $f_{\hat{G}_n}$ (Right) when $\Sigma_i \equiv I_2$

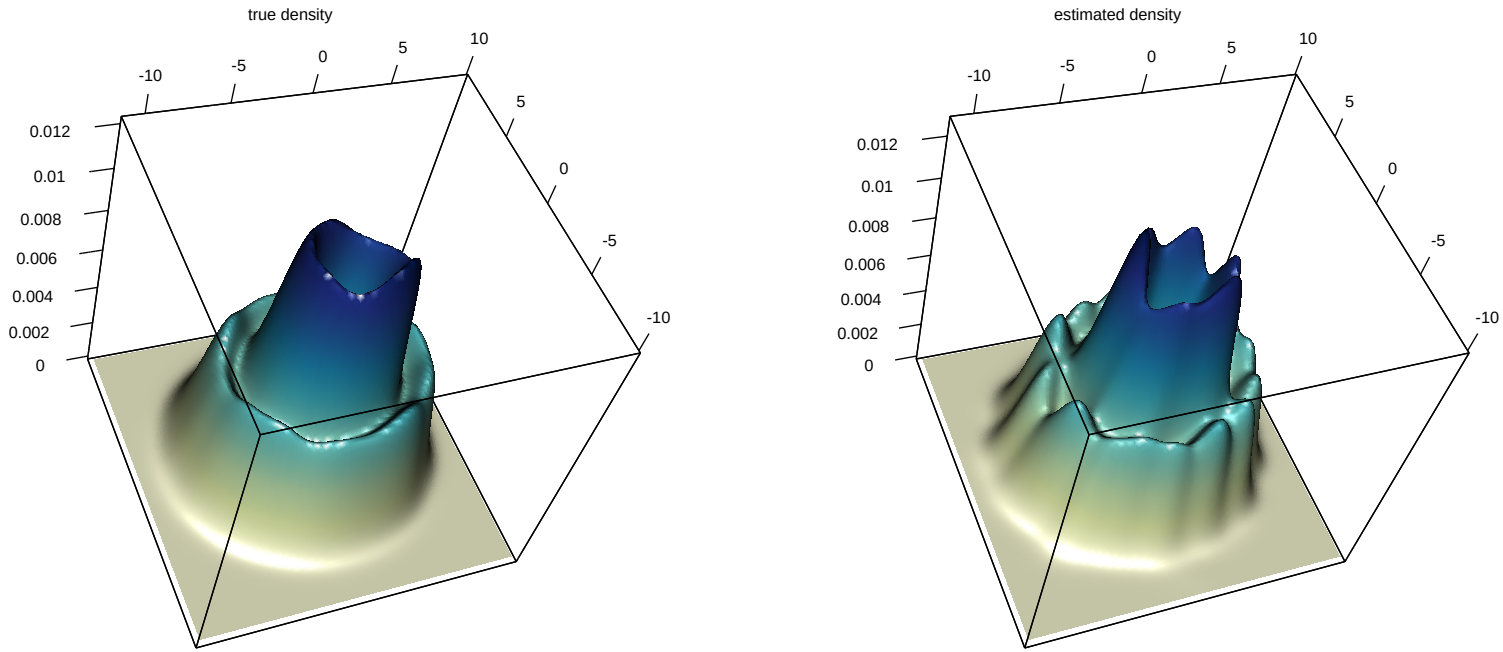


Figure: Here $n = 10^4$. G^* is uniformly distributed on two concentric circles of radii 3 and 6.

