Estimating the mean of high valued observations in high dimensions

Eitan Greenshtein, SAMSI, Research Triangle Park, NC (eitan.greenshtein@gmail.com)

Junyong Park, Department of Mathematics and Statistics, University of Maryland Baltimore County, MD (junpark@math.umbc.edu)

Ya'acov Ritov, Jerusalem, Israel (yaacov.ritov@gmail.com)

Abstract: Let $Y_i \sim N(\mu_i, 1)$, $i = 1, ..., n$, be independent random variables. We study the problem of estimating the quantity $S = \sum_{\{i|C<Y_i\}} \mu_i$. We emphasize the case where n is large, the vector $(\mu_1, ..., \mu_n)$ is sparse, and the value of C is large. Our approach is nonparametric empirical Bayes, where μ_i are assumed i.i.d from an unknown G. The performance of our suggested estimator is studied both theoretically and through simulations. We also obtain some results related to the local false discovery rates corresponding to high valued points Y_i.

Keywords: FDR; Sparse vector; Empirical Bayes.

---

1 Introduction

A standard semiparametric problem starts with a family $\mathcal{P}$ of distributions, a parameter $\vartheta : \mathcal{P} \to \mathbb{R}^d$, and a simple random sample $\boldsymbol{Y}_n = Y_1, \ldots, Y_n$ from $P \in \mathcal{P}$. The statistical challenge is identifying an estimator $\hat{\vartheta}(\mathbb{P}_n)$ of $\vartheta(P)$, where $\mathbb{P}_n$ is the empirical distribution of $\boldsymbol{Y}_n$. Many times, $\hat{\vartheta}(\mathbb{P}_n) = \vartheta(\mathbb{P}_n)$. Cf. Bickel and Lehmann (1975). In this paper we consider an estimator of a parameter which is a function of P, but also of the sample $\boldsymbol{Y}_n$ itself, $\vartheta = \vartheta(\mathbb{P}_n, P)$. In fact, many standard estimators are 'second order' statistics, and hence are naturally written in the above form. Thus, the sample variance is estimating $\vartheta(\mathbb{P}_n, P) = \mathrm{E}_P(Y - \mathrm{E}_{\mathbb{P}_n} Y)^2$. It should be noted that this expression, although a function of the sample, is not an estimator, since it involves the unknown underlined distribution P. The sample variance depends both on P and $\mathbb{P}_n$. However, the dependency on the latter disappears when the asymptotic influence function is considered, and the estimator of the sample variance is equivalent with $o_p(n^{-1/2})$ term to $\vartheta^*(P) = \mathrm{E}_P(Y - \mathrm{E}_P Y)^2$ which is estimated non-parametrically by $\vartheta^*(\mathbb{P}_n)$. However, there are situations where the parameter of interest are bone-fide

of the type $\vartheta = \vartheta(\mathbb{P}_n, P)$. Such a parameter was considered in Skinner and Shlomo (2006) in the context of identification risk in data disclosure where the parameter of interest is a function of the number of subjects in a given cell both in the population as well as in sample. Our paper deals with another situation. See below.

Let Y be an observed random variable from a real distribution $F = G * N(0,1)$ where G is some an underlined distribution. The deconvolution of F, or the estimate of G under standard conditions was investigated by many. Cf. Fan (1991). A different description of the same problem is the following. Suppose $(\mu_1, Y_1), \ldots, (\mu_n, Y_n)$ are i.i.d. pairs, with common distribution P given by $\mu_i \sim G$ and $Y_i \mid \mu_i \sim N(\mu_i, 1)$. Assume that only $Y_1, \ldots, Y_n$ are observed. The μ_is are unobserved, and in fact their existence in only stipulated by the observers. That is, we observe only $Y_1, \ldots, Y_n$, an i.i.d. sample from F as above. One can consider instead of estimating G directly, estimating the equivalent function $\mathrm{E}_P(\mu \mid Y)$. This type of an estimator appears in many semiparametric mixture model, see Bickel, Klaassen, Ritov, and Wellner (1993), for example. As a concrete example, this function appears explicitly in the errors in variables model with normal noise. See Bickel and Ritov (1987).

A similar problem was investigated earlier from an empirical Bayes point of view, or as a compound decision problem. In this case we observe n independent observations, $Y_i \sim N(\mu_i, 1)$, $i = 1, ..., n$. That is, unlike the previous model, the μ_is are now the unknown parameters of the problem. The Y_is are considered to be exchangeable, that is, their order doesn't carry any particular information. The purpose now is to estimate $\mu_1, \ldots, \mu_n$ with some average loss function. It is known that the solution of the compound decision problem is asymptotically equivalent to a Bayesian problem where the μ_i are *i.i.d.*. Thus this problem with quadratic loss function is equivalent to the deconvolution problem mentioned above.

