Estimation of a Two-component Mixture Model

Bodhisattva Sen[1,2]
University of Cambridge, Cambridge, UK
Columbia University, New York, USA

Indian Statistical Institute, Kolkata, India

6 August, 2012

[1] Joint work with Rohit Patra, Columbia University, USA

[2] Supported by NSF grants DMS-0906597, AST-1107373, DMS-1150435 (CAREER)

Mixture model with two components

$$F(x) = \alpha F_s(x) + (1-\alpha)F_b(x) \tag{1}$$

- F_b is a *known* distribution function (DF).
- *Unknowns*: mixing proportion $\alpha \in (0,1)$ and DF F_s $(\neq F_b)$.
- Problem: Given a random *sample* from $X_1, X_2, \ldots, X_n \overset{ind.}{\sim} F$, we wish to (nonparametrically) estimate F_s and the parameter α.

Previous work

- Most of the previous work on this problem assume some constraint (*parametric* models) on the form of F_s; see e.g., Lindsay (AoS, 1983), Lindsay and Basak (JASA, 1993).
- Bordes et al. (AoS, 2006) assume that both the components belong to an unknown *symmetric location-shift family*.
- In the *multiple testing* setup, this problem has been addressed mostly to estimate α under suitable assumptions; see e.g., Storey (JRSSB, 2002), Genevese & Wasserman (AoS, 2004), Meinshausen & Rice (AoS, 2006).

Applications

- In *multiple* testing problems, e.g., microarray analysis, neuroimaging (fMRI).
- Any situation where numerous *independent hypotheses tests* are performed.
- The p-values are *uniformly* distributed on [0,1], under H_0, while their distribution associated with H_1 is *unknown*; see e.g., Efron (2010).
- Estimate the *proportion of false null hypotheses* α; also needed to control multiple error rates, such as the FDR of Benjamini & Hochberg (JRSSB, 1995).
- In *contamination* problems – application in astronomy.

An application in astronomy

- We analyse the *radial velocity* (RV; line of sight velocity) distribution of stars in *Carina*, a dwarf spheroidal (*dSph*) galaxy.
- The data have been obtained by Magellan and MMT telescopes and consist of radial velocity measurements for $n = 1215$ stars from Carina, *contaminated* with Milky Way stars in the field of view.
- We would like to understand the *distribution* of the line of sight velocity of stars in Carina.
- For the contaminating stars from the Milky Way in the field of view, we assume a *non-Gaussian* velocity distribution F_b that is known from the Besancon Milky Way model (Robin et. al, 2003), calculated along the line of sight to Carina.
- Here α is the *proportion* of stars from Carina.

Outline

1. Estimation of α
2. Lower bounds for α
3. Estimating of F_s and its density f_s

When α is known

- We observe an i.i.d. sample from $F = \alpha F_s + (1-\alpha)F_b$.
- A *naive* estimator of F_s would be

$$\hat{F}^{\alpha}_{s,n} = \frac{\mathbb{F}_n - (1-\alpha)F_b}{\alpha},$$

where $\mathbb{F}_n$ is the empirical DF of the observed sample.

- $\hat{F}^{\alpha}_{s,n}$ is *not* a valid DF: need not be *non-decreasing* or *lie* in $[0,1]$.

- Find the *closest* DF: impose the *known* shape constraint of *monotonicity*, accomplished by minimizing

$$\int \{W(x) - \hat{F}^{\alpha}_{s,n}(x)\}^2 \, d\mathbb{F}_n(x) \equiv \frac{1}{n}\sum_{i=1}^{n}\{W(X_i) - \hat{F}^{\alpha}_{s,n}(X_i)\}^2$$

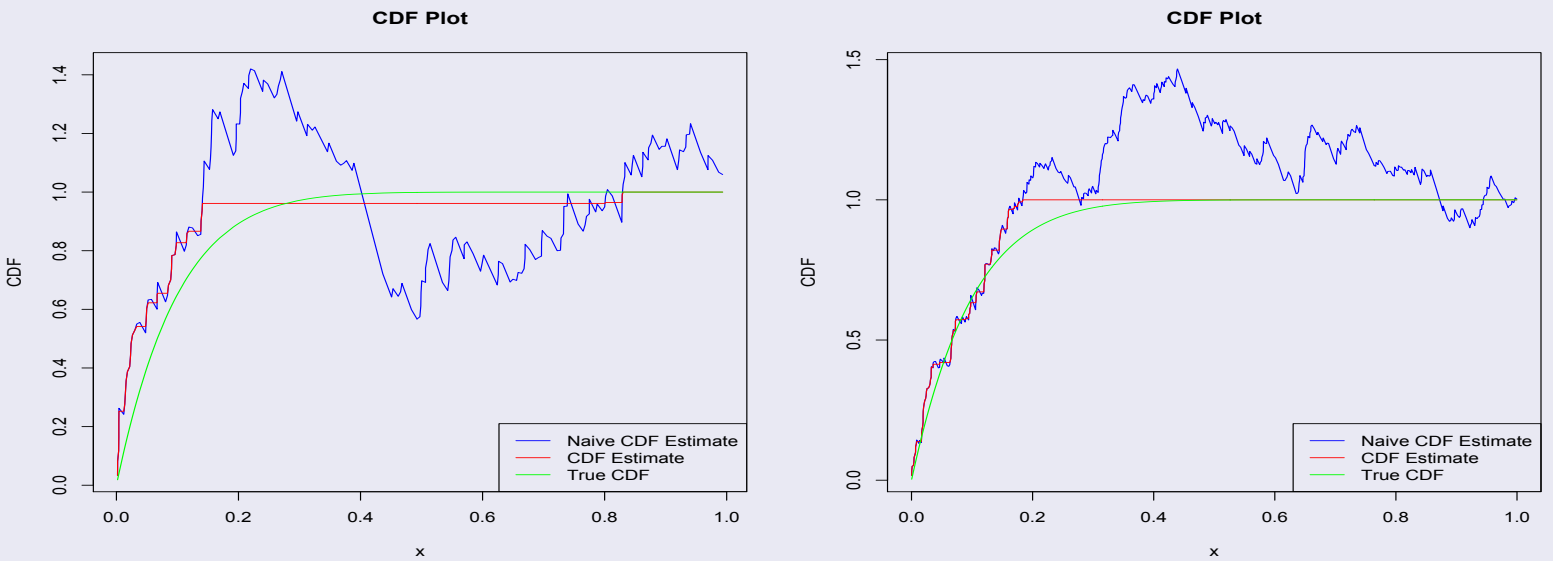
over all DFs W.

