Nonparametric Estimation in a Two-component Mixture Model with Covariates

Nabarun Deb
Columbia University, New York

Joint Statistical Meeting 2019

29 July, 2019

Joint work with **Sujayam Saha** (Google)
Adityanand Guntuboyina (University of California at Berkeley) and
Bodhisattva Sen (Columbia University)

Preprint available at https://arxiv.org/abs/1810.07897

Mixture model with two-components

- **Data**: $Y_1, Y_2, \ldots, Y_n \overset{i.i.d.}{\sim} f$, $\qquad f$ density (pdf) on $\mathbb{R}$.
- **Two-groups model**: $\quad f(y) = \pi f_s(y) + (1 - \pi) f_b(y), \quad y \in \mathbb{R}.$
- f_b is a known density function.
- **Unknowns**: Mixing proportion $\pi \in [0, 1]$ and pdf f_s ($\neq f_b$).
- **Goals**: Estimate π and f_s (nonparametrically), under certain structural assumptions.

Mixture model with two-components

- **Data**: $Y_1, Y_2, \ldots, Y_n \overset{i.i.d.}{\sim} f$, f density (pdf) on $\mathbb{R}$.
- **Two-groups model**: $f(y) = \pi f_s(y) + (1-\pi) f_b(y), \quad y \in \mathbb{R}.$
- f_b is a known density function.
- **Unknowns**: Mixing proportion $\pi \in [0,1]$ and pdf f_s ($\neq f_b$).
- **Goals**: Estimate π and f_s (nonparametrically), under certain structural assumptions.

Applications

- In multiple testing problems — the z-scores are normally distributed under H_0 (i.e., f_b is known), while their distribution under H_1 is unknown (Storey [2002], Genovese and Wasserman [2004b], Langaas et al. [2005], Meinshausen and Rice [2006], Efron [2010] ...) where π denotes the proportion of false null hypotheses
- In contamination problems — application in astronomy

Prostate data [Efron (2010)]

- Genetic expression levels for $n = 6033$ genes for $m_1 = 50$ control subjects and $m_2 = 52$ prostate cancer patients

- **Goal**: To discover a small number of "interesting" genes whose expression levels differ between the cancer and control patients

- Such genes, once identified, might be further investigated for a causal link to prostate cancer development

Prostate data [Efron (2010)]

- Genetic expression levels for $n = 6033$ genes for $m_1 = 50$ control subjects and $m_2 = 52$ prostate cancer patients

- **Goal**: To discover a small number of "interesting" genes whose expression levels differ between the cancer and control patients

- Such genes, once identified, might be further investigated for a causal link to prostate cancer development

- The two-sample t-statistic for testing significance of gene i is

$$t_i = \frac{\bar{x}_i(2) - \bar{x}_i(1)}{s_i} \sim t_{100} \qquad [\text{under } H_{0i} : \mu_i(1) = \mu_i(2)],$$

  where s_i is an estimate of the standard error of $\bar{x}_i(1) - \bar{x}_i(2)$.

- Reject H_{0i} if $|t_i| > c_\alpha$ (as $H_{Ai} : \mu_i(1) \neq \mu_i(2)$)

Z-score modeling

- $t_i = \frac{\bar{x}_i(2) - \bar{x}_i(1)}{s_i} \approx Z_i + \frac{\mu_i(2) - \mu_i(1)}{\sigma_i}, \qquad Z_i \sim N(0,1)$ (approx).
- Let $\Delta_i := \frac{\mu_i(2) - \mu_i(1)}{\sigma_i}$ — effect-size.
- Thus, $t_i \sim N(\Delta_i, 1)$ (approx).

Z-score modeling

- $t_i = \frac{\bar{x}_i(2) - \bar{x}_i(1)}{s_i} \approx Z_i + \frac{\mu_i(2) - \mu_i(1)}{\sigma_i}, \qquad Z_i \sim N(0,1)$ (approx).
- Let $\Delta_i := \frac{\mu_i(2) - \mu_i(1)}{\sigma_i}$ — effect-size.
- Thus, $t_i \sim N(\Delta_i, 1)$ (approx).
- Assume that Δ_i's are i.i.d. $(1-\pi)\delta_0 + \pi G$ (G unknown DF).
- Then $t_1, \ldots, t_n$ are i.i.d. (approx) and $t_i \approx Z_i + \Delta_i$:

$$t_i \sim (1-\pi)\phi(\cdot) + \pi \int \phi(\cdot - u) dG(u) = (1-\pi) f_b + \pi f_s$$

  where $f_b := \phi(\cdot)$ and

$$f_s = \int \phi(\cdot - u) dG(u)$$

  is a Gaussian location mixture. See Scott et al. [2015] for a related example.
- We will come back to this model later in the talk.

