Semiparametric Estimation of the Proportion of True Null Hypotheses

David Ruppert

Cornell University

http://www.orie.cornell.edu/~davidr

August 2005

Joint work with Dan Nettleton and Gene Hwang

Problem:

- One has a large collection of p-values, $p_1, \ldots, p_n$
- Need to know the proportion that came from a true H_0
  - useful, e.g., to estimate the false discovery rate

Recent paper:

Langaas, Ferkingstad, and Lindqvist (2005, JRSS-B)

- Surveys earlier work on this topic
  - Estimator of Schweder and Spjøtvoll

$$\widehat{\pi}(\lambda) = \frac{\#\{p_j > \lambda\}}{n(1-\lambda)}$$

    * λ estimated by a bootstrapping (Storey, 2002) or spline-smoothing (Storey and Tibshirani, 2003)

- Proposes new estimators
  - Estimate the marginal density of the p-values at 0 by
    * Grenander decreasing density estimator
    * longest-constant interval estimator
    * convex-decreasing estimator

Topic of this talk – semiparametric estimator:

- $\{(p_i, \mu_i)\}_{i=1}^n$ are iid
- let

$$\pi_0 = P(\mu_i \in \text{null region})$$

- $g(\mu)$ = density of μ_i under H_1

- marginal cdf of p_i is

$$F_p(p\,;\pi_0) = \pi_0 p + (1-\pi_0)\int_0^\infty F_{p|\mu}(p\,;\mu)g(\mu)d\mu \qquad (1)$$

- denote marginal pdf by $f_p(p\,;\pi_0)$.

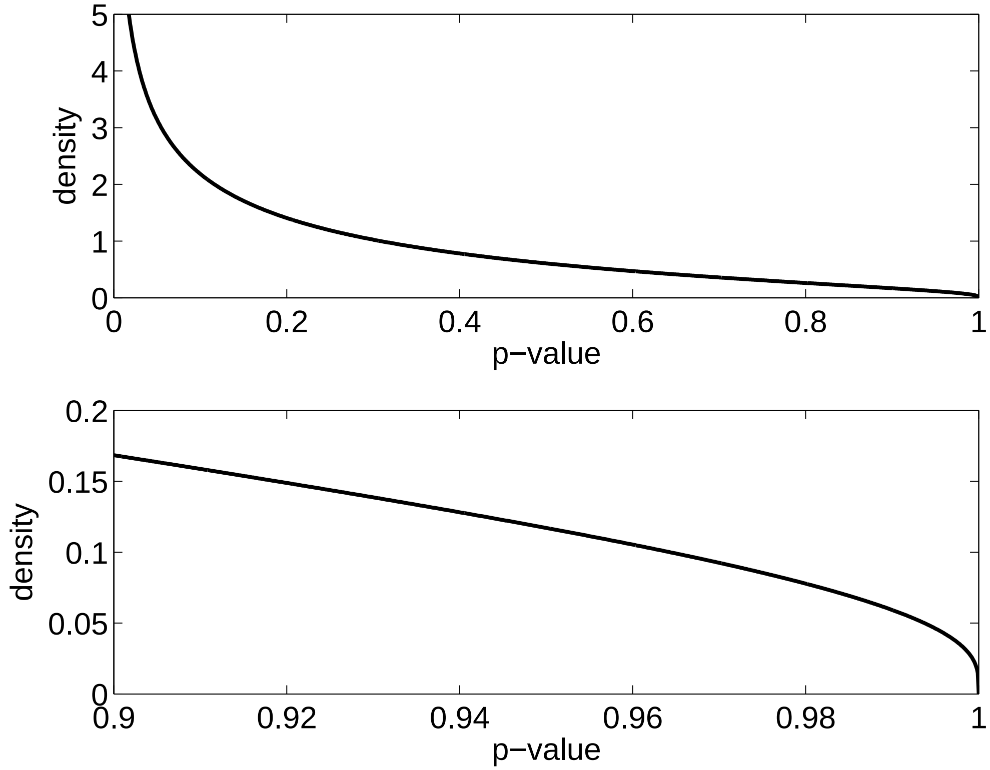

Figure 1: Density of the p-value from a z-test of $H_0 : \mu = 0$ versus $H_1 : \mu > 0$ when $\mu = 1$. The lower plot zooms in on the region where the density is concave.