Accuracy of $f_{\hat{G}_n,\Sigma_i}$ for f_{G^*,Σ_i}

- How accurate is $f_{\hat{G}_n,\Sigma_i}$ for estimating f_{G^*,Σ_i}?
- As the distribution of Y_i varies with i, we consider estimation quality of the NPMLE in terms of the average Hellinger distance
- We study accuracy via risk under average squared Hellinger distance:

$$\mathfrak{H}^2(\hat{G}_n, G^*) := \frac{1}{n}\sum_{i=1}^{n}\int\left(\sqrt{f_{\hat{G}_n,\Sigma_i}} - \sqrt{f_{G^*,\Sigma_i}}\right)^2$$

and prove upper bounds for $\mathbb{E}\left[\mathfrak{H}^2(\hat{G}_n, G^*)\right]$

Accuracy of $f_{\hat{G}_n,\Sigma_i}$ for f_{G^*,Σ_i}

- How accurate is $f_{\hat{G}_n,\Sigma_i}$ for estimating f_{G^*,Σ_i}?
- As the distribution of Y_i varies with i, we consider estimation quality of the NPMLE in terms of the average Hellinger distance
- We study accuracy via risk under average squared Hellinger distance:

$$\mathfrak{H}^2(\hat{G}_n, G^*) := \frac{1}{n}\sum_{i=1}^{n}\int\left(\sqrt{f_{\hat{G}_n,\Sigma_i}} - \sqrt{f_{G^*,\Sigma_i}}\right)^2$$

  and prove upper bounds for $\mathbb{E}\left[\mathfrak{H}^2(\hat{G}_n, G^*)\right]$

- We argue that $f_{\hat{G}_n,\Sigma_i}$ is a very good estimator for f_{G^*,Σ_i} when G^* satisfies natural assumptions (such as being discrete)
- Our work is heavily inspired by Saha and Guntuboyina (2020) and Zhang (2009)

When G^* has compact support

- Suppose that G^* has compact support $S \subset \mathbb{R}^d$
- Assume: $\underline{a}^2 I_d \lesssim \Sigma_i \lesssim \bar{a}^2 I_d$ (for fixed $\underline{a}, \bar{a} > 0$)
- Then (for $S^a := S + a\,B(0,1) = \cup_{x \in S} B(x,a)$),

$$\mathbb{E}\left[\mathfrak{H}^2(\hat{G}_n, G^*)\right] \leq C_{d,\underline{a},\bar{a}} \frac{\mathrm{Vol}(S^{\bar{a}})}{n} (\log n)^{d+1}$$

for a positive constant $C_{d,\bar{a},\underline{a}}$ depending on $d, \bar{a}, \underline{a}$ alone

When G^* has compact support

- Suppose that G^* has compact support $S \subset \mathbb{R}^d$
- Assume: $\underline{a}^2 I_d \lesssim \Sigma_i \lesssim \bar{a}^2 I_d$ (for fixed $\underline{a}, \bar{a} > 0$)
- Then (for $S^a := S + a\,B(0,1) = \cup_{x \in S} B(x,a)$),

$$\mathbb{E}\left[\mathfrak{H}^2(\hat{G}_n, G^*)\right] \leq C_{d,\underline{a},\bar{a}} \frac{\mathrm{Vol}(S^{\bar{a}})}{n} (\log n)^{d+1}$$

  for a positive constant $C_{d,\bar{a},\underline{a}}$ depending on $d, \bar{a}, \underline{a}$ alone

- The risk of the NPMLE is $\lesssim \mathrm{Vol}(S^{\bar{a}})/n$ (ignoring logarithmic factors). Extensions to non-compact support are also possible
- NPMLE is completely tuning-free and does not use any knowledge of the support of G^*.

When G^* is a discrete distribution

- Suppose that G^* is a discrete distribution with k^* atoms. Then,

$$\mathbb{E}\left[\mathfrak{H}^2\left(\hat{G}_n, G^*\right)\right] \leq C_{d,\bar{a},\underline{a}}\left(\frac{k^*}{n}\right)(\log n)^{d+1}$$

  for a positive constant $C_{d,\bar{a},\underline{a}}$ depending on $d, \bar{a}, \underline{a}$ alone

- Shows that the risk of $f_{\hat{G}_n,\Sigma_i}$ is k^*/n up to logarithmic factors in n
- This is remarkable because the NPMLE does not a priori know k^*
- Minimax lower bounds show that no estimator can estimate k^*-component Gaussian mixtures at a rate that is better than k^*/n