We consider a special variant of this model. We assume that the vector $\mu = (\mu_1, \ldots, \mu_n)$ is sparse, in the sense that most of the μ_i's are 0. A few of the Y_is are selected for a further investigation, and they should correspond to those with relatively large value of μ_i. Without any auxiliary information, the natural (and in fact the only conceivable) selection procedure is selecting those items with large value of Y_i. We want to investigate the total amount of signal in the selected lot. Formally, we study the estimation of

$$S_C = \sum \mu_i \mathbb{I}(Y_i > C), \tag{1.1}$$

for some given fixed C.

Our estimation problem is a special case of the more general problem of estimating $\sum_{i=1}^{n} U(X_i, \theta)$, for observed X_i, $X_i \sim F_{\theta_i}$, $\theta_i \in \Theta$, where θ_i is unknown, and a given function U. Note, this is not a standard estimation problem since the above quantity is random. See Zhang (2005), for various examples and applications and further references. See also Robbins and Zhang (1988). Our suggested technique, for the concrete estimation problem (1), is not studied in the fore mentioned papers, yet, the Empirical Bayes approach is common to our paper and to those papers.

We give now some motivation for estimating (1.1). Suppose candidates are selected according to their achievement value Y_i in a screening test. Let $\mathbb{S} = \{i : \ Y_i > C\}$ the sub-sample of the selected items. They are tested again after treatment to yield the value $\tilde{Y}_i^*$. A naive approach would compare $\bar{Y}_{\mathbb{S}}$ to $\bar{Y}_{\mathbb{S}}^*$, the mean of those in $\mathbb{S}$ before and after the treatment. However, a standard regression to the mean argument shows that $\bar{Y}_{\mathbb{S}}$ is stochastically larger than $\bar{Y}_{\mathbb{S}}^*$. Hence we need a fair estimator of the before treatment value S. See, the motorist example described in Zhang (2005), where drivers with especially bad driving record are trained and the effectiveness of the training should be estimated. An amusing related example, which demonstrates the consequence of ignoring the regression to the mean effect, is by Herbert Robbins; it is about a "successful" training of coins, where their ability to land on tail is "improved".

More generally, consider now a case where n is large and the vector $\mu = (\mu_1, ..., \mu_n)$ is sparse. Such a situation occurs when there are many variables involved in a study, but only a few are expected to be meaningful, i.e., only a few of the coordinates μ_i do not equal (nearly) zero. This is the typical situation /in the analysis of fMRI and gene arrays. Cf. Erickson and Sabatti (2005), and is typical to many data mining applications. A few hopefully"Meaningful" variables are selected. One may be interested in S, with the interpretation of the "overall amount of signal" one was able to capture. Large μ_is count more than small ones, even if both of them are different form 0. This question may be more interesting than just how many of μ_is are different from 0. A question answered by, for example, the FDR (False Discovery Rate) methodology, see Benjamini and Hochberg (1995). However, under the sparse setup that we study, there are simple implications to the Local False Discovery Rate, denoted fdr, which was suggested by Efron, et al. (2002), see also Storey (2003). The density estimation technique that we use in Section 2, together with the sparsity assumptions and Empirical Bayes interpretation, suggest a Bayesian ranking of the large valued observations/'discoveries' through a consistent estimator of the fdr. The ranking is in terms of the strength of evidence against $H_0^i : \ \mu_i = 0$.

In Section 2 we treat the estimation of (1.1). The treatment involves Empirical Bayes considerations and density estimation. In Section 3 we present some simulation results. In Section 4 we study the concept of Local False Discovery Rate.

2 Estimation

We take an Empirical Bayes approach where μ_i are modeled as independent observations from an unknown distribution G. We emphasize the case where most of the mass of G is near zero, that is, the vector μ is sparse. This assumption is expressed through equations (2.10) and (2.11) below. However, we stress that the method we develop is general, and it covers non-sparse cases.