- model g as $g(\mu\,;\boldsymbol{\beta})$
  - $g(\cdot\,;\cdot)$ is a known function
  - $\boldsymbol{\beta}$ is a vector of parameters
  - will use linear splines
- let $F_p(\cdot\,;\pi_0,\boldsymbol{\beta})$ be given by (1) with $g(\mu)$ replaced by $g(\mu\,;\boldsymbol{\beta})$.

Weighted penalized least-squares

- let l_i, c_i, r_i, and $w_i = r_i - l_i$ be the left edge, center, right edge, and width of the ith bin, $i = 1, \ldots, N_{\text{bin}}$
- let $M_1, \ldots, M_{N_{\text{bin}}}$ be the bin counts
- $$y_i = \frac{M_i}{n w_i}$$

  is an unbiased estimate of

  $$m_i(\pi_0, \boldsymbol{\beta}) = \frac{F_p(l_i\,; \pi_0, \boldsymbol{\beta}) - F_p(r_i\,; \pi_0, \boldsymbol{\beta})}{w_i} \approx f_p(c_i\,; \pi_0)$$

- estimate $(\pi_0, \boldsymbol{\beta})$ by minimizing the penalized sum of squares is

$$SS(\pi_0, \boldsymbol{\beta}\,; \lambda) = \sum_{i=1}^{N_{\mathrm{bin}}} \{y_i - m_i(\pi_0, \boldsymbol{\beta})\}^2 + \lambda Q(\boldsymbol{\beta}) \qquad (2)$$

  - $\lambda \geq 0$
  - $Q(\boldsymbol{\beta})$ is a roughness penalty

Spline model for g

- g will be modeled as a linear spline and estimated using the B-spline basis
- g is assumed to have support contained in $[0, \mu^*]$
- spline will have K knots, $0 = \kappa_1, \ldots, \kappa_K = \mu^*$, equally spaced between 0 and μ^*

- B-splines are normalized to be densities
  - not essential, but helpful
  - any convex combination is a density
- let

$$g(\mu, \boldsymbol{\beta}) = \sum_{k=1}^{K-1} \beta_k B_k(\mu), \qquad (3)$$

where $\beta_k \geq 0$ for all k and $\sum_{k=1}^{K-1} \beta_k = 1$.