Regression in a two-component mixture model

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. (X, Y) where

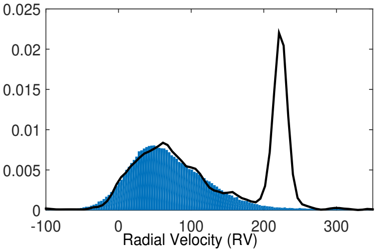
- Y: comes from a two-component mixture model
- X ($\in \mathbb{R}^d$): may provide information about membership

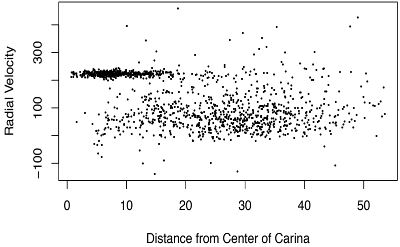
Regression in a two-component mixture model

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. (X, Y) where

- Y: comes from a two-component mixture model
- X $(\in \mathbb{R}^d)$: may provide information about membership

- **Astronomy example** (Walker et al. [2009]): Radial velocity (RV) of stars ($n = 1266$) from Carina (dSph), contaminated by Milky Way stars
- Neural synchrony detection (Scott et al. [2015]); genomic studies (Ignatiadis et al. [2016] ...)



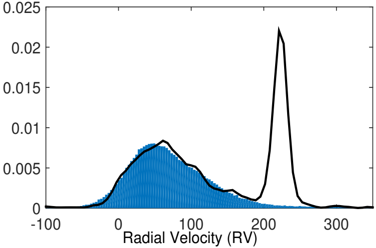


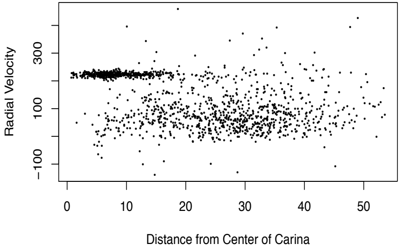
Regression in a two-component mixture model

Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. (X, Y) where

- Y: comes from a two-component mixture model
- X $(\in \mathbb{R}^d)$: may provide information about membership

- **Astronomy example** (Walker et al. [2009]): Radial velocity (RV) of stars ($n = 1266$) from Carina (dSph), contaminated by Milky Way stars
- Neural synchrony detection (Scott et al. [2015]); genomic studies (Ignatiadis et al. [2016] ...)





Question: How do we model the data (i.e., incorporate the covariates)?

Model (Scott et al. [2015], Walker et al. [2009])

- Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ where

$$Y|X = x \sim \pi(x) f_s + (1 - \pi(x)) f_b$$

1. f_b — known pdf on $\mathbb{R}$
2. f_s — unknown pdf on $\mathbb{R}$ belonging to a (non)-parametric class $\mathfrak{F}$
3. $\pi : \mathbb{R}^d \to [0, 1]$ is an unknown (non)-parametric function; $\pi \in \Pi$

Model (Scott et al. [2015], Walker et al. [2009])

- Let $(X_1, Y_1), \ldots, (X_n, Y_n)$ be i.i.d. $(X, Y) \in \mathbb{R}^d \times \mathbb{R}$ where

$$Y|X = x \sim \pi(x) f_s + (1 - \pi(x)) f_b$$

1. f_b — known pdf on $\mathbb{R}$
2. f_s — unknown pdf on $\mathbb{R}$ belonging to a (non)-parametric class $\mathfrak{F}$
3. $\pi : \mathbb{R}^d \to [0, 1]$ is an unknown (non)-parametric function; $\pi \in \Pi$

- Suppose H is the unobserved latent variable (to (X, Y)), i.e.,

$$H = \begin{cases} 1, & \text{if } Y \text{ comes from } f_s \\ 0, & \text{if } Y \text{ comes from } f_b \end{cases}$$

- $H|X = x \sim \text{Bernoulli}(\pi(x))$; $\quad Y|H = 1 \sim f_s$ and $Y|H = 0 \sim f_b$

- Identifiability issues with this model?

Identifiability

- **Two-groups model**: Suppose $\pi \in \Pi :=$ {constant functions in $[0, 1]$} and $f_s \in \mathfrak{F}$ (convex family of densities) — Not identifiable (see Patra and Sen [2016], Genovese and Wasserman [2004a]).

Identifiability

- **Two-groups model**: Suppose $\pi \in \Pi := \{\text{constant functions in } [0,1]\}$ and $f_s \in \mathfrak{F}$ (convex family of densities) — Not identifiable (see Patra and Sen [2016], Genovese and Wasserman [2004a]).

- **Two-groups model with covariates**: Suppose $\pi \in \Pi := \{\text{non-decreasing functions in } [0,1] \text{ bounded above } (< 1)\}$ and $f_s \in \mathfrak{F}$ (family of non-increasing densities) — Identifiable.

