Generalized Shrinkage Estimators Adaptive to Sparsity and Asymmetry of High Dimensional Parameter Spaces

by

M. Zhang and D. Zhang
Purdue University

M. T. Wells
Cornell University

Technical Report #08-01

Department of Statistics
Purdue University

March 2008

Generalized Shrinkage Estimators Adaptive to Sparsity and Asymmetry of High Dimensional Parameter Spaces

Min Zhang, Dabao Zhang

Department of Statistics

Purdue University

and

Martin T. Wells

Department of Statistical Science

Cornell University

October, 2007

Abstract

In this article we address some of the fundamental issues that arise in analyzing high dimensional data and simultaneously making inference of many parameters. Shrinkage and thresholding methods have been quite successful in improving various component-wise estimators. Recently, Johnstone and Silverman (2004) developed a class of shrinkage and thresholding estimators via an empirical Bayes approach, which adapts to sparsity of the parameter space. Here we introduce and develop another class of generalized shrinkage and thresholding estimators, which are adaptive to both sparsity and asymmetry of the parameter space. We use the Bayesian approach merely as a tool to place a measure on the sparse and asymmetric parameter space, and therefore construct better decision rules adaptive to the scenario at hand. The proposed estimators have the bounded shrinkage property under a slightly broader condition than the one given by Johnstone and Silverman (2004). An empirical Bayes construction is presented for estimating multivariate normal mean. Theoretical and simulation studies demonstrate excellent performance of the proposed estimators, especially for both sparse and asymmetric high dimensional parameter space.

Keywords: Asymmetric parameter space, Bayesian estimator, empirical Bayes, shrinkage, sparse parameter space, thresholding.

1 Introduction

The compound decision problems studied by Robbins (1951) address some of the fundamental issues in analyzing high dimensional data: simultaneously making inference of p decision problems with each of which has (1) a common probability structure; (2) an observation which is independent of all the others; and (3) an unknown parameter. In the case of estimating a multivariate normal mean, Stein (1956) showed that component-wise admissible estimators failed to provide an admissible multivariate estimator when $p > 2$. Later, the results by James and Stein (1961) as well as others, led to minimax parametric Bayesian and frequentist approaches to this compound decision problem (Efron and Morris, 1972a, 1972b, 1973a, 1973b; Strawderman, 1971; Fourdrinier et al., 1998).

It is well-known that Bayesian approaches play critical roles in solving Robbins' compound decision problems. Making no appeal to dependence between the parameter values, we take a Bayesian approach merely as a way to represent the topology of the parameter space, and therefore construct coherent and more adaptive decision rules (Copas, 1969). Indeed, Samuel (1965) showed that, in the compound decision problem, all Bayes rules with the asymptotic risk convergence property must correspond to a prior distribution offering statistical dependence between the parameters.

Motivated by genomic and proteomic data analysis, we investigate estimation of a high-dimensional parameter $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_p)$ from the noisy data $\mathbf{Y}_p = (y_1, y_2, \cdots, y_p)$, where

$$y_i - \mu_i \overset{iid}{\sim} \varphi(\cdot), \qquad \varphi(\cdot) \text{ is a symmetric log-concave density function.} \tag{1.1}$$

As is often the case, a large number of the components in $\boldsymbol{\mu}$ are zero, hence $\boldsymbol{\mu}$ resides in a relatively sparse subspace. The common approaches to various problems in *omics* data analysis, such as hard and soft thresholding, are not adaptive to sparsity of the parameter space because of their fixed thresholds. Based on Stein's unbiased risk estimator (SURE) in estimation of a multivariate normal mean (Stein 1981), Donoho and Johnstone (1995) proposed the SURE method. This estimator performs well when the size of non-zero parameters is small, however, its performance is poor when the size of non-zero parameters is large. Abramovich et al. (2000) proposed a method based on Benjamini and Hochberg's

(1995) idea of controlling the false discovery rate (FDR) in multiple tests. This approach is adaptive to sparse parameter space, but its performance is very sensitive to the choice of FDR parameter. Fan and Li (2001) proposed a penalized least squares estimator with the smoothly clipped absolute deviation (SCAD) penalty, which is also adaptive to sparse parameter space but relies on two critical thresholding parameters. Due to these limitations, Johnstone and Silverman (2004) recently proposed an empirical Bayes method that is adaptive to the high-dimensional sparse parameter space, and they also demonstrated some nice theoretical properties and numerical advantages of their estimator.

Existing shrinkage and thresholding approaches inherently assume that the parameters are symmetrically located within the parameter space around the origin. Under this symmetry assumption, we presumably have the same chance to observe negative and positive values. However, many real data analyses do not support this observation. For example, the genomics dataset shown in Section 3.3.1 has only 4,936 negatives out of a total of 16,734 endogenous genes, in which the proportional of negative values is estimated to be 0.295 with standard error 0.0035. In extremal cases, we may have either all positive or all negative parameters as shown by the proteomics data in Section 3.3.2. See other examples in Li et al. (2003), Xu (2003), Finehout et al. (2005) and Zhang et al. (2005,2006). To solve this practical issue, here we investigate a more general class of estimators, which have the traditional shrinkage and thresholding approaches as special cases and are also adaptive to sparse and asymmetric parameter spaces.