Empirical Bayes Estimation of Normal Means

Empirical Bayes Estimation of Normal Means

- Consider the problem of estimating $\theta_1, \ldots, \theta_n$ where

$$Y_i = \theta_i + Z_i \qquad \text{with } Z_i \overset{\text{ind}}{\sim} N_d(0, \Sigma_i)$$

- As $\theta_1, \ldots, \theta_n \overset{\text{iid}}{\sim} G^*$, a simple Bayesian approach to this problem estimates each θ_i by its oracle posterior mean

$$\theta_i^* := \mathbb{E}[\theta_i | Y_i] = \frac{\int \theta \, \phi_{\Sigma_i}(Y_i - \theta) dG^*(\theta)}{f_{G^*, \Sigma_i}(Y_i)}$$

Empirical Bayes Estimation of Normal Means

- Consider the problem of estimating $\theta_1, \ldots, \theta_n$ where

$$Y_i = \theta_i + Z_i \qquad \text{with } Z_i \overset{\text{ind}}{\sim} N_d(0, \Sigma_i)$$

- As $\theta_1, \ldots, \theta_n \overset{\text{iid}}{\sim} G^*$, a simple Bayesian approach to this problem estimates each θ_i by its oracle posterior mean

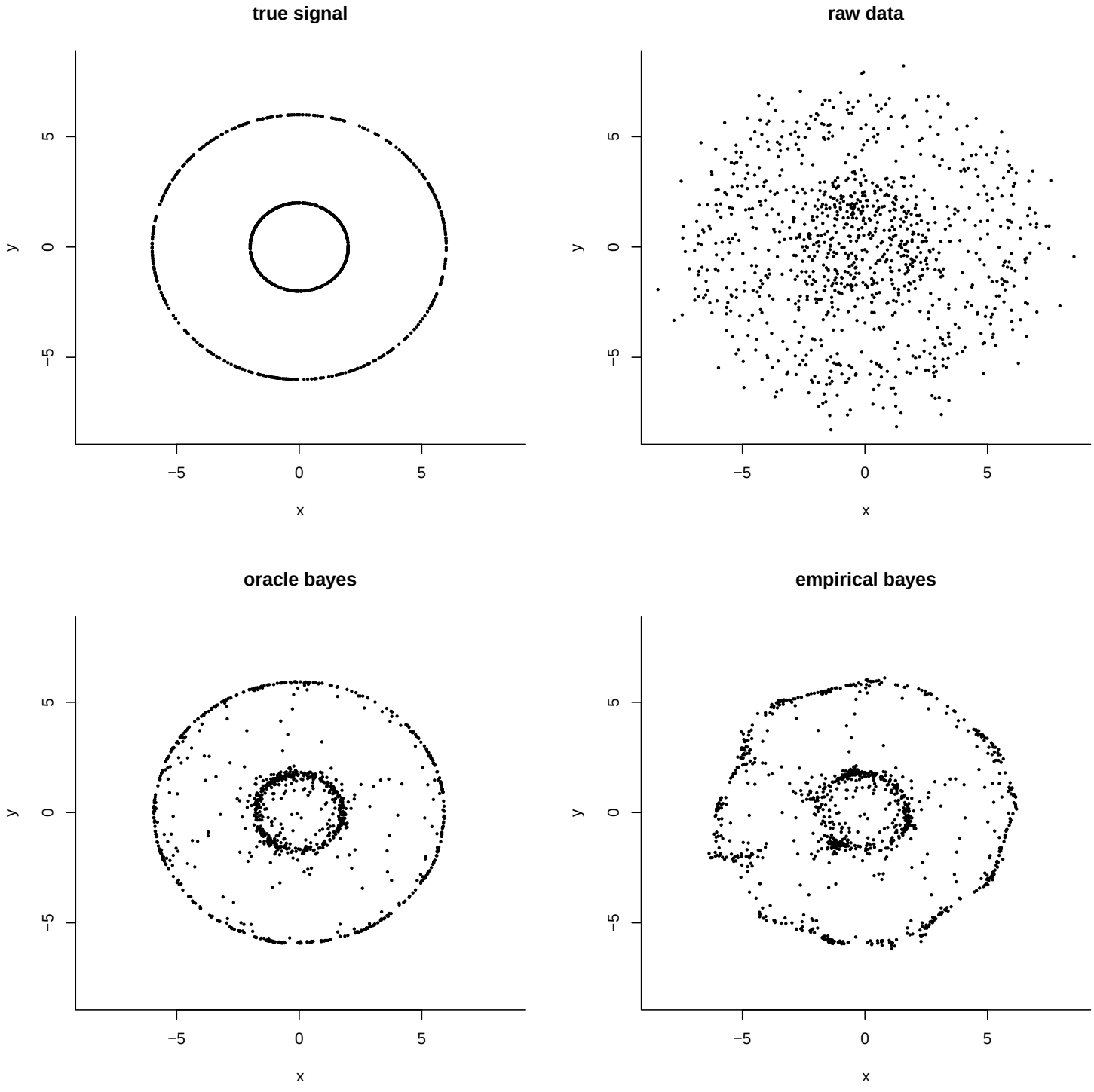
$$\theta_i^* := \mathbb{E}[\theta_i | Y_i] = \frac{\int \theta \, \phi_{\Sigma_i}(Y_i - \theta) dG^*(\theta)}{f_{G^*, \Sigma_i}(Y_i)}$$