Identifiability

- **Two-groups model**: Suppose $\pi \in \Pi := \{\text{constant functions in } [0,1]\}$ and $f_s \in \mathfrak{F}$ (convex family of densities) — Not identifiable (see Patra and Sen [2016], Genovese and Wasserman [2004a]).
- **Two-groups model with covariates**: Suppose $\pi \in \Pi := \{\text{non-decreasing functions in } [0,1] \text{ bounded above } (< 1)\}$ and $f_s \in \mathfrak{F}$ (family of non-increasing densities) — Identifiable.
- **Discrete or continuous covariates**: In general, even for "nice" function classes Π (e.g., logistic function/probbit function), the presence of discrete (say binary) covariates may not restore identifiability.

Identifiability

- **Two-groups model**: Suppose $\pi \in \Pi := \{\text{constant functions in } [0,1]\}$ and $f_s \in \mathfrak{F}$ (convex family of densities) — Not identifiable (see Patra and Sen [2016], Genovese and Wasserman [2004a]).
- **Two-groups model with covariates**: Suppose $\pi \in \Pi := \{\text{non-decreasing functions in } [0,1] \text{ bounded above } (< 1)\}$ and $f_s \in \mathfrak{F}$ (family of non-increasing densities) — Identifiable.
- **Discrete or continuous covariates**: In general, even for "nice" function classes Π (e.g., logistic function/probbit function), the presence of discrete (say binary) covariates may not restore identifiability.
- A general version of identifiability conditions have been presented in the paper.

- **Model**: $Y|X = x \sim \pi(x) f_s + (1 - \pi(x)) f_b,$ f_b known

- **Unknowns**: $\pi \in \Pi$ and $f_s \in \mathfrak{F}$

- Note: $Y|H = 1 \sim f_s$ and $Y|H = 0 \sim f_b$ (H is the latent variable)

Goals

- Estimate $\pi(\cdot)$ and the density $f_s(\cdot)$

- Another important quantity to estimate is the posterior probability of the latent variable being 0 ("null")

$$\mathbb{P}(H = 0|Y, X) = \frac{(1 - \pi(X)) f_b(Y)}{(1 - \pi(X)) f_b(Y) + \pi(X) f_s(Y)}$$

- In multiple testing this is the local false discovery rate $LFDR(\cdot, \cdot)$

- Obtain accurate estimates of $LFDR(\cdot, \cdot)$

- **Model**: $Y|X = x \sim \pi(x) f_s + (1 - \pi(x)) f_b, \qquad f_b$ known
- $\pi \in \Pi$ and $f_s \in \mathfrak{F}$ are unknown

Some natural assumptions on $f_s(\cdot) \in \mathfrak{F}$

- Arbitrary location mixture of unit-variance Gaussians, i.e.,

$$f_s(y) = \int \phi(y - u) dG(u) \qquad (G \text{ unknown DF});$$

  arises in multiple testing problems when modeling the z-scores (where G is the distribution of the nonzero effect sizes)

- Any decreasing density on $[0, 1]$ (useful in modeling p-values)

- **Model**: $Y|X = x \sim \pi(x) f_s + (1 - \pi(x)) f_b,$ f_b known
- $\pi \in \Pi$ and $f_s \in \mathfrak{F}$ are unknown

Some natural assumptions on $f_s(\cdot) \in \mathfrak{F}$

- Arbitrary location mixture of unit-variance Gaussians, i.e.,

$$f_s(y) = \int \phi(y - u) dG(u) \qquad (G \text{ unknown DF});$$

  arises in multiple testing problems when modeling the z-scores (where G is the distribution of the nonzero effect sizes)

- Any decreasing density on $[0, 1]$ (useful in modeling p-values)

Some natural assumptions on $\pi(\cdot) \in \Pi$

- Parametric models, i.e., $\pi(x) = (1 + e^{-\beta^\top x})^{-1}$ (Scott et al. [2015])

- Nonparametric models for $\pi(\cdot)$: monotonicity, regression splines, piecewise constancy (Walker et al. [2009], Scott et al. [2015], Li and Barber [2016])

Estimation: (Nonparametric) Maximum Likelihood

- Suppose $f_s \in \mathfrak{F}$, e.g., $\mathfrak{F} = \{\int \phi(\cdot - u) dG(u) : G \text{ is DF}\}$
- Suppose $\pi \in \Pi$, e.g., $\Pi = \{(1 + e^{-\beta^\top x})^{-1} : \beta \in \mathbb{R}^d\}$
- Denote the log-likelihood by

$$\ell(\pi, f_s) := \sum_{i=1}^{n} \log \Big[(1 - \pi(X_i)) f_b(Y_i) + \pi(X_i) f_s(Y_i)\Big], \quad \pi \in \Pi, f_s \in \mathfrak{F}$$