Let μ_i, $i = 1, \ldots, n$, be i.i.d. G, and given them, let $Y_1, \ldots, Y_n$ be independent, $Y_i \sim N(\mu_i, 1)$. Recall (1.1), $S_C = \sum_{i=1}^n \mu_i I(Y_i > C)$. Then:

$$\begin{aligned} \eta \equiv E(S_C) &= \sum_{i=1}^n E(\mu_i I(Y_i > C)) \\ &= \sum_{i=1}^n E(\mu_i | Y_i > C) P(Y_i > C) \\ &= np\, E(\mu_i | Y_i > C) \end{aligned}$$

where $p = P(Y_i > C)$. By equation (1.2.2) in Brown(1971),

$$E(\mu_i | Y_i = y) = y + \frac{f'(y)}{f(y)} \tag{2.2}$$

where $f(\cdot)$ is the marginal density of Y, q $f(y) = \int \varphi(y - \mu) dG(\mu)$, where φ is the standard normal probability density function. Note that this equation is valid only under the assumption of normality. However, the assumption that the Y_is are independent can be relaxed considerably. Hence

$$\begin{aligned} E(\mu_i | Y_i > C) &= \int_C^\infty E(\mu_i | Y_i = y) \frac{f(y)}{p} dy \\ &= \int_C^\infty \left(y + \frac{f'(y)}{f(y)} \right) \frac{f(y)}{p} dy \\ &= \frac{1}{p} \int_C^\infty y f(y) dy + \frac{1}{p} \int_C^\infty f'(y) dy \\ &= \frac{1}{p} \int_C^\infty y f(y) dy - \frac{1}{p} f(C). \end{aligned}$$

Therefore,

$$\eta = npE(\mu_i | Y_i > C) = n \left(\int_C^\infty y f(y) dx - f(C) \right) \tag{2.3}$$

The estimator $\hat{S}_C$ for S_C, which is suggested in the following, is motivated through $E\hat{S}_C \approx \eta = ES_C$. It is of the form:

$$\hat{S}_C = n \left(\int_C^{\widehat{\infty}} y f(y) dx - \hat{f}(C) \right)$$

which will be made explicit in the sequel. Zhang (2005), treats formally the issue of estimating η versus estimating S_C.

First we estimate $\int_C^\infty y f(y) dy$ by the natural estimator, denoted M,

$$M = \frac{1}{n} \sum_{\{i: Y_i > C\}} Y_i. \tag{2.4}$$

It remains to estimate $f(C)$. We describe now one of the many possibilities, a standard kernel density estimator. Let

$$\hat{f}_h(C) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{C - Y_i}{h}), \tag{2.5}$$

where $K = \varphi$ is a Gaussian kernel.

The choice of the bandwidth h is discussed in the following. The slightly non-standard part of this discussion of the bandwidth choice, is that we are interested in estimating the density $f(C)$ for a large C, i.e., deep in the tail. Consider the mean square error (MSE) of $\hat{f}(C)$:

$$\mathrm{MSE}(\hat{f}_h) = \mathrm{Bias}(\hat{f}_h)^2 + \mathrm{var}(\hat{f}_h).$$

Standard calculations in kernel estimation, see, e.g., Silverman (1992), imply that:

$$\mathrm{Bias}(\hat{f}_h(C)) \backsimeq f''(C)h^2, \tag{2.6}$$

while

$$\mathrm{var}(\hat{f}_h(C)) \backsimeq \frac{1}{nh} f(C). \tag{2.7}$$

Here, we use the notation $\backsimeq$ to imply equality up to a bounded factor for $h \to 0$. As in the standard case, the optimal bandwidth h approaches 0 as $n \to \infty$. Thus, again as in the standard development, by equating the squared bias and the variance we approximate the optimal bandwidth, denoted h_{opt} and its corresponding squared error risk. The resulting quantities are:

$$h_{opt} \backsimeq \left[\frac{f(C)}{n(f''(C))^2}\right]^{\frac{1}{5}}, \tag{2.8}$$

and

$$E(\hat{f}_{h_{opt}}(C) - f(C))^2 \backsimeq \left[\frac{f(C)(f''(C))^{\frac{1}{2}}}{n}\right]^{\frac{4}{5}}. \tag{2.9}$$

2.1 Asymptotics for a sparse vector of means and large C

As explained in the introduction, we are especially interested in the case where C is large. Hence, in order to apply the above we need to estimate $f(C)$ and $f''(C)$ for large values of C. In the asymptotics that follows, we consider $C \equiv C_n = (2\alpha \log(n))^{1/2}$ $0 < \alpha < 1$. For such points $f(C_n)$ and $f''(C_n)$, might be of order $n^{-\alpha}$. The usual interpretation of (2.8) and (2.9), as having bandwidth h of the order $n^{-1/5}$ while the corresponding square error risk of the order $n^{-4/5}$, is not valid anymore. Thus, a special attention should be given to the issue of estimation of $f(C)$

and $f''(C)$, in order to plug into (2.8) and (2.9), for the purpose of getting estimates for h_{opt} and its corresponding risk. An accuracy of order (say) $n^{-1/2}$ in estimation of $f(C)$, could be very misleading, if the order of magnitude of $f(C)$ itself is smaller than $n^{-1/2}$.