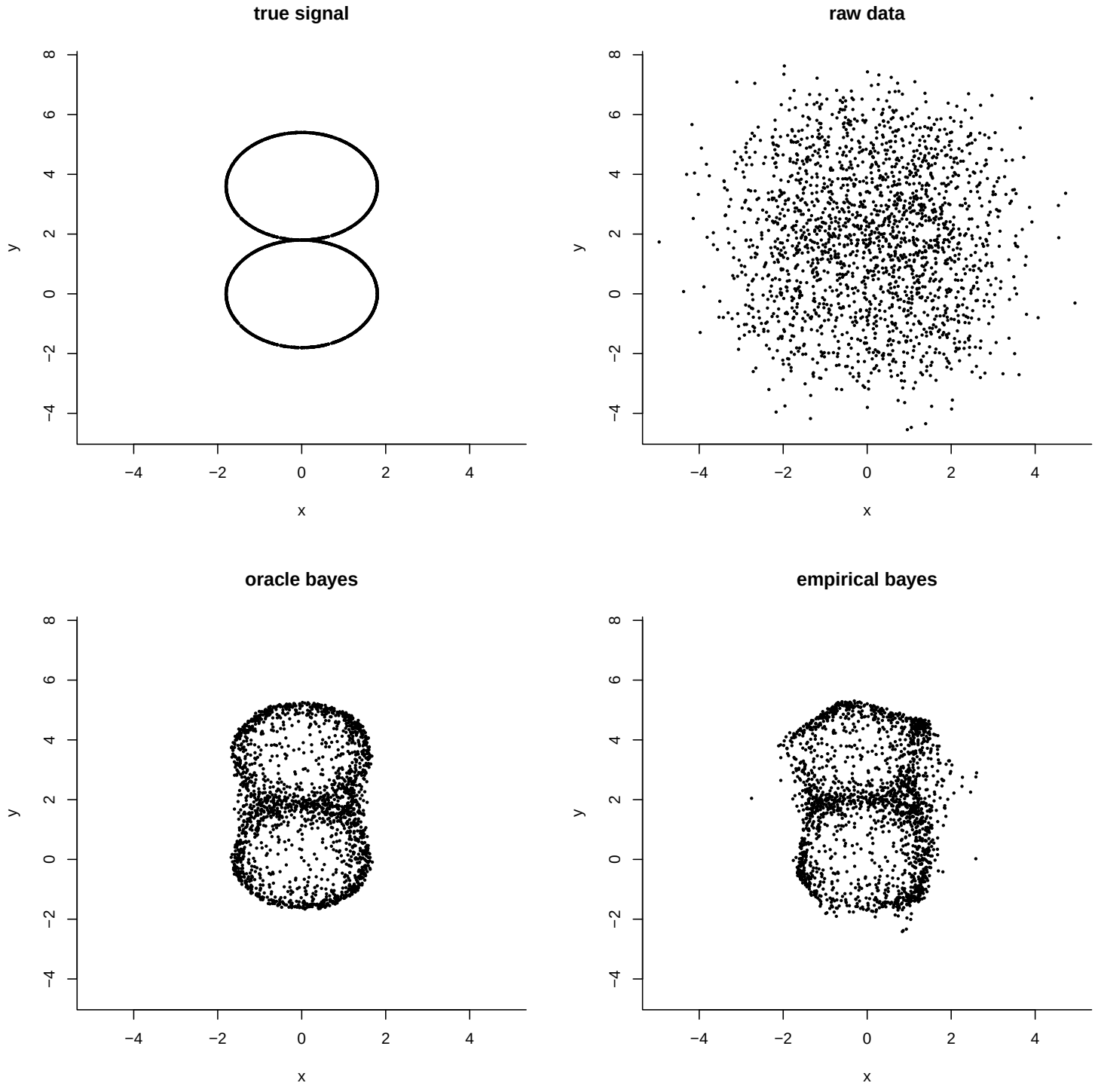
- Natural empirical Bayes estimator of θ_i^* is