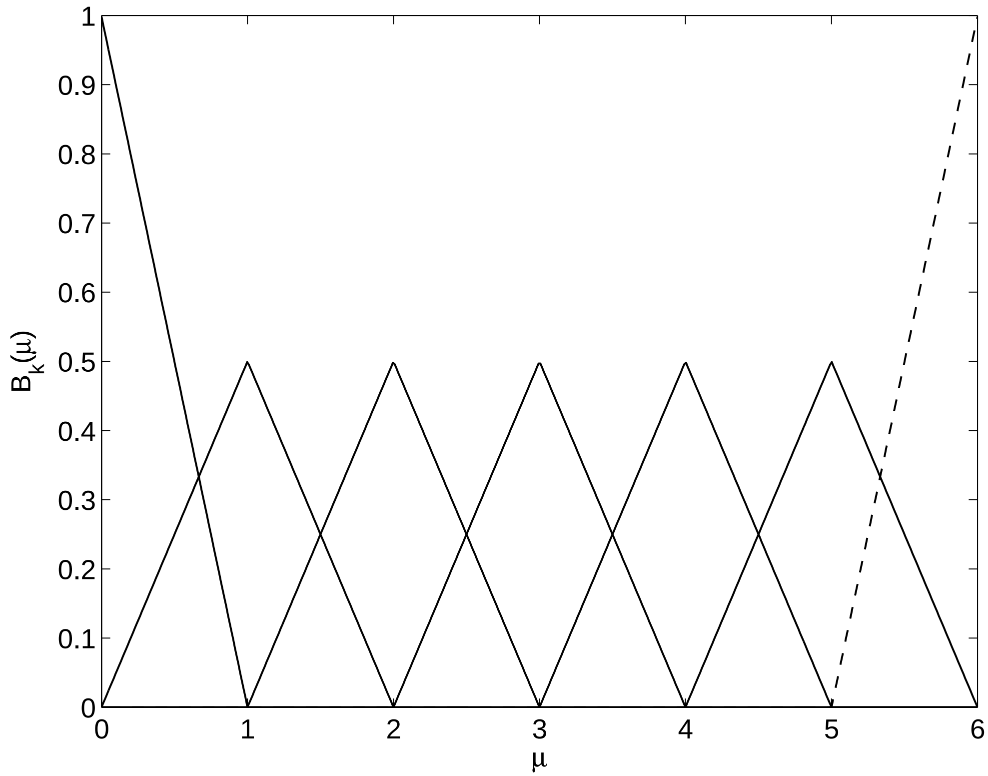

Figure 2: B-splines with 7 knots and $\mu^* = 6$ used to model g. Each B-spline is normalized to be a density. The B-spline with support [5, 6] is shown as a dashed line and is not used in the model for g because it is discontinuous at 6.

- define $\theta_1 = \pi_0$ and $\theta_{k+1} = (1-\pi_0)\beta_k$ for $k = 1, \ldots, K-1$
- define $\boldsymbol{\theta} = (\theta_1, \ldots, \theta_K)^\mathsf{T}$
- let $Z_1(p) = p$ be the (uniform) cdf of the p-values under H_0
- for $k = 1, \ldots, K-1$, let $Z_{k+1}(p) = \int F_{p|\mu}(p;\mu)B_k(\mu)d\mu$ be the marginal cdf of a p-value if the density of μ is B_k
- the marginal cdf of a p-value is modeled as

$$F_p(p\,;\boldsymbol{\theta}) = \sum_{k=1}^{K} \theta_k Z_k(p), \tag{4}$$

where

$$\theta_k \geq 0,\ \forall\, k,\ \text{and} \ \sum_{k=1}^{K} \theta_k = 1 \tag{5}$$

- The roughness penalty is

$$
\begin{aligned}
Q(\boldsymbol{\theta}) &= (2\theta_1 - \theta_2)^2 + \sum_{k=2}^{K-1} (\theta_k - \theta_{k+1})^2 \\
&= \{d(1-\pi_0)\}^2 \sum_{k=1}^{K-1} \{g(\kappa_k) - g(\kappa_{k+1})\}^2
\end{aligned}
$$

- the sum of squares is

$$
\begin{aligned}
SS(\boldsymbol{\theta};\lambda) &= \sum_{i=1}^{N_{\text{bin}}}\left\{y_i - \sum_{k=0}^{K-1}\theta_k Z_{i,k+1}\right\}^2 \\
&+ \lambda\left\{(2\theta_2-\theta_3)^2 + \sum_{k=3}^{K-1}(\theta_k-\theta_{k+1})^2\right\} \\
&= \|\mathbf{y}-\mathbf{Z}\boldsymbol{\theta}\|^2 + \lambda\boldsymbol{\theta}^{\mathsf{T}}\left\{(\mathbf{DA})^{\mathsf{T}}\mathbf{DA}\right\}\boldsymbol{\theta},
\end{aligned}
$$

  where

  - $\mathbf{y} = (y_1,\ldots,y_{N_{\text{bin}}})^{\mathsf{T}}$
  - $\mathbf{Z}$ is the $N_{\text{bin}} \times K$ matrix whose i,jth element is $Z_{i,j} = \{Z_j(r_i) - Z_j(l_i)\}/w_i$
  - $\mathbf{A} = \text{diag}(0,2,1,\ldots,1)$
  - $\mathbf{D}$ is a $(K-2)\times K$ "differencing matrix" whose ith row has $+1$ in column $i+1$, -1 in column $i+2$, 0 elsewhere