Isotonized estimator

- Let

$$\check{F}^{\alpha}_{s,n} = \arg \min_{W \text{ DF}} \frac{1}{n} \sum_{i=1}^{n} \{W(X_i) - \hat{F}^{\alpha}_{s,n}(X_i)\}^2.$$

- $\check{F}^{\alpha}_{s,n}$ is the *isotonized* estimator; *uniquely* defined at the data points X_i, for all $i = 1, \ldots, n$.



Plot of $\hat{F}^{\check{\alpha}_n}_{s,n}$, $\check{F}^{\check{\alpha}_n}_{s,n}$ and F_s for $n = 300$ and 500, when $\alpha = 0.3$.

Computation

- Recall

$$\check{F}^{\alpha}_{s,n} = \arg \min_{W\uparrow \text{ DF}} \frac{1}{n} \sum_{i=1}^{n} \{\hat{F}^{\alpha}_{s,n}(X_{(i)}) - W(X_{(i)})\}^2.$$

- The above optimization problem is the same as minimizing

$$\|\mathbf{V} - \theta\|^2 \text{ over } \theta = (\theta_1, \ldots, \theta_n) \in \Theta_{inc}$$

where $\Theta_{inc} = \{\theta \in \mathbb{R}^n : 0 \le \theta_1 \le \theta_2 \le \ldots \le \theta_n \le 1\}$,
$\mathbf{V} = (V_1, V_2, \ldots, V_n)$, $V_i := \hat{F}^{\alpha}_{s,n}(X_{(i)})$, $i = 1, 2, \ldots, n$.

- The estimator $\hat{\theta}$ is *uniquely* defined by the *projection theorem*; it is the L_2 projection of $\mathbf{V}$ on a *closed convex cone* in $\mathbb{R}^n$.

- Can easily be computed using the *pool-adjacent-violators* algorithm (PAVA); see Section 1.2 of Robertson et al. (1988).

Identifiability

- When α is *unknown*, the problem is *non-identifiable*.
- If $F = \alpha F_s + (1-\alpha)F_b$ for some F_b (known) and α (unknown), then the mixture model can be re-written as

$$F = (\alpha+\gamma)\left(\frac{\alpha}{\alpha+\gamma}F_s + \frac{\gamma}{\alpha+\gamma}F_b\right) + (1-\alpha-\gamma)F_b,$$

  for $0 \leq \gamma \leq 1-\alpha$, and the term $(\alpha F_s + \gamma F_b)/(\alpha+\gamma)$ can be thought of as the *nonparametric* component.
- A trivial solution occurs when we take $\alpha + \gamma = 1$, in which case $\check{F}^1_{s,n} = \hat{F}^1_{s,n} = \mathbb{F}_n$.
- Hence, α is *not* uniquely defined.

Identifiability

- We redefine the mixing proportion as

$$\alpha_0 := \inf \left\{ \gamma \in (0,1) : \frac{F - (1-\gamma)F_b}{\gamma} \text{ is a valid DF} \right\}.$$

- Intuitively, this definition makes sure that the "signal" distribution F_s does *not* include any contribution from the known "background" F_b.
- We consider the estimation of α_0 as defined above.

