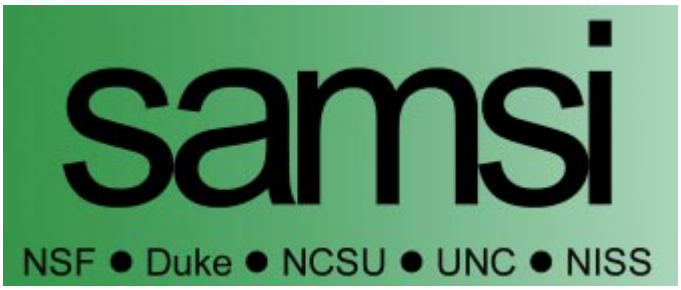

Estimating the mean of high valued observations in high dimensions

Eitan Greenshtein, Junyong Park and
Ya'acov Ritov

Technical Report #2006-7
November 22, 2006

This material was based upon work supported by the National Science Foundation under Agreement No. DMS-0112069. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation.

Statistical and Applied Mathematical Sciences Institute
PO Box 14006
Research Triangle Park, NC 27709-4006
www.samsi.info

Estimating the mean of high valued observations in high dimensions

Eitan Greenshtein

SAMSI

Junyong Park

Department of Mathematics and Statistics

Uinversity of Maryland

Ya'acov Ritov

Department of Statistics, Hebrew University

November 22, 2006

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.

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.

1 Introduction

Let $Y_i \sim N(\mu_i, 1)$, $i = 1, ..., n$ be independent observations, most of them are noise, that is the μ_i's are mostly 0. A few of them are selected for a further investigation. Without other information the selected items will be those with large value of $|Y_i|$. We want to investigate the total amount of signal in the selected lot. Formally, the main purpose of this paper is to study the estimation of the quantity

$$S = \sum_{\{i|C<Y_i\}} \mu_i, \tag{1}$$

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 now give some motivation for estimating (1).

Selection Bias and absence of a control group. The following example is an amusing version due to Herbert Robbins, of the 'motorist example' described in Zhang (2005). Suppose we want to "train" coins to land on head after being tossed. Given a large number of coins each one is tossed K times, let Y_i be the number of times coin i landed on tail. Especially "bad coins" with values of Y_i such that $Y_i > C$ are being "trained", and then tossed again the same amount of times. In order to judge whether the training was helpful, we need to compare the sum of their current means with that of their old means. In particular, estimating the sum of their old means has to do with (1). In the original motorist example, the aim is to determine whether instructions that were given to drivers with especially bad driving records in the previous year, were helpful. The 'coins' version stresses the problem of selection bias.

Such a situation is more general. Given n candidates for a clinical trial,

suppose that for ethical reasons, a treatment is given only to the subset of the n persons, consisted of those having a certain measurement $Y_i > C$. The purpose of the treatment is to reduce the corresponding $\mu_i = EY_i$. Then in order to study the treatment effect we need to estimate (1). A standard paired t-test will be affected by the selection bias.

Data mining and a sparse vector of means. Consider 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. Such a situation is typical in data mining applications where one studies and searches for a correlation of a variable of interest, with many candidate explanatory variables, in a 'shooting in the dark' manner. In such a case, attention is focused on μ_i corresponding to high valued $|Y_i|$. We concentrate, w.l.o.g., on the high valued positive Y_i. One may be interested in the "overall amount of signal" corresponding to high-valued observations Y_i, satisfying $C < Y_i$, for a given (high-valued) C.

In Section 2 we treat the estimation of (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. Note, the problem of estimating the fraction of non-zero μ_i within a subset of Y_i, has to do with the very extensively studied topic, of False Discovery Rate denoted FDR, see Benjamini and Hochberg (1995). In the above we suggest

to study the "overall amount of signal" rather than the FDR. 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 saprsity 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 $H_0^i: \ \mu_i = 0$.

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 (10) and (11) below. However, we stress that the method we develop is general, and it covers non-sparse cases.