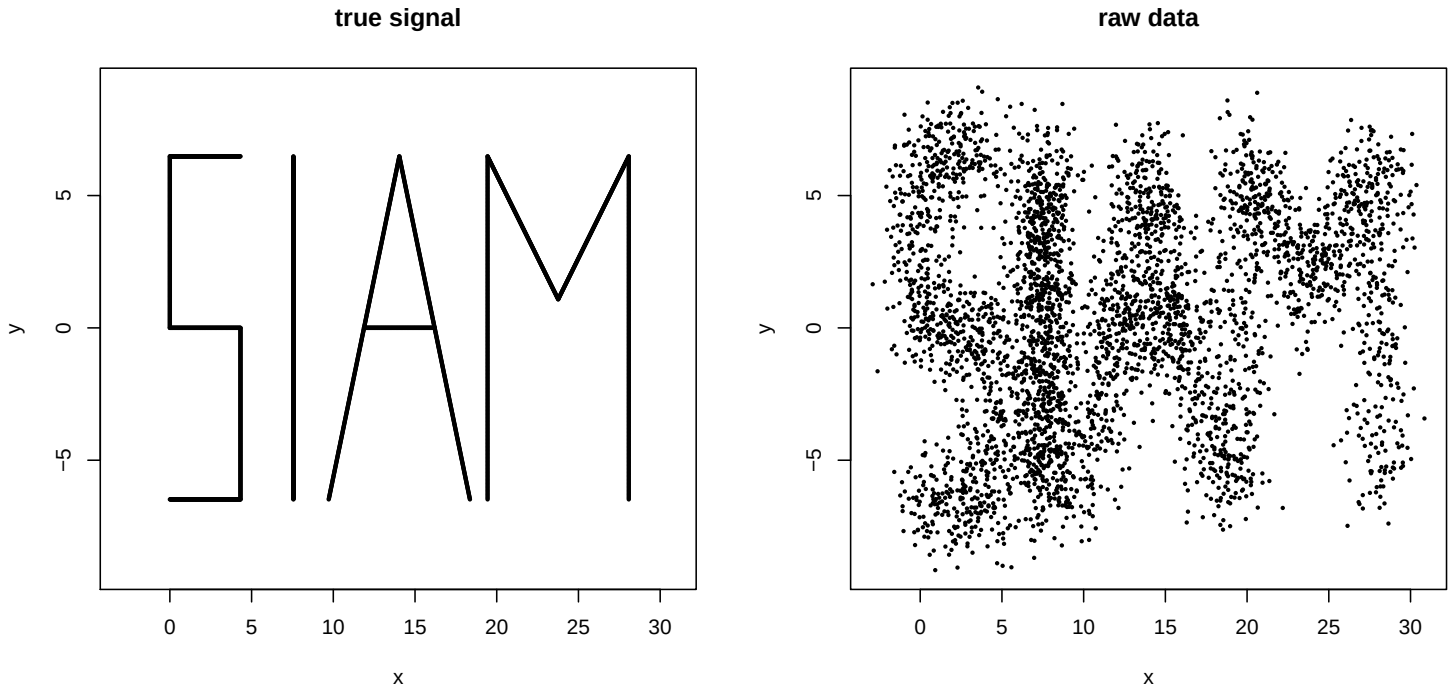
$$\hat{\theta}_i := \frac{\int \theta \phi_{\Sigma_i}(Y_i - \theta) d\hat{G}_n(\theta)}{f_{\hat{G}_n, \Sigma_i}(Y_i)}.$$