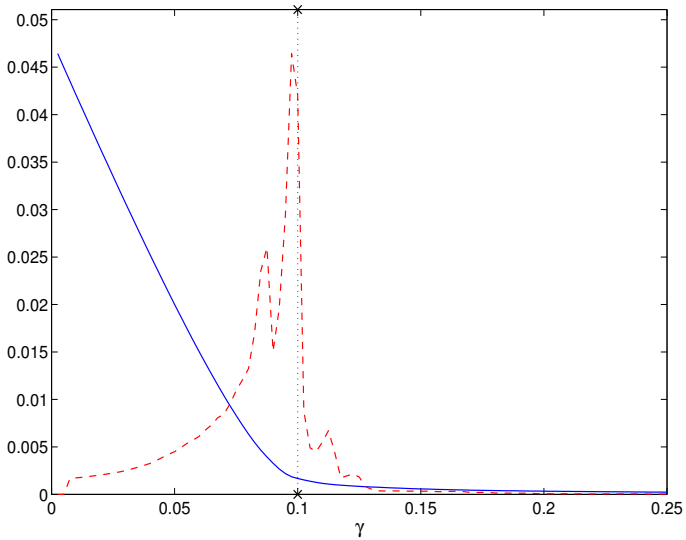
Lemma

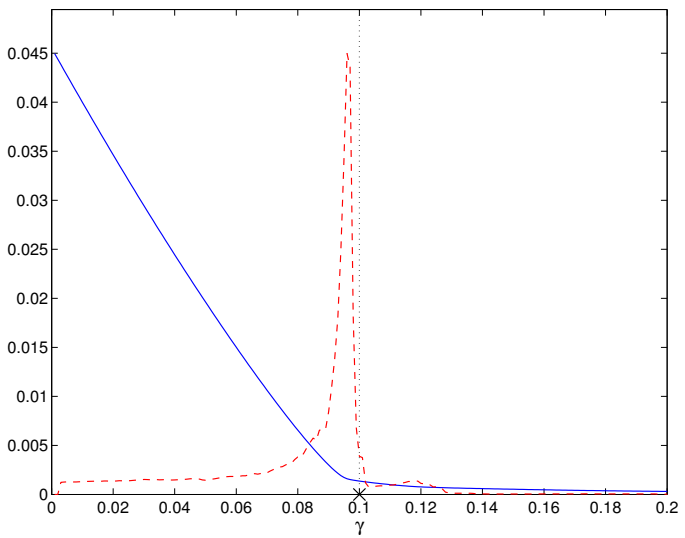
Suppose that F_s and F_b are absolutely continuous, i.e., they have densities f_s and f_b, respectively. Then $\alpha_0 < \alpha$ iff there exists $c > 0$ such that $f_s(x) \geq c f_b(x)$, for all $x \in \mathbb{R}$.

- If the *support* of F_s is *strictly contained* in that of F_b, then the problem is *identifiable*.

Estimation of α_0

- When $\gamma = 1$, $\hat{F}^{\gamma}_{s,n} = \mathbb{F}_n = \check{F}^{\gamma}_{s,n}$.
- When γ is much *smaller* than α_0, the regularization of $\hat{F}^{\gamma}_{s,n}$ modifies it, and thus $\hat{F}^{\gamma}_{s,n}$ and $\check{F}^{\gamma}_{s,n}$ are quite *different*.
- Measure distance by d_n – the $L_2(\mathbb{F}_n)$ distance, i.e., if $g, h : \mathbb{R} \to \mathbb{R}$ are two functions, then $d_n(g,h) = \sqrt{\int \{g(x) - h(x)\}^2 \, d\mathbb{F}_n(x)}$.





Plot of $\gamma \; d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ (in solid blue) when $\alpha_0 = 0.1$ and $n = 5000$.

- We will study $\gamma\ d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n}) = d_n(\mathbb{F}_n, \gamma\check{F}^{\gamma}_{s,n} + (1-\gamma)F_b)$.

Lemma

$\gamma\ d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ is a non-increasing convex function of γ in $(0,1)$.

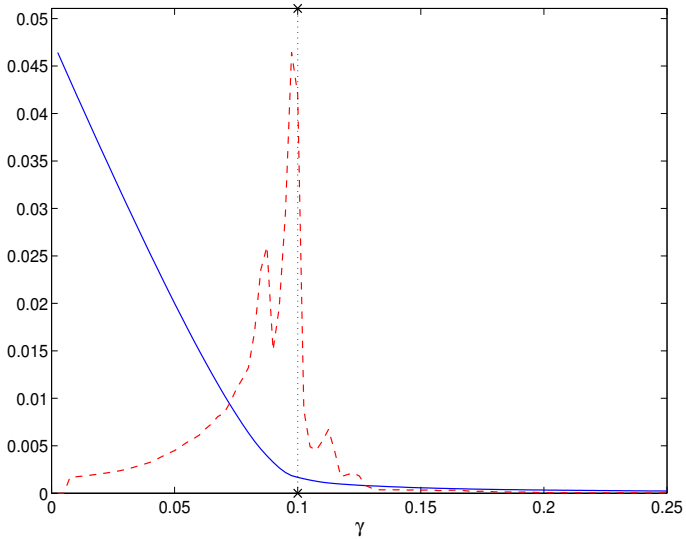
Lemma

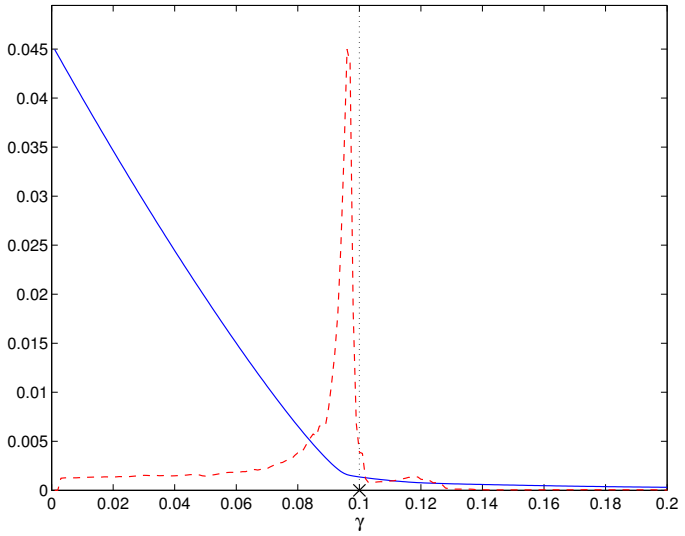
For $1 \geq \gamma \geq \alpha_0$, $\gamma\ d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n}) \leq d_n(F, \mathbb{F}_n)$. Thus,

$$\gamma\ d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n}) \overset{a.s.}{\to} \begin{cases} 0, & \gamma \geq \alpha_0, \\ > 0, & \gamma < \alpha_0. \end{cases}$$