2 Generalized Shrinkage and Thresholding Estimators

2.1 Definitions

Formally, as defined in Johnstone and Silverman (2004), a function $\delta(y,\tau)$ is called a *shrinkage estimator* if and only if (i) $\delta(\cdot,\tau)$ is antisymmetric (i.e., symmetric but with opposite sign); (ii) $\delta(\cdot,\tau)$ is increasing on $\mathbb{R}$ for each $\tau \geq 0$; and (iii) $\delta(y,\tau) \in [0,y]$, for all $y \geq 0$. The shrinkage estimator $\delta(y,\tau)$ is further called a *thresholding estimator* with threshold τ if and only if

$$\delta(y,\tau) = 0 \Leftrightarrow |y| \le \tau.$$

With a given threshold τ, there are usually two types of thresholding estimators: *hard threshold* refers to estimating the i-th parameter μ_i with

$$\hat{\mu}_i = \delta_{\text{hard}}(y_i,\tau) = y_i 1_{(-\infty,-\tau)\cup(\tau,\infty)}(y_i);$$

and *soft threshold* refers to estimating the i-th parameter μ_i with

$$\hat{\mu}_i = \delta_{\text{soft}}(y_i,\tau) = sign(y_i)(|y_i| - \tau)1_{(-\infty,-\tau)\cup(\tau,\infty)}(y_i),$$

where $1_A(x)$ equals to one if $x \in A$, and zero otherwise.

Johnstone and Silverman (2004) investigate the case that $\varphi(\cdot)$ is a Pólya frequency density function of order three (abbreviated PF_3) and propose a general Bayesian approach. A mixture prior is constructed to have an atom of probability at zero and a density γ with tails heavier than those of the noise density $\varphi(\cdot)$, that is,

$$\mu_i \overset{iid}{\sim} f_{prior}(\mu) = (1-w)\delta_0(\mu) + w\gamma(\mu). \tag{2.1}$$

When the tails of the noise density $\varphi(\cdot)$ are not heavier than exponential, the posterior median $\hat{\mu}(y_i;w) = \text{median}(\mu_i|y_i,w)$ is both a shrinkage estimator and a threshold estimator. This estimator also has the bounded shrinkage property relative to its threshold $\tau(w)$.

Existing shrinkage and thresholding estimators are antisymmetric, which implies an inherent assumption that the parameter space of interest is symmetric around the origin. Here we introduce a class of generalized shrinkage and thresholding estimators, which have the traditional shrinkage and thresholding estimators as special cases. The generalized shrinkage and generalized thresholding estimators are adaptive to sparse and asymmetric parameter spaces.

Definition 3.1. For $\tau_- \le 0$ and $\tau_+ \ge 0$, $\delta(y,\tau_-,\tau_+)$ is a *generalized shrinkage estimator* if it satisfies the following properties,

$$\begin{cases} \delta(y,\tau_-,\tau_+) \text{ is increasing on } y \in \mathbb{R}; \\ -|y| \le \delta(y,\tau_-,\tau_+) \le |y|, \forall y \in \mathbb{R}; \\ \delta(y,\tau) = \delta(y,-\tau,\tau) \text{ is antisymmetric for any } \tau \ge 0. \end{cases}$$

Definition 3.2. $\delta(y, \tau_-, \tau_+)$ is a *generalized thresholding estimator* if

$$\begin{cases} \delta(y, \tau_-, \tau_+) \text{ is a generalized shrinkage estimator;} \\ \delta(y, \tau_-, \tau_+) = 0 \text{ if and only if } \tau_- \leq y \leq \tau_+. \end{cases}$$

It is obvious that $\delta(y, \tau) = \delta(y, -\tau, \tau)$ is a shrinkage estimator if $\delta(y, \tau_-, \tau_+)$ is a generalized shrinkage estimator; on the other hand, $\delta(y, \tau) = \delta(y, -\tau, \tau)$ is a thresholding estimator if $\delta(y, \tau_-, \tau_+)$ is a generalized thresholding estimator. This implies that any generalized shrinkage/thresholding estimator can simply reduce to a shrinkage/thresholding estimator by setting negative and positive thresholds at the same sizes, when the data are truly symmetric. If, on the other hand, non-zero parameters can only be positive (or negative), it is more desirable to construct a generalized shrinkage/thresholding estimator $\delta(y, 0, \tau_+)$ (or $\delta(y, \tau_-, 0)$).

Corresponding to the hard thresholding estimator and soft thresholding estimator, we can define, with the thresholds (τ_-, τ_+), $\tau_- \leq 0$ and $\tau_+ \geq 0$, two types of generalized thresholding estimators: *generalized hard threshold* refers to estimating the i-th parameter μ_i with

$$\hat{\mu}_i = \delta_{\text{hard}}(y_i, \tau_-, \tau_+) = y_i 1_{(-\infty, \tau_-) \cup (\tau_+, \infty)}(y_i),$$

and *generalized soft threshold* refers to estimating the i-th parameter μ_i with

$$\hat{\mu}_i = \delta_{\text{soft}}(y_i, \tau_-, \tau_+) = (y_i - \tau_-) 1_{(-\infty, \tau_-)}(y_i) + (y_i - \tau_+) 1_{(\tau_+, \infty)}(y_i).$$

2.2 A Bayesian Approach

Here we construct generalized shrinkage and thresholding estimators for $\boldsymbol{\mu}$ in model (1.1) using a Bayesian approach. With a unimodal and symmetric distribution function $\gamma(\cdot)$, we denote $\gamma_+(\mu) = 2\gamma(\mu) 1_{[0,\infty)}(\mu)$ and $\gamma_-(\mu) = 2\gamma(\mu) 1_{(-\infty, 0]}(\mu)$. Consider a Bayesian estimator of $\boldsymbol{\mu}$ by assuming that the components of $\boldsymbol{\mu}$ have the prior,

$$\mu_i \overset{iid}{\sim} (1 - w_- - w_+)\delta_0(\mu) + w_- \gamma_-(\mu) + w_+ \gamma_+(\mu), \tag{2.2}$$

where $\delta_0(\cdot)$ is Dirac's delta function. Here w_- and w_+ are the weights for the negative and positive parts with density distributions $\gamma_-(\mu)$ and $\gamma_+(\mu)$, respectively.