- Maximum likelihood estimator (MLE):

$$(\hat{\pi}, \hat{f}_s) = \operatorname*{argmax}_{\pi \in \Pi, f_s \in \mathfrak{F}} \ell(\pi, f_s)$$

- Non-convex problem; use EM algorithm (or alternating maximization)

The EM algorithm

The complete data log-likelihood of $\{(X_i, Y_i, H_i)\}_{i=1}^n$ is

$$\sum_{i=1}^{n} \left\{ H_i \log\left[\pi(X_i) f_s(Y_i)\right] + (1 - H_i) \log\left[(1 - \pi(X_i)) f_b(Y_i)\right] \right\}$$

E-step

- As H_i's are unobserved we replace H_i's by their cond. expectations:

$$w_i := \mathbb{E}(H_i | Y_i = y, X_i = x) = \frac{\pi(x) f_s(y)}{\pi(x) f_s(y) + (1 - \pi(x)) f_b(y)}$$

- We plug-in current estimates of f_s and π to obtain $\hat{\mathbf{w}} = (\hat{w}_1, \ldots, \hat{w}_n)$

The EM algorithm

The complete data log-likelihood of $\{(X_i, Y_i, H_i)\}_{i=1}^n$ is

$$\sum_{i=1}^n \Big\{ H_i \log\left[\pi(X_i) f_s(Y_i)\right] + (1 - H_i) \log\left[(1 - \pi(X_i)) f_b(Y_i)\right] \Big\}$$

E-step

- As H_i's are unobserved we replace H_i's by their cond. expectations:

$$w_i := \mathbb{E}(H_i | Y_i = y, X_i = x) = \frac{\pi(x) f_s(y)}{\pi(x) f_s(y) + (1 - \pi(x)) f_b(y)}$$

- We plug-in current estimates of f_s and π to obtain $\hat{\mathbf{w}} = (\hat{w}_1, \ldots, \hat{w}_n)$

M-step

- Due to the particular form of the expected log-likelihood, this joint maximization breaks into two isolated maximization problems:

$$\begin{aligned} \hat{\pi}_{\mathrm{EM}}(\hat{\mathbf{w}}, \Pi) &:= \operatorname*{argmax}_{\pi \in \Pi} \sum \left[\hat{w}_i \log \pi(X_i) + (1 - \hat{w}_i) \log\left(1 - \pi(X_i)\right)\right] \\ \hat{f}_{\mathrm{EM}}(\hat{\mathbf{w}}, \mathfrak{F}) &:= \operatorname*{argmax}_{f_s \in \mathfrak{F}} \sum \hat{w}_i \log f_s(Y_i) \end{aligned}$$

- Suppose $\pi(x) = (1 + e^{-\beta^\top x})^{-1}$; $f_s(y) = \int \phi(y-u)dG(u)$, G is DF

- The logistic likelihood problem can be solved using gradient descent:

$$\hat{\pi}_{\mathrm{EM}}(\hat{\mathbf{w}}, \Pi) = \operatorname*{argmax}_{\pi \in \Pi} \sum \left[\hat{w}_i \log \pi(X_i) + (1 - \hat{w}_i) \log\left(1 - \pi(X_i)\right)\right]$$

- Suppose $\pi(x) = (1 + e^{-\beta^\top x})^{-1}$; $f_s(y) = \int \phi(y-u)dG(u)$, G is DF

- The logistic likelihood problem can be solved using gradient descent:

$$\hat{\pi}_{\mathrm{EM}}(\hat{\mathbf{w}}, \Pi) = \operatorname*{argmax}_{\pi \in \Pi} \sum \left[\hat{w}_i \log \pi(X_i) + (1 - \hat{w}_i) \log \left(1 - \pi(X_i)\right)\right]$$

Solving the Gaussian location mixture problem

- Solving for any arbitrary Gaussian location mixture is a Kiefer-Wolfowitz MLE (Kiefer and Wolfowitz [1956]):

$$\hat{f}_{\mathrm{EM}}(\hat{\mathbf{w}}, \mathfrak{F}) = \operatorname*{argmax}_{f_s = \int \phi(\cdot - u)dG(u), G \text{ is DF}} \sum_{i=1}^{n} \hat{w}_i \log f_s(Y_i)$$

- An infinite dimensional convex program (Lindsay [1995])

- Resulting $\hat{G}$ is supported on at most n points in $\mathrm{ConvexHull}(Y_1, \ldots, Y_n)$

- Can be approximated by optimizing G over discrete distributions with support in a grid in $\mathrm{ConvexHull}(Y_1, \ldots, Y_n)$

- Suppose $\mathfrak{F} = \{\int \phi(y-u)dG(u) : G \text{ is DF}\}$
- Suppose $\pi \in \Pi$, where Π is a VC subgraph class of functions with VC dimension V (e.g., $\Pi = \{(1+e^{-\beta^\top x})^{-1} : \beta \in \mathbb{R}^d\}$)
- **Truth**: $(\pi^0, f_s^0) \in \Pi \times \mathfrak{F}$