- Note, $\gamma\ d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n}) \leq \gamma\ d_n\left(\hat{F}^{\gamma}_{s,n}, \frac{F-(1-\gamma)F_b}{\gamma}\right)$ from which

$$\gamma\ d_n\left(\frac{\mathbb{F}_n - (1-\gamma)F_b}{\gamma}, \frac{F-(1-\gamma)F_b}{\gamma}\right) = d_n(F, \mathbb{F}_n).$$





Estimation of α_0

- We would like to compare $\hat{F}^{\gamma}_{s,n}$ and $\check{F}^{\gamma}_{s,n}$, and choose the *smallest* γ for which their distance is still *small*.
- Let

$$\hat{\alpha}_n = \inf \left\{ \gamma \in (0,1] : \gamma \; d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n}) \leq \frac{c_n}{\sqrt{n}} \right\}. \tag{2}$$

Theorem (Consistency of $\hat{\alpha}_n$)

If $c_n/\sqrt{n} \to 0$ and $c_n \to \infty$, then $\hat{\alpha}_n \xrightarrow{P} \alpha_0$.

Lower bound for α_0

- Construct a *finite sample* (honest) *lower confidence bound* $\hat{\alpha}_n$ with the property

$$P(\alpha_0 \geq \hat{\alpha}_n) \geq 1 - \beta, \qquad \text{for all } n, \tag{3}$$

  for a specified confidence level $(1 - \beta)$, $0 < \beta < 1$.

- Would allow one to assert, with a specified level of confidence, that the proportion of "signal" is *at least* $\hat{\alpha}_n$.

- It can also be used to *test* the hypothesis that there is *no "signal"* at level β by rejecting when $\hat{\alpha}_n > 0$.

- Genovese & Wasserman (AoS, 2004) and Meinshausen & Rice (AoS, 2006) usually yield *approximate* conservative lower bounds.

Theorem

Let H_n be the DF of $\sqrt{n}d_n(\mathbb{F}_n, F) := \sqrt{n \int \{\mathbb{F}_n(x) - F(x)\}^2 \, d\mathbb{F}_n(x)}$. Let $\hat{\alpha}_n$ be defined as in (2) with c_n defined as the $(1-\beta)$-*quantile* of H_n. Then

$$P(\alpha_0 \geq \hat{\alpha}_n) \geq 1 - \beta, \qquad \text{for all } n.$$

Proof of Theorem

$$\begin{aligned} P(\alpha_0 < \hat{\alpha}_n) &= P\left(\alpha_0 \; d_n(\hat{F}_{s,n}^{\alpha_0}, \check{F}_{s,n}^{\alpha_0}) \geq \frac{c_n}{\sqrt{n}}\right) \\ &\leq P\left(\alpha_0 \; d_n(\hat{F}_{s,n}^{\alpha_0}, F_s^{\alpha_0}) \geq \frac{c_n}{\sqrt{n}}\right) \\ &= P\left(\sqrt{n} \; d_n(\mathbb{F}_n, F) \geq c_n\right) \\ &= 1 - H_n(c_n) \\ &= \beta, \end{aligned}$$

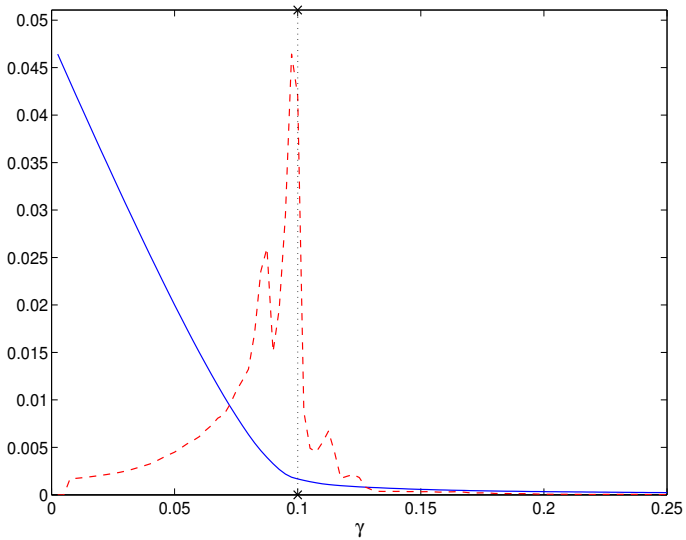
as $\alpha_0 d_n(\hat{F}_{s,n}^{\alpha_0}, F_s^{\alpha_0}) = \alpha_0 d_n\left(\frac{\mathbb{F}_n - (1-\alpha_0)F_b}{\alpha_0}, \frac{F-(1-\alpha_0)F_b}{\alpha_0}\right) = d_n(\mathbb{F}_n, F).$