Given the model and prior specification, we can obtain the posterior distribution of each parameter. Let $\tilde{w}_-$ and $\tilde{w}_+$ denote the posterior probabilities of μ_i being positive and negative respectively. Then it follows that

$$\begin{cases} \tilde{w}_+(y_i; w_-, w_+) = \frac{w_+ g_+(y_i)}{(1-w_- - w_+)\varphi(y_i) + w_+ g_+(y_i) + w_- g_-(y_i)} \\ \tilde{w}_-(y_i; w_-, w_+) = \frac{w_- g_-(y_i)}{(1-w_- - w_+)\varphi(y_i) + w_+ g_+(y_i) + w_- g_-(y_i)}, \end{cases}$$

where

$$\begin{cases} g_-(y_i) = \int_{-\infty}^{0} \varphi(y_i - \mu)\gamma_-(\mu) d\mu \\ g_+(y_i) = \int_{0}^{\infty} \varphi(y_i - \mu)\gamma_+(\mu) d\mu. \end{cases} \quad (2.3)$$

Let f_- and f_+ denote the posterior conditional densities for the positive and negative parts respectively, i.e.,

$$\begin{cases} f_+(\mu_i | y_i; w_-, w_+) \triangleq [\mu_i | y_i, \mu_i > 0] = \frac{\varphi(y_i - \mu_i)\gamma_+(\mu_i)}{g_+(y_i)} \\ f_-(\mu_i | y_i; w_-, w_+) \triangleq [\mu_i | y_i, \mu_i < 0] = \frac{\varphi(y_i - \mu_i)\gamma_-(\mu_i)}{g_-(y_i)}. \end{cases}$$

Then the posterior distribution of the parameter μ_i, given the observed value y_i, is

$$\begin{aligned} \mu_i | y_i, w_-, w_+ \sim & \{1 - \tilde{w}_-(y_i; w_-, w_+) - \tilde{w}_+(y_i; w_-, w_+)\}\delta_0(\mu_i) \\ & + \tilde{w}_+(y_i; w_-, w_+) f_+(\mu_i | y_i; w_-, w_+) \\ & + \tilde{w}_-(y_i; w_-, w_+) f_-(\mu_i | y_i; w_-, w_+). \end{aligned}$$

For fixed w_- and w_+, a Bayesian estimator of μ_i (under componentwise absolute error loss) is to use its posterior median, i.e.,

$$\hat{\mu}(y_i; w_-, w_+) = \text{median}(\mu_i | y_i; w_-, w_+). \quad (2.4)$$

Obviously, the performance of the estimator $\hat{\mu}(y_i; w_-, w_+)$ depends on the choices of the hyperparameters (w_-, w_+), the density distribution of the noise, and (γ_-, γ_+) in the prior (2.2). Here (w_-, w_+) describe not only the asymmetry but also the sparsity of the parameter space. Intuitively, an optimal (w_-, w_+) may be elicited by maximizing the marginal likelihood function. As we will show below that the estimator $\hat{\mu}(y_i; w_-, w_+)$ is a generalized shrinkage estimator and a generalized thresholding estimator when (w_-, w_+) lie in a region predetermined by the following constant,

$$a = \frac{\varphi(0)}{g_+(0) + \varphi(0)} \in (0, 1). \quad (2.5)$$

Since $\varphi(\cdot)$ is log-concave, it has at most an exponential tail. The symmetric and unimodal $\gamma(\cdot)$ implies a decreasing function γ_+ on $\mathbb{R}_+ = [0,\infty)$ and an increasing function γ_- on $\mathbb{R}_- = (-\infty, 0]$. Therefore, the constant a implies the flatness of the priors $\gamma_+(\cdot)$ and $\gamma_-(\cdot)$ relative to the density of the noise. For each $a \in [0,1]$, we define the simplex,

$$\mathcal{S}(a) = \{(w_-, w_+) \in [0,1]^2 : (2a-1)w_- + w_+ \le a, w_- + (2a-1)w_+ \le a\}. \tag{2.6}$$

Theorem 2.1. *With the constant a in (2.5) and the simplex $\mathcal{S}(a)$ in (2.6): (i) $\hat{\mu}(y; w_-, w_+)$ is a generalized shrinkage estimator if and only if $(w_-, w_+) \in \mathcal{S}(a)$; (ii) $\hat{\mu}(y; w_-, w_+)$ is a generalized thresholding estimator if and only if $(w_-, w_+) \in \mathcal{S}(a)$.*

We will prove this theorem in Section 4. It is interesting to observe that, for any $a \in (0,1)$, we always have $\mathcal{S}(0) = \{(w/2, w/2) : 0 \le w \le 1\} \subset \mathcal{S}(a)$. Hence, the antisymmetric estimator considered by Johnstone and Silverman (2004) is a special case of our estimator. Note that Johnstone and Silverman (2004) develop their shrinkage/threshold estimator by requiring a PF_3 density $\varphi(\cdot)$ of the noise. Theorem 2.1 shows that their estimator is still a shrinkage estimator as well as a thresholding estimator when φ is log-concave, or equivalently PF_2.

Theorem 2.1 also states a sufficient and necessary condition for our proposed Bayesian estimator $\hat{\mu}(y; w_-, w_+)$ to be a generalized shrinkage/threshold estimator, i.e., $(w_-, w_+) \in \mathcal{S}(a)$. This can be visualized using Figure 1. As mixing weights, (w_-, w_+) can be ultimately any point within the area under the line $w_+ + w_- = 1$, which is defined by $\mathcal{S}(1)$ and corresponds to the grey area in Figure 1. Theorem 2.1 says that, in order for the Bayesian estimator $\hat{\mu}(y; w_-, w_+)$ to have the same sign as the observed data y, (w_-, w_+) needs to be chosen from the shaded area, i.e., the intersection area under the two solid lines that defines $\mathcal{S}(a)$. However, the antisymmetric Bayesian estimator developed by Johnstone and Silverman (2004) essentially requires $(w_-, w_+) \in \mathcal{S}(0)$, i.e., the lower part of the dashed line lying completely in the shaded area (see Figure 1). Obviously, the Bayesian estimator developed here gains more flexibility by offering a much larger space for (w_-, w_+).