- $(\hat{\pi}, \hat{f}_s)$ is s.t. $\ell(\hat{\pi}, \hat{f}_s) \geq \ell(\pi^0, f_s^0)$ (e.g., $(\hat{\pi}, \hat{f}_s) = \underset{\pi\in\Pi, f_s\in\mathfrak{F}}{\operatorname{argmax}}\, \ell(\pi, f_s)$)

- Suppose $\mathfrak{F} = \{\int \phi(y-u)dG(u) : G \text{ is DF}\}$
- Suppose $\pi \in \Pi$, where Π is a VC subgraph class of functions with VC dimension V (e.g., $\Pi = \{(1+e^{-\beta^\top x})^{-1} : \beta \in \mathbb{R}^d\}$)
- **Truth**: $(\pi^0, f_s^0) \in \Pi \times \mathfrak{F}$
- $(\hat{\pi}, \hat{f}_s)$ is s.t. $\ell(\hat{\pi}, \hat{f}_s) \geq \ell(\pi^0, f_s^0)$ (e.g., $(\hat{\pi}, \hat{f}_s) = \operatorname*{argmax}_{\pi\in\Pi, f_s\in\mathfrak{F}} \ell(\pi, f_s)$)

Theorem (Deb, Saha, Guntuboyina and S. (2018))

Letting d_H denote the Hellinger distance, define

$$d^2((\hat{\pi}, \hat{f}_s), (\pi, f_s)) := \frac{1}{n}\sum_{i=1}^{n} d_H^2\left((1-\pi(X_i))f_b + \pi(X_i)f_s, (1-\hat{\pi}(X_i))f_b + \hat{\pi}(X_i)\hat{f}_s\right).$$

If Π has VC dimension V and G is supported on $[-M, M]$,

$$\mathbb{E}\left[d^2((\hat{\pi}, \hat{f}_s), (\pi^0, f_s^0))\right] = \mathcal{O}\left(\frac{M+V}{n}(\log n)^2\right).$$

- Suppose $\mathfrak{F} = \{\int \phi(y-u)dG(u) : G \text{ is DF}\}$
- Suppose $\pi \in \Pi$, where Π is a VC subgraph class of functions with VC dimension V (e.g., $\Pi = \{(1+e^{-\beta^\top x})^{-1} : \beta \in \mathbb{R}^d\}$)
- **Truth**: $(\pi^0, f_s^0) \in \Pi \times \mathfrak{F}$
- $(\hat{\pi}, \hat{f}_s)$ is s.t. $\ell(\hat{\pi}, \hat{f}_s) \geq \ell(\pi^0, f_s^0)$ (e.g., $(\hat{\pi}, \hat{f}_s) = \underset{\pi \in \Pi, f_s \in \mathfrak{F}}{\operatorname{argmax}} \ell(\pi, f_s)$)

Theorem (Deb, Saha, Guntuboyina and S. (2018))

Letting d_H denote the Hellinger distance, define

$$d^2((\hat{\pi}, \hat{f}_s), (\pi, f_s)) := \frac{1}{n}\sum_{i=1}^{n} d_H^2\left((1-\pi(X_i))f_b + \pi(X_i)f_s, (1-\hat{\pi}(X_i))f_b + \hat{\pi}(X_i)\hat{f}_s\right).$$

If Π has VC dimension V and G is supported on $[-M, M]$,

$$\mathbb{E}\left[d^2((\hat{\pi}, \hat{f}_s), (\pi^0, f_s^0))\right] = \mathcal{O}\left(\frac{M+V}{n}(\log n)^2\right).$$

- Almost parametric (n^{-1}) rate of convergence
- Implications in estimating (denominator of) the posterior $LFDR(\cdot, \cdot)$
- The model need not be identifiable for the result to hold

Marginal Method – II

- Recall: $Y|X = x \sim (1 - \pi(x))f_b + \pi(x)f_s, \qquad f_b$ known
- Regression of Y on X: $\mathbb{E}(Y|X = x) = (1 - \pi(x))\mu_b + \pi(x)\mu_s$

Marginal Method – II

- Recall: $Y|X = x \sim (1-\pi(x))f_b + \pi(x)f_s$, $\quad f_b$ known
- Regression of Y on X: $\mathbb{E}(Y|X=x) = (1-\pi(x))\mu_b + \pi(x)\mu_s$
- Whenever $\mathbb{E}_{Y\sim f_b}[Y] =: \mu_b \neq \mu_s := \mathbb{E}_{Y\sim f_s}[Y]$, this poses a (non-linear) regression problem (μ_b known, μ_s unknown):

$$(\hat{\pi}, \hat{\mu}_s) := \operatorname*{argmin}_{\pi\in\Pi, \mu_s\in\mathbb{R}} \sum_{i=1}^{n} \Big(Y_i - \mu_b - \pi(X_i)(\mu_s - \mu_b) \Big)^2$$