We will now turn to deal with the setup we have in mind, where the vector μ is sparse. Our formal asymptotic treatment is in a context of *triangular array*. That is, at stage n, the n random variables $\mu_1, ..., \mu_n$, are i.i.d from a distribution G^n, G^n depending on n. Note, when assuming a fixed G, as $n \to \infty$, we can not achieve the sparsity setup we want to study. Under the setup that we have in mind the proportion of non-zero signals is $o(1)$, as $n \to \infty$. We will drop the super-script n; we write the mixture density, corresponding to G^n, simply as f rather than f^n. Similarly we write C rather than C_n.

The following assumptions, while implying sparsity, are also convenient for our derivation. Assume:

$$f(C) < \kappa_1 \varphi(C), \tag{2.10}$$

$$|f''(C)| < \kappa_2 |\varphi''(C)|. \tag{2.11}$$

The bounds κ_i $i = 1, 2$, are uniform in n.

Remark 1

(i) Typically under (2.10) we expect that (2.11) is also satisfied, e.g., if $\mu_i < 2C$ for all i, then (2.10) implies (2.11).

(ii) Assumption (2.10) holds for points $C \equiv C_n$ is such that meaningful portion of the detected item correspond to zero signal. Formally, if $\mathcal{S} = \{i : Y_i > C\}$, $P(|\{i \in \mathcal{S}, \mu_i > 0\}|/|\mathcal{S}| < 1 - \epsilon) \to 1$, where $\epsilon > 0$. This is the more interesting and challenging case, where it is essential to find a way to 'screen out' the zero signals.

Define now $h^0(n)$ by

$$h^0(n) \sim \left[\frac{\varphi(C)}{n(\varphi''(C))^2}\right]^{\frac{1}{5}}, \tag{2.12}$$

Denote the kernel estimator induced by the $h^0(n)$, by

$$\hat{f} \equiv \hat{f}_{h^0(n)}.$$

Finally, define our estimator $\hat{S_C}$ as:

$$\hat{S_C} = n(M - \hat{f}), \tag{2.13}$$

where M is defined in (2.4) .

Theorem 1: Under conditions (2.10) and (2.11),

$$\hat{S_C} = \eta + O_p(n^{\frac{3}{5}}[\varphi(C)(|\varphi''(C)|)^{\frac{1}{2}}]^{\frac{2}{5}}) = \eta + O_p(C[n\varphi(C)]^{\frac{3}{5}}) \tag{2.14}$$

$$\hat{S_C} = S_C + O_p(C[n\varphi(C)]^{\frac{3}{5}}) \tag{2.15}$$

Proof: The proof of the first equality in (2.14) follows from the above, when observing that the variance of $\hat{S_C}$ is of the order of the variance of $\hat{f}(C)$ (i.e., $\mathrm{var}(\sum_{i=1}^{n} Y_i I(Y_i > C))/n = O(\mathrm{var}(\hat{f}(C)))$. The second equality in (2.14) follows by replacing $|\varphi''(C)|$ by $C^2\varphi(C)$.

The proof of (2.15) follows when observing that $(S_C - \eta) = O_p(C[n\varphi(C)]^{\frac{3}{5}})$. This follows since S_C is an unbiased estimator for η, with standard deviation $O_p([n\varphi(C)]^{\frac{1}{2}})$. The last order for the standard deviation of S_C is obtained as follows. $E[\mu_i I(Y_i > C)]^2 \leq P(Y_i > C)E\mu_i^2 = O_p(\varphi(C))$. Hence $\mathrm{var}(S_C) = O_P(n\varphi(C))$. □

The last theorem gives the same approximation for the order of $(\hat{S_C} - \eta)$ and for that of $(\hat{S_C} - S_C)$. It seems plausible that the second quantity is typically smaller than the first one.

Remark 2. In situations where there is a very sparse and weak signal, the estimator $\hat{S_C}$, might get negative values. The following adjustment of $\hat{S_C}$ makes sense in such a sparse situation. Let

$$\hat{S_C}^{+} = \max(0, \hat{S_C}).$$

In the simulations of the next section we study the adjusted estimator $\hat{S_C}^{+}$.

Remark 3. One may be interested in estimating sum of higher order moments, e.g.,

$$\sum_{i=1}^{n} \mu_i^2 I(Y_i > C).$$

This may be done in a similar fashion, only estimation of further derivatives of f is needed. It involves derivation which is similar to that of equation (1.2.2) of Brown (1971).