- minimizing $SS(\boldsymbol{\theta}; \lambda)$ is equivalent to minimizing

$$\boldsymbol{f}^\top \boldsymbol{\theta} + 0.5\, \boldsymbol{\theta}^\top \mathbf{H} \boldsymbol{\theta} \tag{6}$$

  where

  - $\boldsymbol{f} = -\mathbf{y}^\top \mathbf{Z}$
  - $\mathbf{H} = \mathbf{Z}^\top \mathbf{Z} + \lambda \mathbf{A}^\top \mathbf{D}^\top \mathbf{D} \mathbf{A}$,

- the constraints are

$$\boldsymbol{\theta} \geq 0 \text{ and } \mathbf{1}^\top \boldsymbol{\theta} = 1, \tag{7}$$

  - $\mathbf{1}$ is a K-dimensional vector of ones

- approximate GCV — use GCV for the unconstrained estimator

Two semiparametric estimators of θ:

- $\widehat{\pi_0}_{\text{sem},1} = \widehat{\theta}_1$
- $\widehat{\pi_0}_{\text{sem},2}$ = estimated density at 1 Recall:

$$F_p(p\,;\boldsymbol{\theta}) = \sum_{k=1}^{K} \theta_k Z_k(p)$$

Therefore,

$$\widehat{\pi_0}_{\text{sem},2} = \sum_{k=1}^{K} \widehat{\theta}_k Z'_k(p) \Big|_{p=1}$$

Recall:

- $Z_1(p) = p$ is the (uniform) cdf of the p-values under H_0
- for $k = 1, \ldots, K-1$, let $Z_{k+1}(p) = \int F_{p|\mu}(p;\mu)B_k(\mu)d\mu$ is the marginal cdf of a p-value if the density of μ is B_k
- therefore

$$\widehat{\pi_0}_{\text{sem},2} = \widehat{\theta}_1 + \sum_{k=2}^{K} \widehat{\theta}_k \int f_{p|\mu}(p;\mu)B_k(\mu)d\mu \Big|_{p=1} \geq \widehat{\pi_0}_{\text{sem},1}$$

Simulation study

- one-side z-test
  - $\mu = 0$ versus $\mu > 0$ based on $Z \sim N(\mu, 1)$
- g is beta(b_1, b_2) on $[\mu_{\min}, \mu_{\max}]$
- Gr-M and LCI-M are the Grenander and longest constant interval estimator estimators using M equally-spaced order statistics
  - $M = n$ gives standard Grenander and LCI estimators

Gr-50	Gr-500	Gr-5000	LCI-50	LCI-500	LCI-5000
3.0231	23.3634	95.7562	1.9185	4.2623	12.6780

Table 1: $1000 \times$ MSE for six estimators with $n = 5000$, $\pi_0 = 0.8000$, $\mu_{\min} = 0$, $\mu_{\max} = 4$, $b_1 = 2$, and $b_2 = 2$. Each MSE is based on 25 Monte Carlo simulations. The standard errors of the MSE values are roughly 1/2 the MSE values themselves or smaller.