Let Y_i be i.i.d. samples from $\sim N(\mu_i, 1)$ for $i = 1, \ldots, n$, where $\mu_i \sim G$. Recall, $S = \sum_{i=1}^n \mu_i I(Y_i > C)$. Then:

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

where $p = P(Y_i > C)$.

By equation (1.2.2) in Brown(1971),

$$E(\mu_i|Y_i = y) = y + \frac{f'(y)}{f(y)} \tag{2}$$

where $f(y) = \int \phi(y - \mu)dG(\mu)$. Hence

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

where $p = \int_C^\infty f(y)dy$. Therefore,

$$\eta = npE(\mu_i|Y_i > C) = n\left(\int_C^\infty yf(y)dx - f(C)\right) \tag{3}$$

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

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

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

First we estimate $\int_c^\infty yf(y)dy$ by

$$\frac{1}{n}\sum_{\{i:Y_i>C\}} Y_i. \tag{4}$$

It remains to estimate $f(C)$. The later is estimated using kernel density estimation technique. Let

$$\hat{f}_h(C) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{C-Y_i}{h}), \tag{5}$$

where $K = \phi$ 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., in the tail.

Consider the mean square error (MSE) of $\hat{f}(C)$:

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

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

$$\text{Bias}(\hat{f}_h(C)) \sim f^{''}(C)h^2, \tag{6}$$

while

$$\text{var}(\hat{f}_h(C)) \sim \frac{1}{nh}f(C). \tag{7}$$

Here, we use the notation $\sim$, to imply equality up to a bounded factor for

$h \to 0$. As in the standard case, the 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:

$$h_{opt} \sim \left[\frac{f(C)}{n(f''(C))^2}\right]^{\frac{1}{5}}, \tag{8}$$

and

$$E(\hat{f}_{h_{opt}}(C) - f(C))^2 \sim \left[\frac{f(C)(f''(C))^{\frac{1}{2}}}{n}\right]^{\frac{4}{5}}. \tag{9}$$

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 $C \equiv C_n = (2\alpha \log(n))^{1/2} \quad 0 < \alpha < 1$. For such points $f(C_n)$ and $f''(C_n)$, might be of order $n^{-\alpha}$. Thus, the usual interpretation of (8) and (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 (8) and (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 sparseness, are also convenient for our derivation. Assume:

$$f(C) < \kappa_1 \phi(C), \tag{10}$$

$$|f''(C)| < \kappa_2 |\phi''(C)|. \tag{11}$$

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

Remark 1

(i) Note, typically, under (10) we expect that (11) is also satisfied, e.g., if $\mu_i < 2C$ for all i, then (10) implies (11).

(ii) Assumption (10) holds for points $C \equiv C_n$, for which among the observations Y_i where $Y_i > C$, those corresponding to non-zero signals are not the vast

majority (i.e., their proportion does not approach 1.). This case is the more interesting and challenging, where it is essential to find a way to 'screen out' the zero signals.

Define now $h^0(n)$ by

$$h^0(n) \sim \left[\frac{\phi(C)}{n(\phi''(C))^2}\right]^{\frac{1}{5}}, \tag{12}$$

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

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

Finally, define our estimator $\hat{S} \equiv \hat{S}_C$ as:

$$\hat{S} = n(M - \hat{f}), \tag{13}$$

where M is defined in (4).

Theorem 1: Under conditions (10) and (11),

$$\hat{S} = \eta + O_p(n^{\frac{3}{5}}[\phi(C)(|\phi''(C)|)^{\frac{1}{2}}]^{\frac{2}{5}}) = \eta + O_p(C[n\phi(C)]^{\frac{3}{5}}) \tag{14}$$

$$\hat{S} = S + O_p(C[n\phi(C)]^{\frac{3}{5}}) \tag{15}$$

Proof: The proof of the first equality in (14) follows from the above, when

observing that the variance of $\hat{S}$ is of the order of the variance of $\hat{f}(C)$(i.e., $\text{var}(\sum_{i=1}^{n} Y_i I(Y_i > C))/n = O(\text{var}(\hat{f}(C)))$. The second equality in (14) follows by replacing $|\phi''(C)|$ by $C^2\phi(C)$.