Recall that $f(y) = \int \varphi(y - \mu)dG(\mu)$. When computing the second derivative of f through differentiation inside the integral we obtain:

$$\frac{f''(y)}{f(y)} = -1 + y^2 - 2yE(\mu|Y = y) + E(\mu^2|Y = y).$$

Recall that $E(\mu|Y=y) = y + \frac{f'(y)}{f(y)}$. We obtain that:

$$\begin{aligned}
&\int_C^\infty E(\mu^2|Y=y)f(y)dy \\
&= \int_C^\infty f''(y)dy + \int_C^\infty f(y)dy + \int_C^\infty y^2 f(y)dy + \int_C^\infty 2yf'(y)dy \\
&= -f'(C) + (1-F(C)) + \int_C^\infty y^2 f(y)dy - 2[Cf(C) + 1 - F(C)] \\
&= -f'(C) - (1-F(C)) + \int_C^\infty y^2 f(y)dy - 2Cf(C)
\end{aligned}$$

It follows that the estimation of $\sum \mu_i^2 I(Y_i > C)$ can be done along the lines of the estimation of $\sum \mu_i I(Y_i > C)$, involving a further estimation of $f'(C)$.

3 Simulation

In this section, we present simulation studies for various situation. In addition to the kernel estimator with bandwidth as in (2.12) of the previous section, we also consider the bandwidth, $h = 0.9An^{-1/5}$ where A=min(standard deviation, inter-quartile range/1.34), as suggested in Silverman(1992). Note, the later bandwidth is suggested to be used in general, not necessarily in a sparse setup, or under (2.10) and (2.11). Yet, in our simulations, it gives very similar results to the estimator based on our suggested bandwidth (2.12).

We consider two more estimators for S_C. One is the naive approach of a hard-threshold estimator, which estimates the mean of observations with values Y_i above a threshold C, by their m.l.e (i.e., by the observed Y_i). Define:

$$\hat{S}_{C,hard} = \sum_{\{i:Y_i>C\}} \hat{\mu}_i^{mle} = \sum_{\{i:Y_i>C\}} Y_i$$

The other estimator follows the conditional maximum likelihood approach. This approach was suggested for variable selection by Greenshtein et al. (2006). The estimator is the maximum likelihood estimator of each of the relevant μ_i, conditional on the event $Y_i > C$. That is:

$$\hat{\mu}_i^{con} = \text{argmax}_{\mu_i} \frac{\varphi(Y_i - \mu_i)}{P_{\mu_i}(Y_i > C)} = \text{argmax}_{\mu_i} \frac{\varphi(Y_i - \mu_i)}{1 - \Phi(C - \mu_i)},$$

where φ and Φ are density and cdf of standard normal.

The latter estimator may obtain occasionally very small (negative!) values, when the value of Y_i is greater but very close to C. In fact, in the extreme case, $Y_i = C$, the corresponding conditional m.l.e is $-\infty$. In order to avoid such cases, and to get more meaningful comparisons with the conditional m.l.e., we consider in this section

a parameter space with $\mu_i \geq 0$. Thus: the conditional m.l.e. is $\hat{\mu}_i^{con+} = \max(\hat{\mu}_i^{con}, 0)$ and the corresponding estimator is

$$\hat{S}_{C,con} = \sum_{\{i:Y_i>C\}} \hat{\mu}_i^{con+}.$$

Of course, such an adjustment is reasonable in sparse situations also without formally assuming that $\mu_i \geq 0$.

As in the case of conditional maximum likelihood estimator, we make the same adjustment also to our estimator $\hat{S_C}$. We will consider

$$\hat{S}_C^+ = \max(\hat{S}_C, 0).$$

In our simulation, we let $n = 10^5$. We study the cases where N, the number of nonzero μ, are 0, 200, and 400. We consider three types of the distributions of nonzero μs: (i) point mass at some μ_0 (ii) gamma distribution with various parameters (iii) absolute value of t distribution with d.f. 1 (i.e., Cauchy distribution). We evaluate the performance of the estimators in terms of $(E_\mu(\hat{S}_C - S_C)^2)^{1/2}$ for different C, $C \in (2, 4)$ where $\hat{S}_C$ represents any of the estimators.

In the following graphs, *kernel1* represents density estimator with the above mentioned bandwidth, suggested in Silverman(1992); *kernel2* represents the kernel estimator with bandwidth in (2.12).

Figure 1 shows the performance of the above mentioned four estimators of S_C, when all μ_i's are 0. In the same way, figure 2, 3 and 4 shows the case the nonzero μ's are from gamma, point mass, and $|t|$ with d.f.1 respectively. The reduction in risk in estimating S_C, as C grows, is since there are fewer indices i for which $Y_i > C$, hence estimating the corresponding sum of μ_i has a smaller risk; a reduction in the risk as C grows may also be since the set of μ_i corresponding to $Y_i > C$ is more homogeneous, as C grows (e.g., there are less zero valued μ_i in the set.)