When $\gamma_+(\cdot)$ and $\gamma_-(\cdot)$ approach to the noninformative priors such that $\gamma_+(0) \to 0$ and $\gamma_-(0) \to 0$, the constant a defined by (2.5) essentially goes to one, which puts less constraint

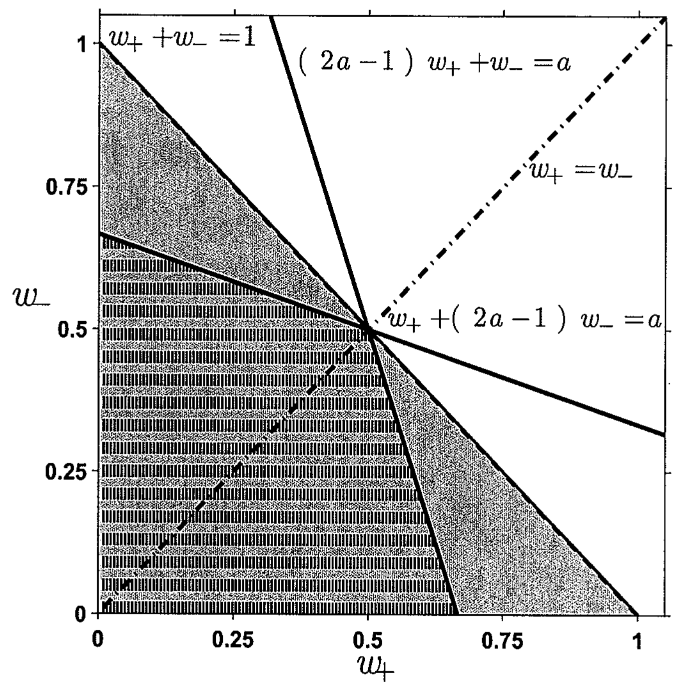

Figure 1: Visualization of Choices for (w_-, w_+)

for the above Bayesian estimator to be a generalized shrinkage/thresholding estimator. However, as shown in the following theorem, the tails of the priors cannot be too heavy in order for this estimator to have the bounded shrinkage property, i.e., the size of the shrinkage needs to be bounded such that the large signals will not be shrunk too much to be indistinguishable from noise.

Theorem 2.2. *Assume that: (i) there exists $\rho > 0$ such that $\varphi(y)\exp\{\rho y\}$ is decreasing for sufficiently large y; and (ii) there exists $\Lambda > 0$ and $M > 0$ such that*

$$\sup_{u>M}\left|\frac{d}{du}\log\gamma_+(u)\right| \le \Lambda < \rho. \tag{2.7}$$

Then, when $(w_-, w_+) \in \mathcal{S}(a)$, there exists a constant c such that, for all w_-, w_+, and y, the generalized shrinkage estimator $\hat{\mu}(y; w_-, w_+)$ has the following bounded shrinkage property

$$\begin{cases} y - \hat{\mu}(y; w_-, w_+) \le \tau_+(w_-, w_+) + c & \text{for } y \ge 0 \\ \hat{\mu}(y; w_-, w_+) - y \le -\tau_-(w_-, w_+) + c & \text{for } y \le 0, \end{cases}$$

where the thresholds $\tau_-(w_-, w_+)$ *and* $\tau_+(w_-, w_+)$ *are defined by* $\tilde{w}_-(\tau_-(w_-, w_+); w_-, w_+) = 0.5$ *and* $\tilde{w}_+(\tau_+(w_-, w_+); w_-, w_+) = 0.5$*, respectively.*

Remark 1: Since $\varphi(\cdot)$ is log-concave, it has at most an exponential tail and the first assumption in the above theorem explicitly states that its tail cannot be heavier than the exponential $\exp\{-\rho y\}$.

Remark 2: Since $\gamma(\cdot)$ is unimodal and symmetric, $\log \gamma_+(u)$ is decreasing when $u > 0$, so the second assumption in the above theorem implies that $\frac{d}{du} \log \gamma_+(u) \geq -\Lambda > -\rho$, i.e., the tail of $\gamma_+(\cdot)$ is heavier than that of the noise density $\varphi(\cdot)$.

Remark 3: With Gaussian error, i.e., $\varphi(\cdot) = \phi(\cdot)$, ρ in the first assumption can be chosen arbitrarily large, so the second assumption essentially places no extra constraint on $\gamma_+(\cdot)$ (or on $\gamma_-(\cdot)$).

Remark 4: As shown later, for any $(w_-, w_+) \in \mathcal{S}(a)$, there exist $\tau_-(w_-, w_+) < 0$ and $\tau_+(w_-, w_+) > 0$ such that $\tilde{w}_+(\tau_+(w_-, w_+); w_-, w_+) = 0.5$ and $\tilde{w}_-(\tau_-(w_-, w_+); w_-, w_+) = 0.5$, respectively, and therefore $\hat{\mu}(y; w_-, w_+) = 0$ if and only if $y \in [\tau_-(w_-, w_+), \tau_+(w_-, w_+)]$.