The proof of (15), is when observing that $(S-\eta) = O_p(C[n\phi(C)]^{\frac{3}{5}})$. This follows since S is an unbiased estimator for η, with standard deviation $O_p([n\phi(C)]^{\frac{1}{2}})$. The last order for the standard deviation of S is obtained as follows. $E[\mu_i I(Y_i > C)]^2 \leq P(Y_i > C)E\mu_i^2 = O_p(\phi(C))$. Hence $\text{var}(S) = O_P(n\phi(C))$.

The last theorem gives the same approximation for the order of $(\hat{S} - \eta)$ and for that of $(\hat{S} - S)$. 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 $\hat{S}$, might get negative values. The following adjustment of $\hat{S}$ makes sense in such a sparse situation. Let

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

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

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).

Let $f(y) = \int \phi(y-\mu)dG(\mu)$. 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:

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

The estimation of $\sum \mu_i^2 I(Y_i > C)$, is now along the lines of the estimation of $\sum \mu_i I(Y_i > C)$, it involves 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 (12) of the previous section, we also consider the bandwidth, $h = 0.9An^{-1/5}$ where A=min(standard deviation, interquartile 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 (10) and (11). Yet, in our simulations, it gives very similar results to the estimator based on our suggested bandwidth (12).

We consider two more estimators for S. 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:

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

The other estimator follows the conditional maximum likelihood approach. This approach was suggested for variable selection by Greenshtein et al. (2006). The estimator is obtained through maximum likelihood conditional on $Y_i > C$. Define:

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

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

The later estimator may obtain occasionally very small (negative!) values,

when the value of Y_i is greater but very close to C, (for $Y_i = C$ 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

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

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 $\hat{S}$. We will consider

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

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 μ: (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 $(E_\mu(\hat{S}-S)^2)^{1/2}$ for different C, $C \in (2, 4)$ where $\hat{S}$ 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 (12).

Figure 1 shows the performance of the above mentioned four estimators of S, 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.

We may see that for moderately high values of C, $\hat{S}^+$ clearly dominates the other estimators (using either kernel), while for very high values of C, $\hat{S}^+$ and $\hat{S}_{con}$ are comparable. The performance of $\hat{S}^+$ 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 (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:

$$\pi_1(n) \to 0. \tag{16}$$

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

$$f(y) = \pi_1 h(y) + \pi_0 \phi(y), \tag{17}$$

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

Hence:

$$\text{fdr}(y) = P_G(\mu_i = 0|Y_i = y) = \frac{\pi_0\phi(y)}{f(y)} = \frac{\phi(y)}{f(y)}(1 + o(1)). \tag{18}$$

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

$$\widehat{\text{fdr}}(y) = \frac{\phi(y)}{\hat{f}(y)}. \tag{19}$$

In the following theorem, we consider asymptotics for a sequence $y_n \equiv y$. As in the previous section, we assume a triangular array setup, in which assumptions (10) and (11), are satisfied, where $y_n \equiv y$ plays the role of $C_n \equiv C$. Note, if we do not assume (10), then asymptotically the value of $\text{fdr}(y)$ approaches 0, while we are interested in estimating $\text{fdr}(y)$ in the non-trivial case.

Theorem 2. Let $y \equiv y_n = (2\alpha \log(n))^{1/2}$, $0 < \alpha < 1$. Assume (10), (11) and (16). Then

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

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 $f(y) \sim f''(y) \sim \phi(y) \sim \phi''(y)$ as follows from (10) and (11). Then we apply (18).

References

Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *JRSSB* **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.

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.

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.

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.