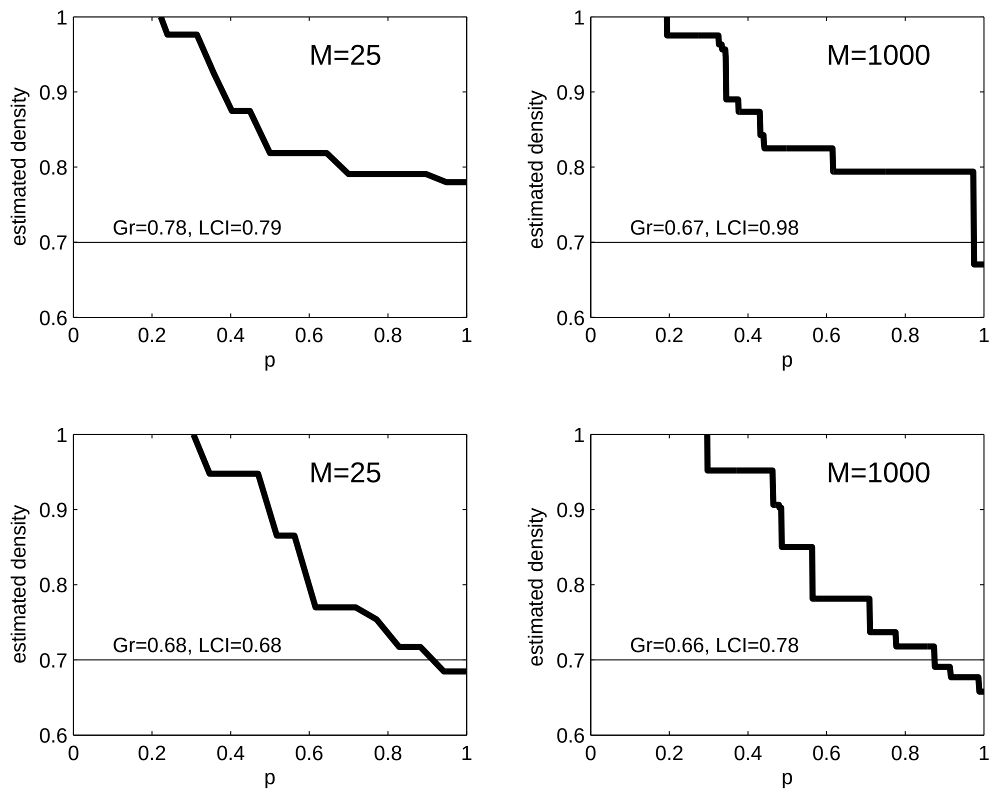

Figure 3: Comparison of Gr-25, Gr-1000, LCI-25, LCI-1000 estimators. The top and bottom rows are different data sets, both from Case #3.

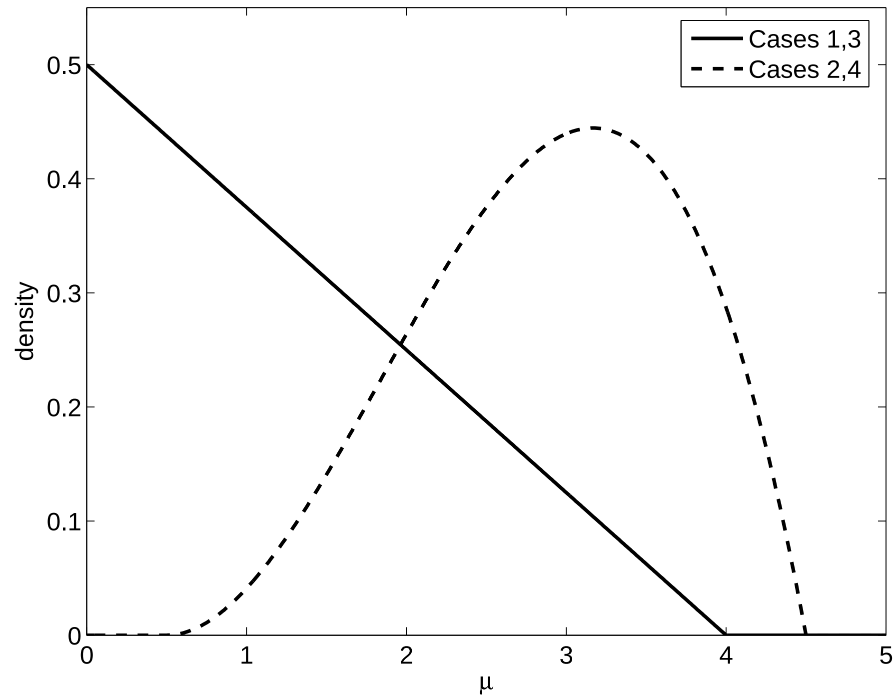

Figure 4: The two non-null densities of μ used in the simulations. Their values of $(\mu_{\min}, \mu_{\max}, b_1, b_2)$ are (0, 4, 1, 2) for Cases 1 and 3, and (0.5, 4.5, 3, 2) for Cases 2 and 4.