Remark 5: As shown by Lemma 4.2 for $y > 0$, $\tilde{w}_+(y; w_-, w_+)$ is an increasing function in both w_- and w_+, which implies that $\tau_+(w_-, w_+)$ is a decreasing function in either w_- or w_+. Similarly, $\tau_-(w_-, w_+)$ is an increasing function in either w_- or w_+. Therefore, we have $0 \leq \tau_+(w_-, w_+) \leq \tau_+(0, w_+)$, $0 \geq \tau_-(w_-, w_+) \geq \tau_-(w_-, 0)$, and

$$\tau_+(\max\{w_-, w_+\}, \max\{w_-, w_+\}) \leq \tau_+(w_-, w_+) \leq \tau_+(\min\{w_-, w_+\}, \min\{w_-, w_+\}),$$
$$\tau_-(\min\{w_-, w_+\}, \min\{w_-, w_+\}) \leq \tau_-(w_-, w_+) \leq \tau_-(\max\{w_-, w_+\}, \max\{w_-, w_+\}).$$

3 Illustrations of the Theory

In the case that $\varphi(\cdot)$ in model (1.1) is the density function of a standard normal distribution, i.e., $\varphi(\cdot) = \phi(\cdot)$, Johnstone and Silverman (2004) developed an empirical Bayes thresholding estimator (which is called EB hereafter) based on a quasi-Cauchy prior for $\boldsymbol{\mu}$, and compared this estimator with others in the literature. The EB estimator showed excellent performance. Here we will adopt this quasi-Cauchy prior to construct a generalized thresholding estimator (which is called GEB hereafter), and compare it with the EB estimator, with a simulation study, in terms of risks and numbers of false discoveries.

3.1 A Construction with Quasi-Cauchy Prior

We can construct the generalized shrinkage/threshold estimators with a quasi-Cauchy prior, i.e., taking

$$\begin{cases} \gamma_+(\mu|\theta_+) = 2(\frac{1}{\theta_+} - 1)^{-1/2}\phi(\frac{\mu}{1/\theta_+ - 1})1_{[0,\infty)}(\mu), & \theta_+ \sim \text{Beta}(0.5, 1), \\ \gamma_-(\mu|\theta_-) = 2(\frac{1}{\theta_-} - 1)^{-1/2}\phi(\frac{\mu}{1/\theta_- - 1})1_{(-\infty,0]}(\mu), & \theta_- \sim \text{Beta}(0.5, 1), \end{cases} \tag{3.1}$$

or equivalently,

$$\begin{cases} \gamma_+(\mu) = \sqrt{\frac{2}{\pi}}\left(1 - \frac{\mu(1-\Phi(\mu))}{\phi(\mu)}\right)1_{[0,\infty)}(\mu), \\ \gamma_-(\mu) = \sqrt{\frac{2}{\pi}}\left(1 + \frac{\mu\Phi(\mu)}{\phi(\mu)}\right)1_{(-\infty,0]}(\mu), \end{cases}$$

which have tails similar to those of Cauchy densities, i.e., much heavier than Gaussian distribution as desired.

Assuming that $\varphi(\cdot) = \phi(\cdot)$ in model (1.1), we then have,

$$g_+(y_i) = \frac{1}{y_i^2\sqrt{2\pi}}\left(2\Phi(y_i) - \exp(-y_i^2/2) - \frac{2y_i\exp(-y_i^2/2)}{\sqrt{2\pi}}\right)$$

and

$$g_-(y_i) = \frac{1}{y_i^2\sqrt{2\pi}}\left(2(1 - \Phi(y_i)) - \exp(-y_i^2/2) + \frac{2y_i\exp(-y_i^2/2)}{\sqrt{2\pi}}\right).$$

Since $\phi(0) = 1/\sqrt{2\pi}$ and $g_+(0) = \lim_{y\downarrow 0} g_+(y) = 1/\sqrt{8\pi}$, we have $a = 2/3$ from (2.5), thus $\mathcal{S}(a)$ is defined by

$$\begin{cases} w_+ + 3w_- \leq 2 \\ 3w_+ + w_- \leq 2. \end{cases}$$

Maximizing the marginal distribution of $\mathbf{Y}_p$ for $(w_-, w_+) \in \mathcal{S}(2/3)$, we then construct a generalized thresholding estimator with the posterior median, which is essentially an empirical Bayes estimator.

3.2 Simulation Study

Here we conduct a simulation study to evaluate the GEB estimator, and compare it with the EB estimator. Assuming $\varphi(\cdot) = \phi(\cdot)$ in model (1.1), we simulate 1000 datasets in each setting of the parameter $\boldsymbol{\mu} = (\mu_1, \mu_2, \cdots, \mu_p)$ with $p = 1000$, and estimate all parameters using both estimators. The risk $R(q) = \sum_{i=1}^{p} E[|\hat{\mu}_i - \mu_i|^q]$ is computed for different values of $q \in (0, 2]$, and the number of false positives (NFP) and the number of false negatives (NFN) are also summarized for comparison.

Define $p_+ = \#\{i : \mu_i > 0\}$ and $p_- = \#\{i : \mu_i < 0\}$. We uniformly specify all positive parameters with values μ_+, and all negative parameters with values μ_-, i.e., $\mu_i = \mu_+$ if $\mu_i > 0$, and $\mu_i = \mu_-$ if $\mu_i < 0$. With sparse non-zero parameters, we consider different cases of asymmetric parameter spaces: one with $|\mu_-| = \mu_+$ but $p_+ \neq p_-$ (see Table 1); one with $|\mu_-| \neq \mu_+$ but $p_+ = p_-$ (see Table 2); another one with $|\mu_-| \neq \mu_+$ and $p_+ \neq p_-$ (see Table 2). For each parameter setting, we report $R(2)$, $R(1)$, NFP, and NFN in Table 1 and Table 2.

With the exception that $|\mu_-| = \mu_+$, the GEB estimator always has smaller risks, in terms of $R(2)$ and $R(1)$, than the EB estimator. The GEB estimator gets much more gain over the EB estimator when the size of non-zero parameters is smaller. As the size of non-zero parameters gets larger, it is much easier to differentiate them from noise, so the GEB estimator can lose its advantage over the EB estimator for large non-zero parameters. With the same size of non-zero parameters, the more p_- and p_+ differ, the more gain of the GEB estimator over the EB estimator in terms of $R(2)$ and $R(1)$. When both $|\mu_-| \neq \mu_+$ and $p_- \neq p_+$, the gain of the GEB estimator over the EB estimator depends more on the difference between p_- and p_+, which is reasonable as only the later is modeled in the specified prior (2.2). When both $|\mu_-| = \mu_+$ and $p_- = p_+$, i.e., the parameter space is symmetric, the EB estimator is slightly better than the GEB estimator in terms of $R(2)$ and $R(1)$. In this symmetric case, the ignorable gain of the EB estimator over the GEB estimator is due to the variation in the observation, which results in slight difference between estimated w_- and w_+.