- Once $\hat{\pi}(\cdot)$ is estimated,

$$\hat{f}_s := \operatorname*{argmax}_{f_s\in\mathfrak{F}} \sum_{i=1}^{n} \log \Big[(1-\hat{\pi}(X_i)) f_b(Y_i) + \hat{\pi}(X_i) f_s(Y_i) \Big]$$

can be solved using the Kiefer-Wolfowitz MLE

Marginal Method – I

- Recall: $Y|X = x \sim (1 - \pi(x))f_b + \pi(x)f_s, \qquad f_b$ known
- Denote $\bar{\pi} := \mathbb{E}_X[\pi(X)]$, overall proportion of non-nulls (signals)
- Observe that marginally, $\qquad Y \sim (1 - \bar{\pi})f_b + \bar{\pi} f_s$

Marginal Method – I

- Recall: $Y|X = x \sim (1-\pi(x))f_b + \pi(x)f_s,\qquad f_b$ known
- Denote $\bar{\pi} := \mathbb{E}_X[\pi(X)]$, overall proportion of non-nulls (signals)
- Observe that marginally, $\qquad Y \sim (1-\bar{\pi})f_b + \bar{\pi}f_s$

When $\bar{\pi}$ is known (problem can be solved easily)

- Maximize the marginal likelihood of Y (Kiefer-Wolfowitz MLE):

$$\hat{f}_s = \operatorname*{argmax}_{f_s \in \mathfrak{F}} \sum_{i=1}^{n} \log\left[(1-\bar{\pi})f_b(Y_i) + \bar{\pi}f_s(Y_i)\right]$$

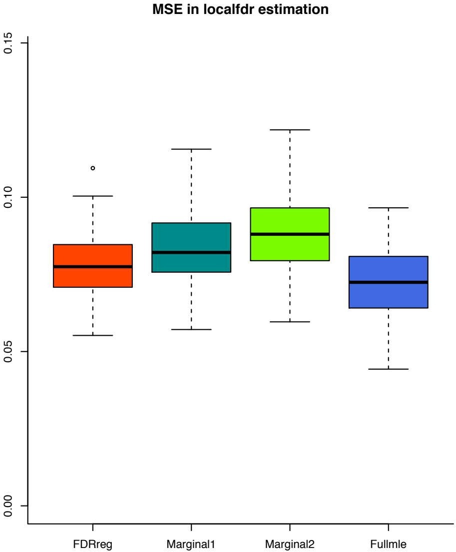
- Maximize the joint likelihood of (X, Y) with $\hat{f}_s$ fixed:

$$\hat{\pi} = \operatorname*{argmax}_{\pi \in \Pi} \sum_{i=1}^{n} \log\left[(1-\pi(X_i))f_b(Y_i) + \pi(X_i)\hat{f}_s(Y_i)\right]$$

- Can take a grid of $\bar{\pi}$ values in practice and choose the one with the highest likelihood

Marginal Methods

- They are computationally simpler and faster.
- They are reasonably accurate (the `fullmle` approach mostly outperforms them).
- They provide good starting points for `fullmle`.



A setting from Scott et al. [2015]:

$$X = (X_1, X_2) \sim (U(0,1), U(0,1))$$
$$\pi(x) = \frac{1}{1 + e^{3 - 1.5x_1 - 1.5x_2}}$$
$$f_s = 0.48\mathcal{N}(\pm 2, 2) + 0.04\mathcal{N}(0, 17)$$
$$Y|X = x \sim (1 - \pi(x))\mathcal{N}(0,1) + \pi(x) f_s(\cdot)$$
$$n = 10000$$

Plot compares the MSEs in estimating the LFDRs at data points for 3 methods[1,2] and FDRreg, a method in Scott et al. [2015].

[1]The `fullmle` method used starting points obtained from the marginal methods
[2]Marginal II method was fitted using the parametric regression of $|Y|$ on X

Summary

- A maximum likelihood procedure that incorporates covariate information in a (nonparametric) two-component mixture model .
- Used NP mixture models to estimate the unknown f_s.

Summary

- A maximum likelihood procedure that incorporates covariate information in a (nonparametric) two-component mixture model .
- Used NP mixture models to estimate the unknown f_s.
- Although our approach is nonparametric, our methods avoid the need to specify tuning parameter(s).
- Almost parametric rate of estimation.