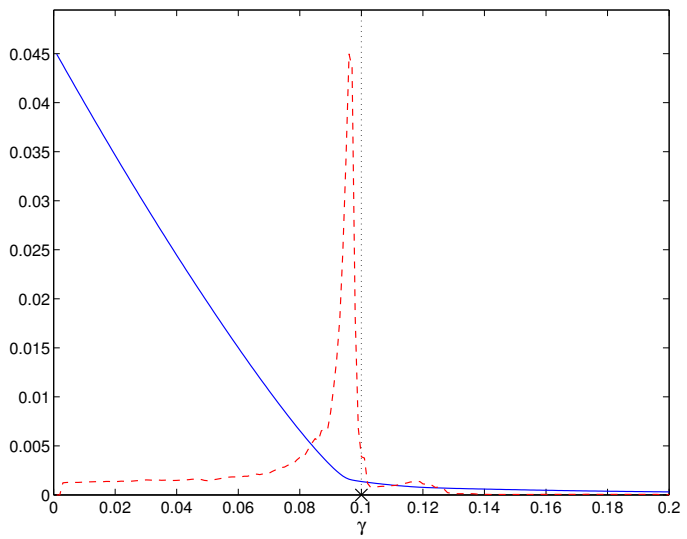
- H_n is *distribution-free* and can be easily simulated.
- Requires *no tuning* parameters.
- Lower bound holds *for all n*.
- For moderately large n (e.g., $n \geq 500$) the distribution H_n can be very well *approximated* by that of the (square-root of) *Cramér-von Mises statistic*, defined as
$\sqrt{n}d(\mathbb{F}_n, F) := \sqrt{n \int \{\mathbb{F}_n(x) - F(x)\}^2 dF(x)}$.

Theorem

Letting G_n to be the DF of $\sqrt{n}d(\mathbb{F}_n, F)$, we have

$$\sup_{x \in \mathbb{R}} |H_n(x) - G_n(x)| \to 0 \text{ as } n \to \infty.$$





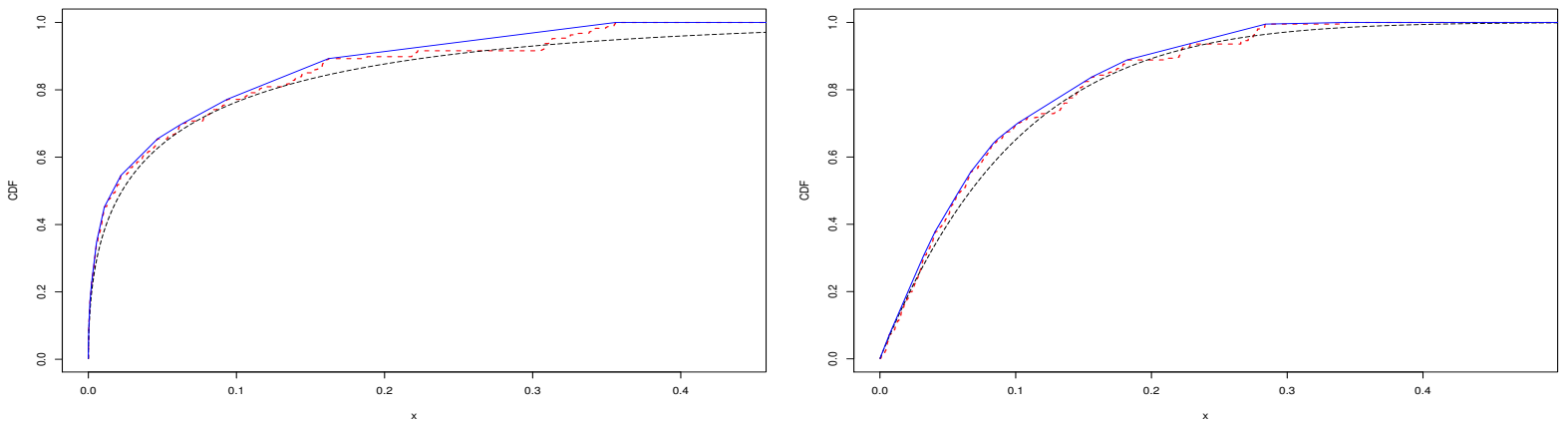
Plot of $\gamma d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ (in solid blue) overlaid with its (scaled) *second derivative* (in dashed red) for $\alpha_0 = 0.1$ and $n = 5000$.

A tuning-parameter-free estimator of α_0

- We can use the *elbow* of $\gamma\ d_n(\hat{F}^{\gamma}_{s,n}, \check{F}^{\gamma}_{s,n})$ to estimate α_0.
- It is the point that has the *maximum curvature*, i.e., the point where the second derivative is maximum.

Estimation of F_s

- Assume now that the model (1) is *identifiable*, i.e., $\alpha = \alpha_0$, and $\check{\alpha}_n$ be an estimator of α_0 ($\check{\alpha}_n$ can be $\hat{\alpha}_n$).
- A natural nonparametric estimator of F_s is $\check{F}_{s,n}^{\check{\alpha}_n}$.



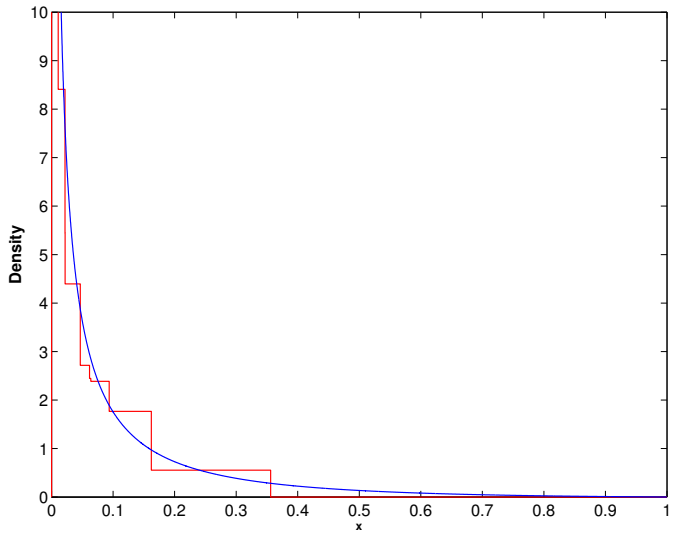
Plot of $\check{F}_{s,n}^{\check{\alpha}_n}$ (in dotted red), F_s (in dashed black) and $F_{s,n}^{\dagger}$ (in blue).