We see that for moderately high values of C, $\hat{S_C}^+$ clearly dominates the other estimators (using either kernel), while for very high values of C, $\hat{S_C}^+$ and $\hat{S}_{Ccon}$ are comparable. The performance of $\hat{S_C}^+$ is nearly the same for the two kernels. In Figure 1, where all μ_i are zero, there is a slight advantage to the kernel suggested in (2.12).

4 The fdr for a Sparse Vector of means

In this section we study the notion of Local False Discovery Rate, denoted fdr, which was suggested by Efron, et al. (2002). Let $\pi_1 \equiv \pi_1(n)$ be the probability under $G \equiv G^n$, that $\mu_i \equiv \mu_i^n$ is not equal to zero, let, $\pi_0 = 1 - \pi_1$. We assume:

$$\pi_1(n) \to 0. \tag{4.16}$$

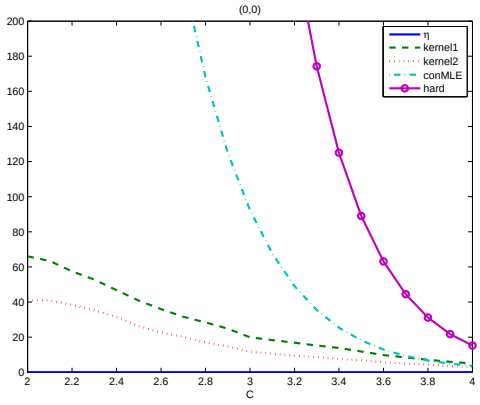

Figure 1: $(N, \mu) = (0, 0)$. Solid line is η. The others are $(E_\mu(\hat{S}_C - S_C)^2)^{1/2}$ where dashed line:kernel1(bandwidth in Silverman(1992)), dotted line:kernel2(bandwidth with (2.12)), dashed-dot line:conditional mle, solid with circle:hard threshold

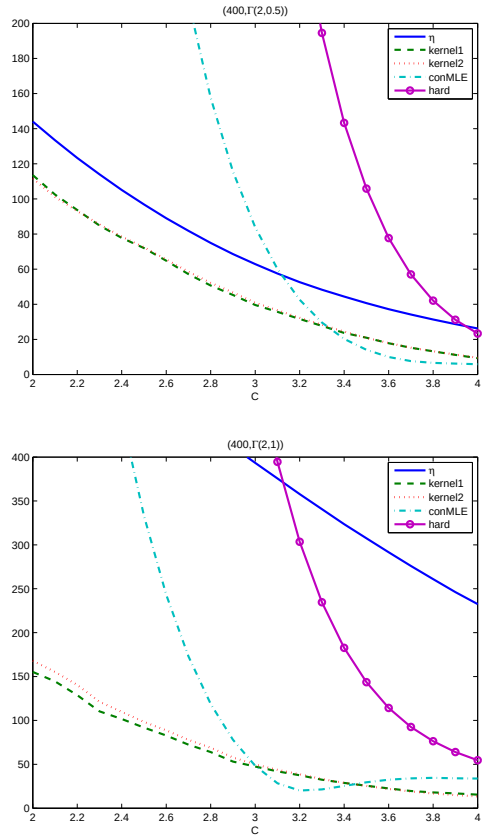

Figure 2: $(N, \mu) = (400, G = \Gamma(2, 0.5))$ and $(N, \mu) = (400, G = \Gamma(2, 1))$