Summary

- A maximum likelihood procedure that incorporates covariate information in a (nonparametric) two-component mixture model .
- Used NP mixture models to estimate the unknown f_s.
- Although our approach is nonparametric, our methods avoid the need to specify tuning parameter(s).
- Almost parametric rate of estimation.
- NPMLE in mixture models deserves more attention.
- See "Two-component Mixture Model in the Presence of Covariates" —Nabarun Deb, Sujayam Saha, Adityanand Guntuboyina and Bodhisattva Sen, at https://arxiv.org/pdf/1810.07897.pdf and the associated R package at https://github.com/NabarunD/NPMLEmix.

Summary

- A maximum likelihood procedure that incorporates covariate information in a (nonparametric) two-component mixture model .
- Used NP mixture models to estimate the unknown f_s.
- Although our approach is nonparametric, our methods avoid the need to specify tuning parameter(s).
- Almost parametric rate of estimation.
- NPMLE in mixture models deserves more attention.
- See "Two-component Mixture Model in the Presence of Covariates" —Nabarun Deb, Sujayam Saha, Adityanand Guntuboyina and Bodhisattva Sen, at https://arxiv.org/pdf/1810.07897.pdf and the associated R package at https://github.com/NabarunD/NPMLEmix.

Thank You! Questions?

References I

Bradley Efron. *Large-scale inference*, volume 1 of *Institute of Mathematical Statistics (IMS) Monographs*. Cambridge University Press, Cambridge, 2010. ISBN 978-0-521-19249-1. doi: 10.1017/CBO9780511761362. URL http://dx.doi.org.ezproxy.cul.columbia.edu/10.1017/CBO9780511761362. Empirical Bayes methods for estimation, testing, and prediction.

C. Genovese and L. Wasserman. A stochastic process approach to false discovery control. *Ann. Statist.*, 32(3):1035–1061, 2004a. ISSN 0090-5364. doi: 10.1214/009053604000000283. URL http://dx.doi.org/10.1214/009053604000000283.

Christopher Genovese and Larry Wasserman. A stochastic process approach to false discovery control. *Annals of Statistics*, pages 1035–1061, 2004b.

Nikolaos Ignatiadis, Bernd Klaus, Judith B Zaugg, and Wolfgang Huber. Data-driven hypothesis weighting increases detection power in genome-scale multiple testing. *Nature methods*, 13(7):577–580, 2016.

J. Kiefer and J. Wolfowitz. Consistency of the maximum likelihood estimator in the presence of infinitely many incidental parameters. *Ann. Math. Statist.*, 27:887–906, 1956. ISSN 0003-4851. doi: 10.1214/aoms/1177728066. URL http://dx.doi.org/10.1214/aoms/1177728066.

Mette Langaas, Bo Henry Lindqvist, and Egil Ferkingstad. Estimating the proportion of true null hypotheses, with application to dna microarray data. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 67(4):555–572, 2005.

Ang Li and Rina Foygel Barber. Multiple testing with the structure adaptive benjamini-hochberg algorithm. *arXiv preprint arXiv:1606.07926*, 2016.

B. G. Lindsay. Mixture models: Theory, geometry and applications. *NSF-CBMS Regional Conference Series in Probability and Statistics*, 5:1–163, 1995. ISSN 19355920. URL http://www.jstor.org/stable/4153184.

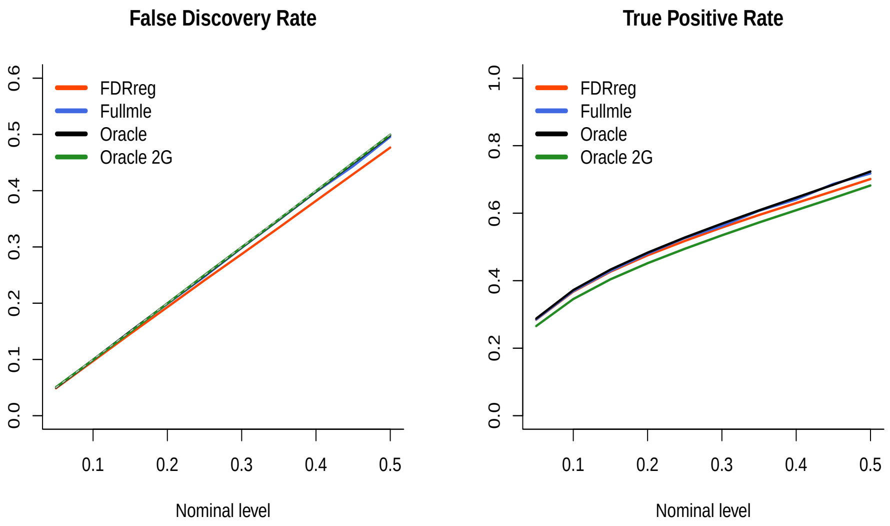
N. Meinshausen and J. Rice. Estimating the proportion of false null hypotheses among a large number of independently tested hypotheses. *Ann. Statist.*, 34(1):373–393, 2006. ISSN 0090-5364. doi: 10.1214/009053605000000741. URL http://dx.doi.org/10.1214/009053605000000741.