In the case that $|\mu_-| \neq \mu_+$, we have different observations on the performance of the two estimators when evaluating them in terms of NFP and NFN. While the NFP of the GEB

Table 1: Comparing the EB and GEB Estimators with $|\mu_-| = \mu_+$

(μ_+,μ_-)	Criterion	(p_+,p_-)							
		(50, 0)		(50, 5)		(50, 20)		(50, 50)	
		EB	GEB	EB	GEB	EB	GEB	EB	GEB
(2,-2)	$R(2)$	195.43	186.37	214.20	206.43	268.83	265.14	372.81	372.82
	$R(1)$	98.19	94.84	107.73	104.87	135.71	134.35	189.62	189.70
	NFP	0.42	0.92	0.49	0.92	0.77	1.01	1.51	1.61
	NFN	47.10	43.70	51.50	48.62	64.16	62.77	87.82	87.70
(3,-3)	$R(2)$	266.86	219.48	285.77	255.29	337.46	329.29	424.59	425.16
	$R(1)$	102.02	90.27	110.24	102.79	133.77	131.88	177.16	177.38
	NFP	2.57	4.02	3.00	3.97	4.48	4.92	8.49	8.62
	NFN	25.22	18.42	26.66	22.18	30.16	28.94	34.60	34.65
(4,-4)	$R(2)$	173.96	141.78	184.91	169.97	215.73	212.67	270.98	271.40
	$R(1)$	72.98	67.08	79.20	76.53	97.61	97.13	133.97	134.15
	NFP	5.24	6.46	6.01	6.96	8.85	9.32	16.46	16.65
	NFN	5.57	3.19	5.64	4.55	5.67	5.44	5.13	5.16
(5,-5)	$R(2)$	102.67	92.40	111.24	107.23	135.89	135.14	185.44	185.69
	$R(1)$	58.15	56.51	63.85	63.23	81.04	80.98	116.22	116.38
	NFP	6.44	7.55	7.43	8.35	10.83	11.39	19.92	20.15
	NFN	0.52	0.23	0.53	0.45	0.44	0.45	0.34	0.34
(6,-6)	$R(2)$	81.84	78.08	89.78	87.85	113.27	112.81	160.63	160.83
	$R(1)$	54.03	53.35	59.56	59.23	76.31	76.28	110.65	110.81
	NFP	6.74	7.75	7.75	8.65	11.27	11.84	20.65	20.84
	NFN	0.03	0.01	0.03	0.03	0.02	0.03	0.01	0.01

estimator is always larger than that of the EB estimator, the NFN of the GEB estimator is always smaller than that of the EB estimator. However, the difference between NFNs of the two estimators is usually larger than the difference between NFPs of the two estimators, especially when the parameter space is strongly asymmetric, e.g., $(p_-, p_+) = (50, 0)$ and $(\mu_-, \mu_+) = (-3, 3)$. Accounting for the sparsity of non-zero parameters, NFN may be appreciated more than NFP as long as NFP is not too large. In this situation, the GEB estimator has its advantage over the EB estimator. When $|\mu_-| = \mu_+$ and $p_- = p_+$, both NFPs and NFNs of the two estimators are close enough to ignore the difference.

In each case of $p_- \neq p_+$, the gain of the GEB estimator in term of the risk $R(q)$ gets smaller when q decreases in $(0, 2]$, and essentially disappears when q approaches to zero (see Figure 2). On the other hand, the GEB estimator usually has smaller NFN but slightly

Table 2: Comparing the EB and GEB Estimators with $|\mu_-| \neq \mu_+$

(μ_+, μ_-)	Criterion	(p_+, p_-) (50, 5) EB	(50, 5) GEB	(50, 25) EB	(50, 25) GEB	(50, 50) EB	(50, 50) GEB
(3,-2)	$R(2)$	281.99	239.85	341.36	317.37	412.95	403.99
	$R(1)$	110.53	100.29	144.15	139.07	185.17	184.13
	NFP	2.72	3.92	3.44	3.82	4.53	4.54
	NFN	29.02	23.36	43.68	41.79	60.91	61.55
(3,-4)	$R(2)$	275.33	248.09	305.62	300.46	340.52	345.89
	$R(1)$	106.88	99.56	126.92	125.43	153.08	154.76
	NFP	3.25	4.22	6.60	6.89	12.31	12.27
	NFN	24.41	19.78	21.44	20.25	18.14	19.30
(3,-5)	$R(2)$	265.24	233.24	269.58	263.51	287.43	298.92
	$R(1)$	104.66	96.36	119.20	117.70	141.82	144.78
	NFP	3.38	4.35	7.31	7.58	13.94	13.71
	NFN	23.50	18.49	18.61	17.54	14.44	16.42
(3,-6)	$R(2)$	261.83	227.27	259.20	252.97	272.06	285.24
	$R(1)$	104.00	95.42	117.05	115.59	138.48	141.73
	NFP	3.41	4.52	7.46	7.74	14.26	14.00
	NFN	23.34	19.19	18.25	17.25	14.02	16.15

larger NFP. This implies that the GEB estimator tends to be reasonably larger than the EB estimator in term of the size.