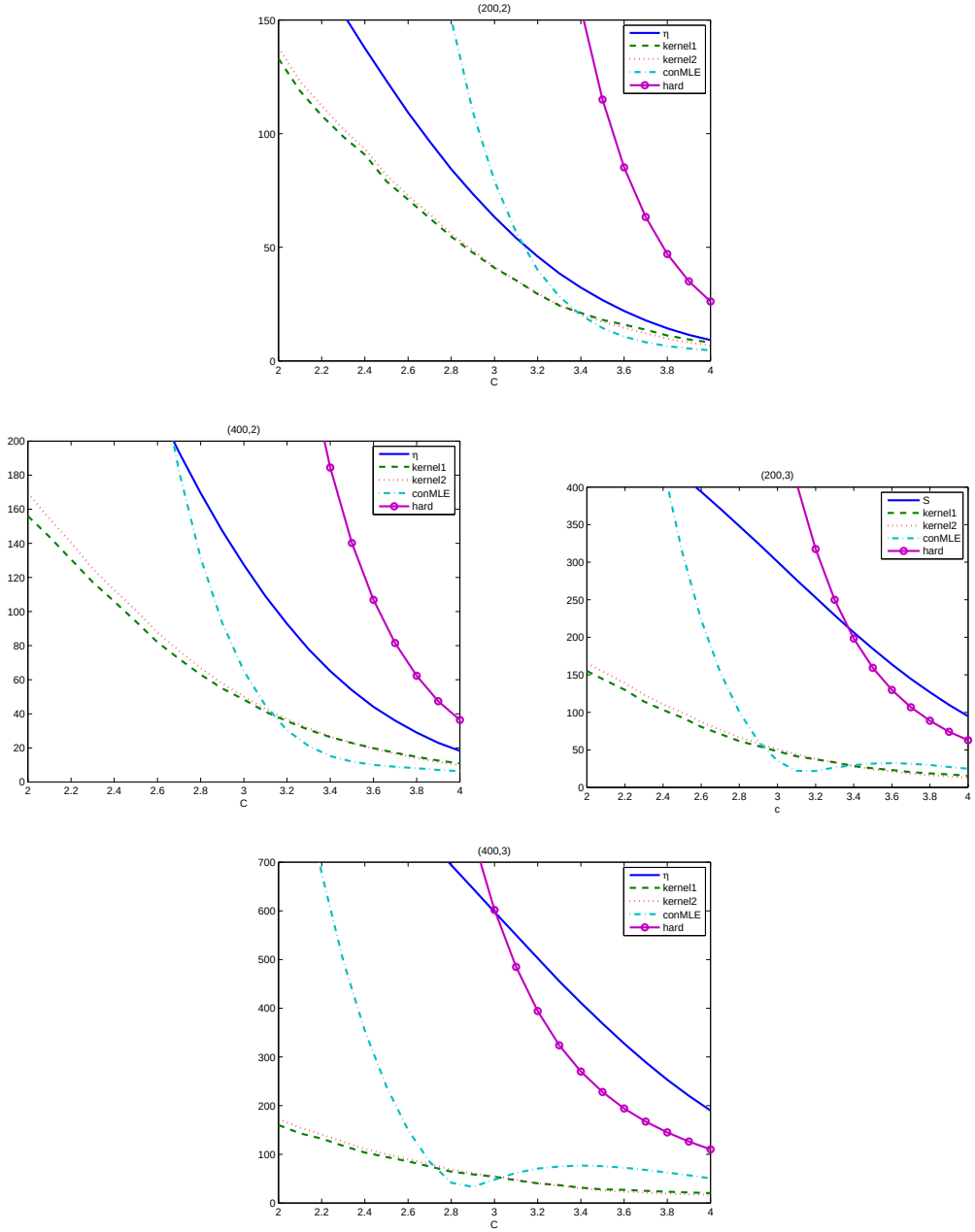

Figure 3: In the first row $(N, \mu) = (200, 2)$ and $(N, \mu) = (400, 2)$. In the second row, $(N, \mu) = (200, 3)$ and $(N, \mu) = (400, 3)$

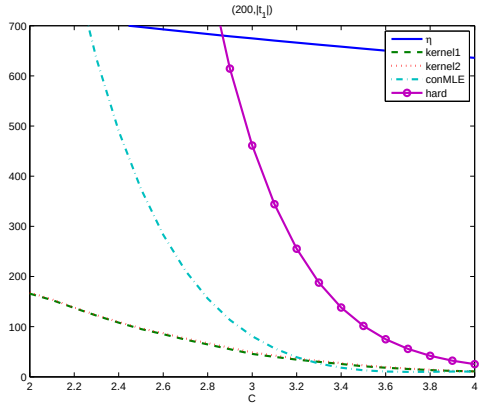

Figure 4: $(N, \mu) = (200, |t_1|)$

Let f be the density of μ_i, under G. Then

$$f(y) = \pi_1 h(y) + \pi_0 \varphi(y), \tag{4.17}$$

where φ is a standard normal density and h is the density of μ_i conditional that it is not zero. Hence:

$$\text{fdr}(y) = P_G(\mu_i = 0 | Y_i = y) = \frac{\pi_0 \varphi(y)}{f(y)} = \frac{\varphi(y)}{f(y)}(1 + o(1)). \tag{4.18}$$

The quantity $0 \leq \text{fdr}(y)) \leq 1$ is suggested (analogously to p-value), as a measure of the evidence against $H_0^i : \ \mu_i = 0$. The smaller is the value of $\text{fdr}(Y_i)$, the stronger is the evidence against H_0^i. On the topic of measuring the evidence against H_0, from frequentist and Bayesian point of view, see Berger, et al. (1994), and references there. See also Storey (2003) which discuss the issue through fdr . Measuring, the evidence against each $H_0^i : \ \mu_i = 0$, is important when planning a future study, having to decide how much effort should be made in further studying each hypothesis H_0^i. As before, we proceed when treating the case $Y_i > 0$, in order to have simpler notations.