Rohit Patra and Bodhisattva Sen. Estimation of a two-component mixture model with applications to multiple testing. *J. R. Stat. Soc. Ser. B. Stat. Methodol.*, 78(4):869–893, 2016. ISSN 1369-7412. doi: 10.1111/rssb.12148. URL http://dx.doi.org/10.1111/rssb.12148.

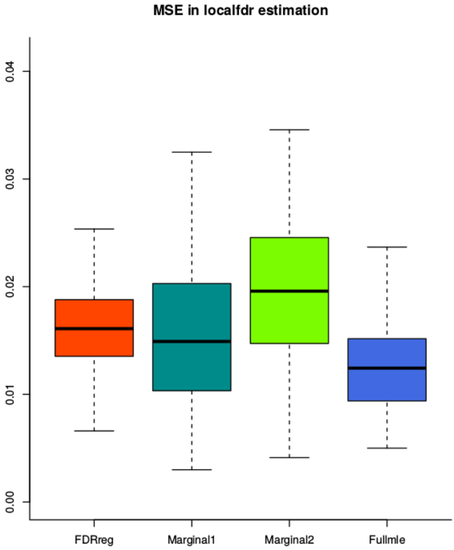
James G Scott, Ryan C Kelly, Matthew A Smith, Pengcheng Zhou, and Robert E Kass. False discovery rate regression: an application to neural synchrony detection in primary visual cortex. *Journal of the American Statistical Association*, 110(510):459–471, 2015.

J. D. Storey. A direct approach to false discovery rates. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 64(3):479–498, 2002. ISSN 1369-7412. doi: 10.1111/1467-9868.00346. URL http://dx.doi.org/10.1111/1467-9868.00346.

Matthew G Walker, Mario Mateo, Edward W Olszewski, Bodhisattva Sen, and Michael Woodroofe. Clean kinematic samples in dwarf spheroidals: An algorithm for evaluating membership and estimating distribution parameters when contamination is present. *The Astronomical Journal*, 137(2):3109, 2009.



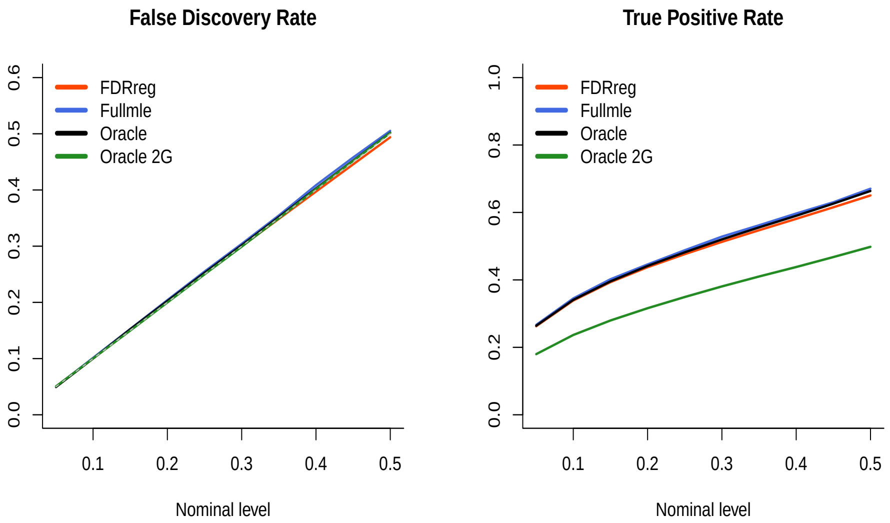
The observed FDR and true positive rate for the `fullmle` and FDRreg methods. We also compare with the "oracle" (that knows the true f_s^0 and π^0), and also the "oracle" ignoring the covariates.



Another setting from Scott et al. [2015]:

$$
\begin{aligned}
X &= (X_1, X_2) \sim (U(0,1), U(0,1)) \\
\pi(x) &= \frac{1}{1 + e^{-3.25+3.5x_1^2-3.5x_2^2}} \\
f_s &= 0.48\mathcal{N}(\pm 2, 2) + 0.04\mathcal{N}(0, 17) \\
Y|X = x &\sim (1-\pi(x))\mathcal{N}(0,1) + \pi(x) f_s(\cdot) \\
n &= 10000
\end{aligned}
$$

Plot compares the mean squared errors (MSEs) in estimating the LFDRs at data points for the 3 methods and FDRreg, a method in Scott et al. [2015].



The observed FDR (and true positive rate) for the `fullmle` and FDRreg methods. We also compare with the "oracle" (that knows the true f_s and π), and also the "oracle" (true) ignoring the covariates.