	Case #1	Case #2	Case #3	Case #4
π_0	0.95	0.95	0.7	0.7
$\mu_{\min}, \mu_{\max}, b_1, b_2$	0, 4, 1, 2	0.5, 4.5, 3, 2	0, 4, 1, 2	0.5, 4.5, 3, 2
$\widehat{\pi_0}_{\text{sem},1}$, $K = 8$, wt	0.3759	0.3555	1.6042	0.8240
$\widehat{\pi_0}_{\text{sem},2}$, $K = 8$, wt	0.3014	0.1878	2.3984	0.3466
$\widehat{\pi_0}_{\text{sem},1}$, $K = 16$, wt	0.4694	0.2974	1.8562	1.0424
$\widehat{\pi_0}_{\text{sem},2}$, $K = 16$, wt	0.2961	0.1635	2.5801	0.4004
Gr-10	0.6609	0.8513	4.4478	0.4323
Gr-50	4.0509	4.4439	2.5229	2.3228
Gr-250	16.8678	17.7898	6.5489	9.6928
LCI-10	0.7541	0.7012	4.6142	0.4575
LCI-50	2.2739	1.6569	5.4461	1.4849
LCI-250	3.1757	2.3155	10.2505	2.1253

Table 2: 1000 $\times$ MSE. 1500 Monte Carlo samples per case.

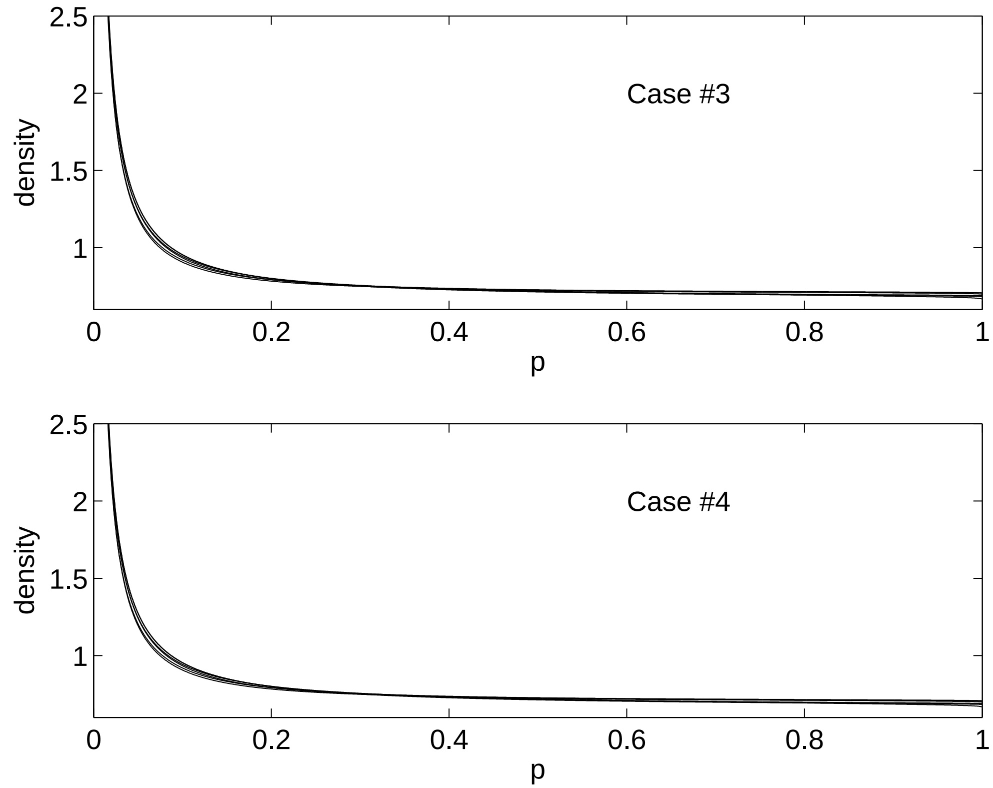

Figure 5: Semiparametric estimates of f_p, the density of the p-values, from six independent data sets from Cases #3 and #4.

Features of semiparametric estimators:

- accurate (small MSE and bias)
- shape-preserving
- fully automatic
- can be computed very rapidly