3.3 Application to Omics Data

3.3.1 Genomics Data

A common issue in genomic study with microarray data is to identify genes differentially expressed under different conditions. For example, van de Peppel et al. (2003) designed a microarray experiment to examine, in comparison to non-heat-shock cells, the heat-shock response of primarily cultured human umbilical vein endothelial cells (HUVECs). The datasets are available from ArrayExpress (http://www.ebi.ac.uk/aerep/) with accession number E-UMCU-2. Here we only use the dataset collected three hours after heat shock for illustration. In this dataset, the differential expression levels of a total of 16,734 endogenous genes are under investigation (van de Peppel et al. 2003) and are normalized using the 960 external

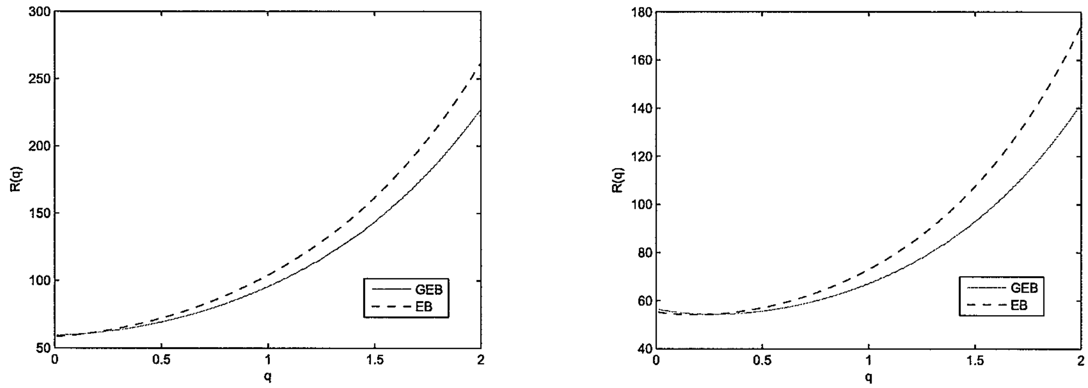

Figure 2: Comparing the Risks of the EB and GEB Estimators with $(\mu_-, \mu_+) = (-6, 3)$ and $(p_-, p_+) = (50, 5)$ (Left), and $(\mu_-, \mu_+) = (-4, 4)$ and $(p_-, p_+) = (0, 50)$ (Right).

control genes (Zhang et al. 2006).

Figure 3 showed the results using the EB and GEB estimators respectively. All genes with $y_i \in (-0.3, 0.3)$ have estimated μ_i equal to zero, and therefore are not shown here. A positive μ_i means that the i-th gene is up-regulated, a negative μ_i means that it is down-regulated, and a zero μ_i means that it is not differentially expressed. In total, using the GEB estimator, 3,275 genes were identified to be up-regulated and 139 genes were identified to be down-regulated. Instead, the EB estimator identified 53 more genes to be down-regulated but 1,947 less genes to be up-regulated, as the symmetry assumption forces the upper and lower thresholds to be the same, as shown in Figure 3. On the other hand, the GEB estimator relaxes this assumption and thus allows unequal thresholds. While for small and large scale y_i, whether positive or negative, the two estimates coincide with each other, their performances on medium scale y_i are quite different due to the asymmetric data.

3.3.2 Proteomics Data

Mass spectrometry (MS) plays an important role in discovering clinically relevant peptides/proteins and eventually understanding biological cancer processes. After preprocessing, experimental MS data present many peaks residing on certain mass-to-charge ratios (m/z), which is a result of either chemical background noise or peptide fragments. A critical step in

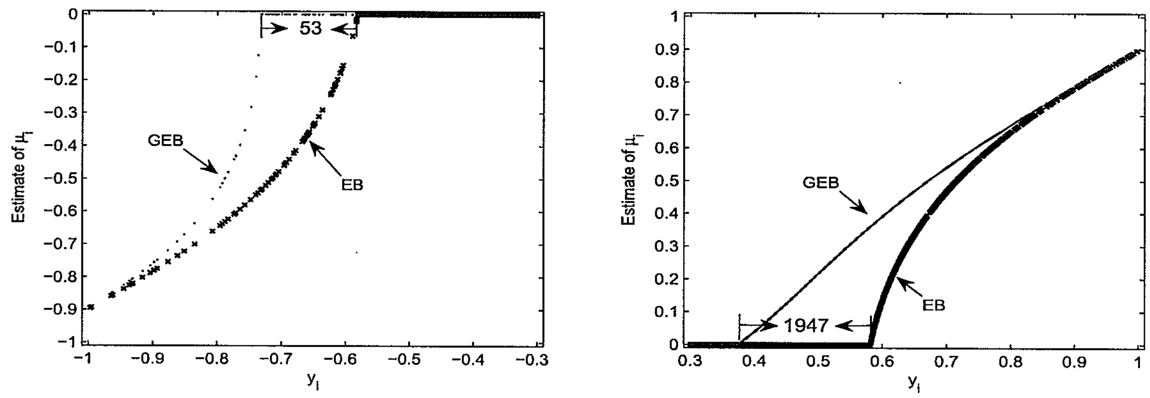

Figure 3: Identified Up-Regulated Genes (Left) and Down-Regulated Genes (Right) Using the EB and GEB Estimators Respectively.

identifying proteins using experimental MS data is to remove those peaks caused by chemical background noise while keeping the peaks corresponding to peptide fragments so as to match them with proteins in databases.

Shown in the top panel of Figure 4 is an experimental MS dataset from Keller et al. (2002). The log-intensity values, after preprocessing and normalization, are shown in the cental panel of Figure 4. Each point in the top panel corresponds to a peak observed in MS. The EB estimator identified 14 peaks with estimated hyperparameters $\hat{w}_- = \hat{w}_+ = 0.0639$, on the other hand, the GEB estimator identified 33 peaks with the estimated hyperparameters $\hat{w}_- = 0$ and $\hat{w}_+ = 0.2767$ (see Figure 4). All the peaks identified here observed positive log-transformed intensities. Apparently, forcing $w_- = w_+$ in the EB estimator has significantly reduced the number of identified peaks, which will also reduce the efficiency of searching proteins in databases.