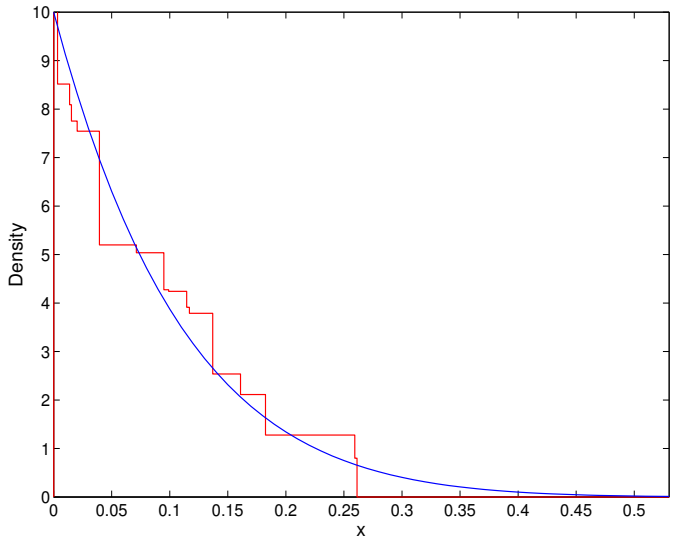
Theorem

Suppose that $\check{\alpha}_n \xrightarrow{P} \alpha_0$. Then, $\sup_{x\in\mathbb{R}} |\check{F}_{s,n}^{\check{\alpha}_n}(x) - F_s(x)| \xrightarrow{P} 0$.

Estimating the density f_s

- Suppose now that F_s has a *non-increasing* density f_s (w.l.o.g., we assume that f_s is non-increasing on $[0, \infty)$).
- Natural assumption in many situations, see e.g., Genovese & Wasserman (AoS, 2004).
- Define $F^{\dagger}_{s,n} := LCM[\check{F}^{\check{\alpha}_n}_{s,n}]$, where for a bounded function $g : [0, \infty) \to \mathbb{R}$, let us represent the *least concave majorant* (LCM) of g by $LCM[g]$.
- $F^{\dagger}_{s,n}$ is a valid DF with a *density* $f^{\dagger}_{s,n}$.
- We can now estimate f_s by $f^{\dagger}_{s,n} = [F^{\dagger}_{s,n}]'$.





Plot of the estimate $f^{\dagger}_{s,n}$ (in solid red) and f_s (in solid blue).

Theorem

Assume F_s has *non-increasing* density f_s on $[0, \infty)$. If $\check{\alpha}_n \xrightarrow{P} \alpha_0$, then,

$$\sup_{x \in \mathbb{R}} |F^{\dagger}_{s,n}(x) - F_s(x)| \xrightarrow{P} 0.$$

Further, if for any $x > 0$, $f_s(x)$ is *continuous* at x, then,

$$f^{\dagger}_{s,n}(x) \xrightarrow{P} f_s(x).$$

Multiple testing

- Estimate the *proportion* of genes that are *differentially* expressed in DNA microarray experiments.
- We wish to test n null hypothesis $H_{01}, H_{02}, \ldots, H_{0n}$ using p-values $X_1, X_2, \ldots, X_n$.
- FWER = Prob (# of false rejections ≥ 1), the probability of committing at least one type I error.
- Benjamini and Hochberg (1995) proposed $FDR = E\left\{\frac{V}{R}\mathbf{1}(R > 0)\right\}$, where V is the number of *false rejections* and R is the number of *total rejections*.
- The BH method guarantees $FDR \leq \beta\alpha_0$; an estimate of α_0 can be used to yield a procedure with FDR approximately equal to β and thus will result in an *increased* power.

Application to multiple testing

- Our method can be used to yield an estimator of α_0.
- We can also obtain a *completely nonparametric* estimator of F_s, the distribution of the p-values arising from the alternative hypotheses.
- The *local false discovery rate* (LFDR) is defined as the function $I : (0, 1) \to [0, \infty)$, such where

$$I(x) = P(H_i = 0 | X_i = x) = \frac{(1 - \alpha_0) f_b(x)}{f(x)}.$$

- The LFDR method can help detect "*interesting*" cases (e.g., $I(x) \leq 0.20$); see Section 5 of Efron (2010).
- If f_s is a assumed to have a *non-increasing* density, we have a natural *tuning-parameter-free* estimator $\hat{I}$ of the LFDR: $\hat{I}(x) = \frac{(1-\check{\alpha}_n) f_b(x)}{\check{\alpha}_n f_{s,n}^{\dagger}(x) + (1-\check{\alpha}_n) f_b(x)}$, for $x \in (0, 1)$.

Simulation: lower bounds

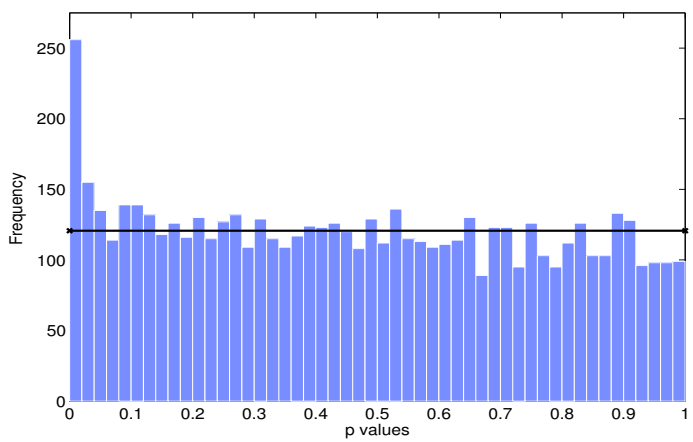
- *Compare* our method with that of Genevose & Wasserman (AoS, 2004), and Meinshausen & Rice (AoS, 2006).
- The method in Meinshausen & Rice (AoS, 2006) are extensions of those proposed in Genevose & Wasserman (AoS, 2004).