- This is the general maximum likelihood empirical Bayes (GMLEB) estimator of Jiang and Zhang (2009) who studied it in $d = 1$

- This estimator is tuning-free and provides excellent shrinkage


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 θ_i^*

$$Y_i = \theta_i + Z_i \qquad Z_i \overset{\text{ind}}{\sim} N_d(0, \Sigma_i), \;\; \theta_1, \ldots, \theta_n \overset{\text{iid}}{\sim} G^*$$

- **Oracle posterior mean**: $\theta_i^* = \mathbb{E}[\theta_i | Y_i] = \frac{\int \theta \, \phi_{\Sigma_i}(Y_i - \theta) dG^*(\theta)}{f_{G^*, \Sigma_i}(Y_i)}$
- **Empirical Bayes** estimator: $\hat{\theta}_i = \frac{\int \theta \phi_{\Sigma_i}(Y_i - \theta) d\hat{G}_n(\theta)}{f_{\hat{G}_n, \Sigma_i}(Y_i)}$
- How accurate is $\hat{\theta}_i$ for estimating θ_i^*?

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

$$Y_i = \theta_i + Z_i \qquad Z_i \overset{\text{ind}}{\sim} N_d(0, \Sigma_i), \;\; \theta_1, \ldots, \theta_n \overset{\text{iid}}{\sim} G^*$$

- **Oracle posterior mean**: $\theta_i^* = \mathbb{E}[\theta_i | Y_i] = \frac{\int \theta \, \phi_{\Sigma_i}(Y_i - \theta) dG^*(\theta)}{f_{G^*, \Sigma_i}(Y_i)}$
- **Empirical Bayes** estimator: $\hat{\theta}_i = \frac{\int \theta \phi_{\Sigma_i}(Y_i - \theta) d\hat{G}_n(\theta)}{f_{\hat{G}_n, \Sigma_i}(Y_i)}$
- How accurate is $\hat{\theta}_i$ for estimating θ_i^*?

Accuracy result when G^* has compact support:

If G^* has compact support $S \subset \mathbb{R}^d$, then

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} \|\hat{\theta}_i - \theta_i^*\|^2\right] \leq C_{d, \underline{a}, \bar{a}} \frac{\text{Vol}(S^{\bar{a}})}{n} (\log n)^{d + \max\{d/2, 4\}}$$

The rate is $\text{Vol}(S^{\bar{a}})/n$ up to log factors (note $\hat{\theta}_i$ uses no knowledge of S)

Accuracy result when G^* is discrete

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

$$\mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\|\hat{\theta}_i - \theta_i^*\|^2\right] \leq C_{d,\underline{a},\bar{a}}\frac{k^*}{n}(\log n)^{d+\max\{d/2,4\}}$$

- We get the $\frac{k^*}{n}$ rate (up to log factors)

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

- Our work is heavily inspired by Saha and Guntuboyina (2020) who obtained similar results for homoscedastic errors

Deconvolution

- **Fundamental question**: How well does $\hat{G}_n$ estimate G^*?