When observing $Y_i > (2\log(n))^{1/2}$, we may be quite confident that the corresponding μ_i is greater than zero, even in a sparse case. The interesting task is to measure the evidence against H_0^i, corresponding to Y_i of the order $(2\alpha \log(n))^{1/2}$ for $0 < \alpha < 1$. For such values of α, it may be easily shown that we may estimate $\text{fdr}(y)$ up to $(1 + o(1))$ error factor. Then we will get the following Theorem 2. Note, unlike p-value, the values $\text{fdr}(y)$ are not necessarily monotone decreasing in y, though most usually they are. Let

$$\widehat{\text{fdr}}(y) = \frac{\varphi(y)}{\hat{f}(y)}. \tag{4.19}$$

In the following theorem, we consider asymptotics for a sequence $y = y_n$. As in the previous section, we assume a triangular array setup, in which assumptions (2.10)

and (2.11), are satisfied, where y_n plays the role of C_n. Note, if we do not assume (2.10), then asymptotically the value of $\mathrm{fdr}(y)$ approaches 0, while we are interested in estimating $\mathrm{fdr}(y)$ in the non-trivial case.

Theorem 2. Let $y \equiv y_n = (2\alpha \log(n))^{1/2}$, $0 < \alpha < 1$. Assume (2.10), (2.11) and (4.16). Then

$$\widehat{\mathrm{fdr}}(y) = \mathrm{fdr}(y)(1 + o_p(1)).$$

Proof. The proof follows from the calculations in the previous section. One may verify that for $\alpha < 1$, $\hat{f}(y) = f(y) + o_p(f(y))$. We use $f(y) \sim f''(y) \sim \varphi(y) \sim \varphi''(y)$ as follows from (2.10) and (2.11). Then we apply (4.18). □

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *JRSS-B* **57** 289-300.

Berger, J.O., Brown, L.D. and Wolpert R.L. (1994). A unified conditional frequentist and Bayesian test for fixed and sequential simple hypothesis testing. *Ann.Stat*, **42** No 4, 1787-1807.

Bickel, Klaassen, Ritov and Wellner (1993), *Efficient and Adaptive Estimation for Semiparametric Models*, John Hopkins University Press and Springer.

Bickel, P. J. and Lehmann, E. L. (1975). Descriptive Statistics for Nonparametric Models I. Introduction. *Ann. Statist.* **3**, 1038–1044.

Bickel, P. J. and Ritov, Y. (1987). Efficient Estimation in the Errors in Variables Model. *Ann. of Statist.* **15**, 513–540.

Brown, L.(1971). Admissible estimators, recurrent diffusions, and insoluble boundary value problems. *Ann.Math.Stat.* **42** No. 3. 855-903.

Efron, B., Tibshirani, R., Storey, J.D. and Tusher, V. (2001). Empirical Bayes analysis of a microarray experiment. *JASA* **96** 1151-1160.

Erickson, S. and Sabatti, C. (2005). Empirical Bayes Estimation of a Sparse Vector of Gene Expression Changes. *Statistical Applications in Genetics and Molecular Biology*: Vol. 4 : Iss. 1, Article 22. Available at: http://www.bepress.com/sagmb/v

Fan, J. (1991). Adaptively Local One-Dimensional Subproblems with Application to a Deconvolution Problem. *Ann. Statist.* **19**, 1257–1272.

Greenshtein, E., Park, J., and Lebanon, G.(2006) Regularization through variable selection and conditional m.l.e, with application to classification in high dimensions. Submitted.

Robbins, H and Zhang, C.H. (1988) Estimating a treatment effect under biased sampling, Proc. Natl.Acad.Sci. Vol.85, pp. 3670-3672.

Silverman, E.W. (1992) Density Estimation for statistics and data analysis Chapman & Hall.

Skinner, Chris and Shlomo, Natalie (2006) Assessing identification risk in survey microdata using log-linear models. Southampton, UK, University of Southampton, Southampton Statistical Sciences Research Institute, 36pp. (S3RI Methodology Working Papers, M06/14) http://eprints.soton.ac.uk/41842/

Storey, J.D. (2003). The positive False discovery rate: a Bayesian interpretation and the q-value. *Ann.Stat.* **31** No. 6, 2013-2035.

Zhang, C.H..(2005). Estimation of sums of random variables: Examples and information bounds.*Ann.Stat*, Vol 33, No. 5. 2022-2041.