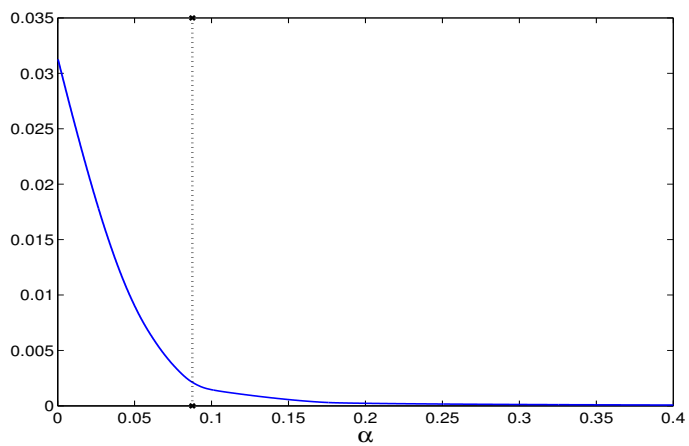
Table: Coverage probabilities of nominal 95% lower confidence bounds for the three methods when $n = 1000$.

α	Setting I			Setting II		
	$\hat{\alpha}_L^0$	$\hat{\alpha}_L^{GW}$	$\hat{\alpha}_L^{MR}$	$\hat{\alpha}_L^0$	$\hat{\alpha}_L^{GW}$	$\hat{\alpha}_L^{MR}$
0	*0.95*	0.98	0.93	*0.95*	0.98	0.93
0.01	0.97	0.98	0.99	0.97	0.97	0.99
0.03	0.98	0.98	0.99	0.98	0.98	0.99
0.05	0.98	0.98	0.99	0.98	0.98	0.99
0.10	0.99	0.99	1.00	0.99	0.98	0.99

Real example: Prostate data

- *Genetic expression* levels for $n = 6033$ genes for $m_1 = 50$ normal 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.

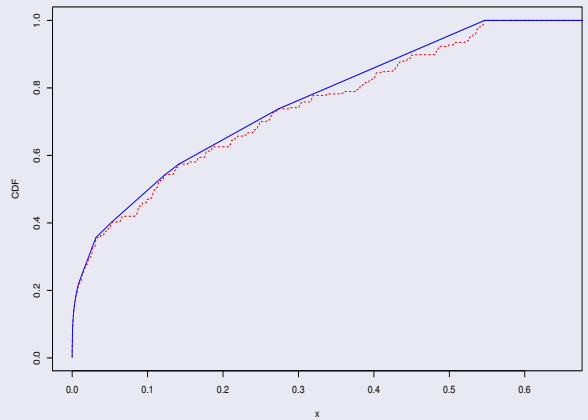




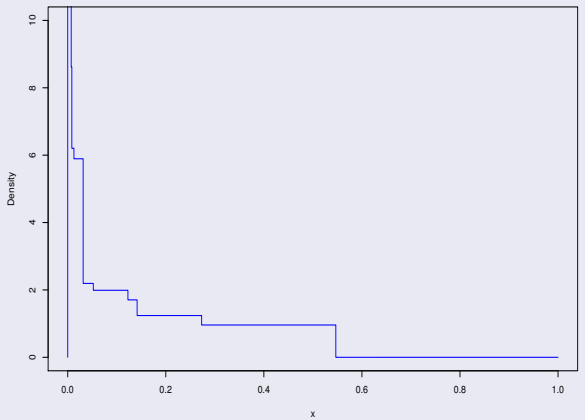
- The *two-sample t*-statistic for testing significance of gene i is

$$t_i = \frac{\bar{x}_i(1) - \bar{x}_i(2)}{s_i} \sim t_{100} \ \ (\text{ under } H_0),$$

where s_i is an estimate of the standard error of $\bar{x}_i(1) - \bar{x}_i(2)$, i.e., $s_i^2 = (1/50 + 1/52)[\sum_{j=1}^{50}\{x_{ij} - \bar{x}_i(1)\}^2 + \sum_{j=51}^{102}\{x_{ij} - \bar{x}_i(2)\}^2]/100$.



$\check{F}_{s,n}^{\check{\alpha}_n}$ (in dotted *red*) and $F_{s,n}^{\dagger}$ (in solid blue).