- Known as the deconvolution problem and has received much attention in statistics; see e.g., Meister (2009), Delaigle (2008), ...

- Yet to our knowledge little is known about its deconvolution error

- **Fundamental question**: How well does $\hat{G}_n$ estimate G^*?

- Known as the deconvolution problem and has received much attention in statistics; see e.g., Meister (2009), Delaigle (2008), ...

- Yet to our knowledge little is known about its deconvolution error

- A natural loss for this problem is the Wasserstein distance from the theory of optimal transport

$$W_2^2(G, H) := \min_{(U,V)\in\Pi_{G,H}} \mathbb{E}\|U - V\|_2^2,$$

  where G, H are two probability measures on $\mathbb{R}^d$

- $\Pi_{G,H}$ denotes the set of couplings of G and H, i.e., joint distributions over $(U, V) \in \mathbb{R}^{2d}$ such that $U \sim G$ and $V \sim H$

- Nguyen (2013) connected the deconvolution error $W_2^2(G, H)$ to the density estimation error between the mixtures, i.e., $\mathfrak{H}^2(f_{G,\Sigma_i}, f_{H,\Sigma_i})$

- We can show that, for G^* compactly supported,

$$\mathbb{E}\left[W_2^2(\hat{G}_n, G^*)\right] \lesssim_{d,\underline{a},\bar{a}} \frac{1}{\log n}$$

- Minimax lower bounds show that no estimator can estimate arbitrary G^* at a rate that is better than $(\log n)^{-1}$

- Nguyen (2013) connected the deconvolution error $W_2^2(G, H)$ to the density estimation error between the mixtures, i.e., $\mathfrak{H}^2(f_{G,\Sigma_i}, f_{H,\Sigma_i})$

- We can show that, for G^* compactly supported,

$$\mathbb{E}\left[W_2^2(\hat{G}_n, G^*)\right] \lesssim_{d,\underline{a},\bar{a}} \frac{1}{\log n}$$

- Minimax lower bounds show that no estimator can estimate arbitrary G^* at a rate that is better than $(\log n)^{-1}$

What if G^* is structured?

- What happens when G^* is a Dirac measure? We can show that

$$\mathbb{E}\left[W_2^2(\hat{G}_n, G^*)\right] \lesssim_{d,\underline{a},\bar{a}} \left(\frac{\log n}{n}\right)^{1/4}$$

- Hints at adaptive properties of the NPMLE, that are yet to be fully explored ...

Summary

- The NPMLE is a very good estimator for Gaussian location mixtures when d is small
- We investigated characterization and basic properties of the NPMLE
- We proved average Hellinger accuracy results for the NPMLE
- The NPMLE is naturally applicable for empirical Bayes estimation
- The NPMLE exhibits adaptive rates when estimating G^* in the Wasserstein loss

Some comments on the computation of the NPMLE

- Can we develop efficient methods for approximately computing the NPMLE that: (i) move beyond gridding for greater scalability when d is moderate, and (ii) give a provably good approximation?

- In ongoing work (with Ying Cui and Yangjing Zhang) we study the computation of the NPMLE using a semismooth Newton (SSN) based augmented Lagrangian method (ALM). This can handle $n \approx 10^6$ and $m \approx 10^3$ for moderate $d \approx 5$.

- Using Wasserstein gradient descent shows improvements over EM algorithm and the ALM (ongoing work)

- How to do empirical Bayes when d is large? For example, when $n = 1000$ and $d = 20$ no method (including EM) seems to work well.

THANK YOU! Questions?