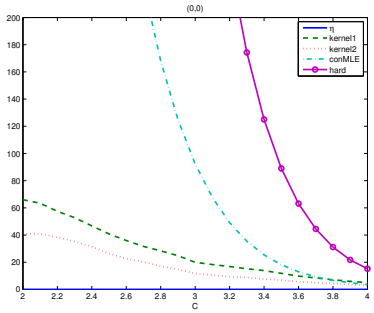

Figure 1: $(N, \mu) = (0, 0)$. Solid line is η. The others are $(E_\mu(\hat{S} - S))^{1/2}$ where dashed line:kernel1(bandwidth in Silverman(1992)), dotted line:kernel2(bandwidth with (12)), dashed-dot line:conditional mle, solid with circle:hard threshold

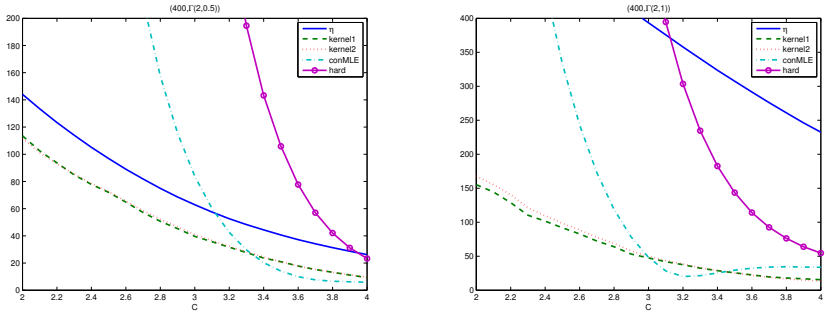

Figure 2: $(N, \mu) = (400, G = \Gamma(2, 0.5))$ and $(N, \mu) = (400, G = \Gamma(2, 1))$

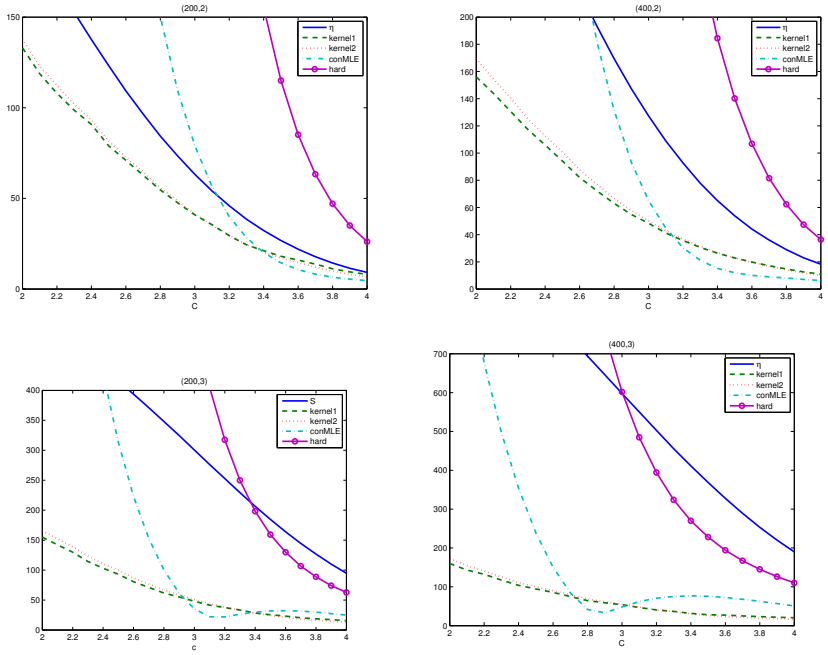

Figure 3: In the first row $(N, \mu) = (200, 2)$ and $(N, \mu) = (400, 2)$. In the second row, $(N, \mu) = (200, 3)$ and $(N, \mu) = (400, 3)$

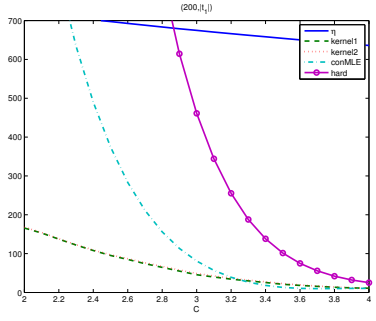

Figure 4: $(N, \mu) = (200, |t_1|)$