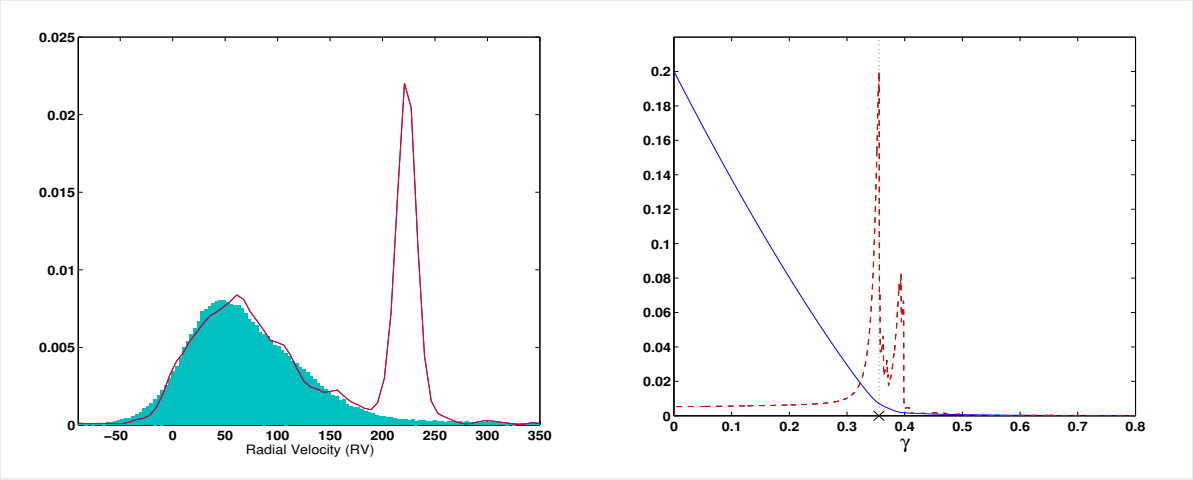


The density $f_{s,n}^{\dagger}$.

- $\hat{\alpha}_n = 0.0877$. The lower bound for α_0 is *0.0512*.

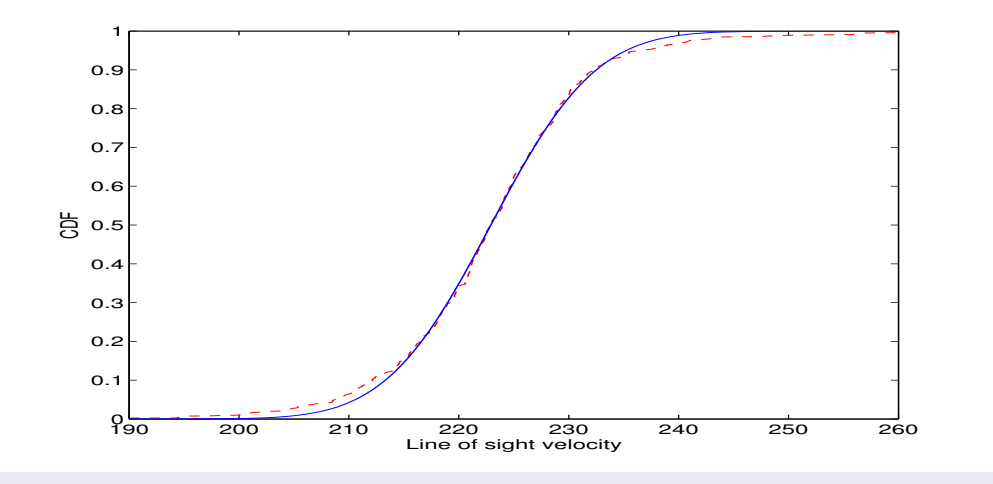
An application in astronomy

- We analyse the *radial velocity* (RV; line of sight velocity) distribution of stars in *Carina*, a dwarf spheroidal (*dSph*) galaxy.
- The data have been obtained by Magellan and MMT telescopes and consist of radial velocity measurements for $n = 1215$ stars from Carina, *contaminated* with Milky Way stars in the field of view.
- We would like to understand the *distribution* of the line of sight velocity of stars in Carina.
- For the contaminating stars from the Milky Way in the field of view, we assume a *non-Gaussian* velocity distribution F_b that is known from the Besancon Milky Way model (Robin et. al, 2003), calculated along the line of sight to Carina.



- In the left panel we have the *histogram* of the radial velocity of the contaminating stars overlaid with the (scaled) *kernel density estimator* of the Carina data set.
- The right panel shows the plot of $\gamma d_n(\hat{F}_{s,n}^{\gamma}, \check{F}_{s,n}^{\gamma})$ (in solid blue) overlaid with its (scaled) *second derivative* (in dashed red).
- Our estimate $\hat{\alpha}_n$ of α_0 for this data set turns out to be 0.356, while the lower bound for α_0 is found to be 0.322.

- Astronomers usually assume the distribution of the radial velocities for these dSph galaxies to be *Gaussian* in nature.
- The figure below shows $\check{F}_{s,n}^{\hat{\alpha}_n}$ (in dashed red) overlaid with the *closest* (in terms of minimising the $L_2(\check{F}_{s,n}^{\hat{\alpha}_n})$ distance) Gaussian distribution (in solid blue).
- Indeed, we see that $\check{F}_{s,n}^{\hat{\alpha}_n}$ is *close* to a normal distribution (with mean 222.9 and standard deviation 7.51).



Summary

- *Consistent* estimation of a *two-parameter mixture* model using techniques from shape-restricted function estimation.
- *Avoids* the need to specify *tuning* parameters.
- Such *shape* constraints arise naturally in many contexts.

References

- Benjamini, Y. & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Statist. Soc. B*, **57**, 289–300.
- Efron, B. (2010). *Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction*. IMS Monographs.
- Genovese, C. & Wasserman, L. (2004). A stochastic process approach to false discovery control. *Ann. Statist.*, **32**, 1035–1061.
- Meinshausen, N. & Rice, J. (2006). Estimating the proportion of false null hypotheses among a large number of independently tested hypothesis. *Ann. Statist.*, **34**, 373–393.
- Robertson,T., Wright, F. T. & Dykstra, R.L. (1988). *Order restricted statistical inference*. Wiley, New York.

Thank You!

Questions?