4 Proofs of the Theorems

4.1 Proof of Theorem 2.1

We will first show some preliminary results of the related functions, and then proceed the proof of Theorem 2.1 by integrating these results.

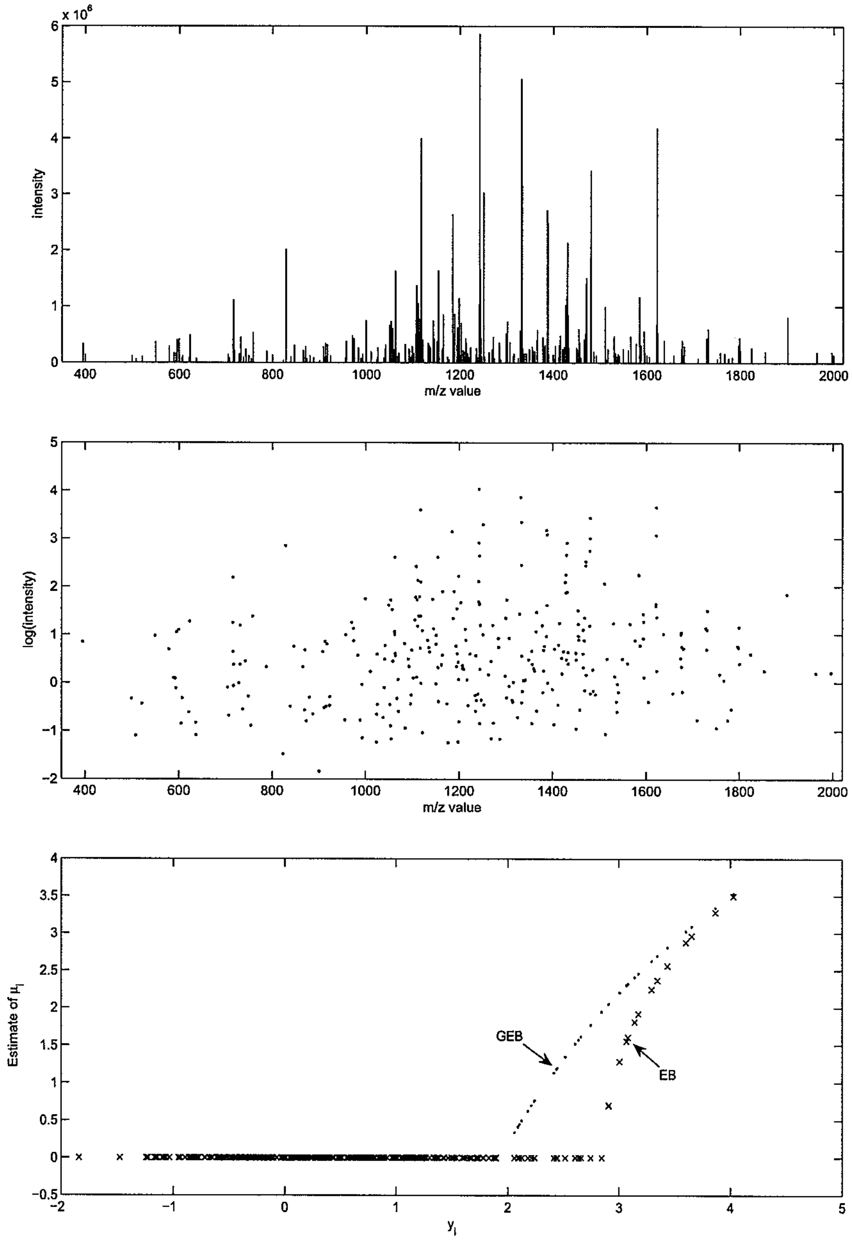

Figure 4: An Experimental MS Dataset (Top), Normalized Peaks (Central) and Identified Peaks (Bottom).

Lemma 4.1. *If $\varphi(\cdot)$ is symmetric and log-concave, and $\gamma(\cdot) = [\gamma_-(\cdot) + \gamma_+(\cdot)]/2$ is unimodal and symmetric, then (i) $g_+(y) = g_-(-y)$, $y \in \mathbb{R}$; (ii) $g_+(y)/\varphi(y)$ is increasing on $\mathbb{R}$; (iii) $g_-(y)/\varphi(y)$ is decreasing on $\mathbb{R}$.*

Proof. (1) It follows the facts that $\varphi(\cdot)$ is symmetric and $\gamma_+(\mu) = \gamma_-(-\mu)$.

(2) Since $\varphi(\cdot)$ is log-concave, it is PF_2. That is, for all $v \geq 0$ and $y_1 \leq y_2$, we have

$$\varphi(y_1)\varphi(y_2 - v) \geq \varphi(y_2)\varphi(y_1 - v),$$

which implies that $\varphi(y - v)/\varphi(y)$ is increasing in y for all $v \geq 0$. Hence, $g_+(y)/\varphi(y)$ is increasing on $y \in \mathbb{R}$ since

$$\frac{g_+(y)}{\varphi(y)} = \int_0^\infty \frac{\varphi(y - v)}{\varphi(y)} \gamma_+(v) dv.$$

(3) Similarly, for all $u \leq 0$ and $y_1 \leq y_2$, we have

$$\varphi(y_1 - u)\varphi(y_2) \geq \varphi(y_1)\varphi(y_2 - u),$$

which implies that $\varphi(y - u)/\varphi(y)$ is decreasing in y for all $u \leq 0$. The conclusion follows the fact that

$$\frac{g_-(y)}{\varphi(y)} = \int_{-\infty}^0 \frac{\varphi(y - u)}{\varphi(y)} \gamma_-(u) du.$$

□

Note that

$$\tilde{w}_+(y; w_-, w_+) = \frac{w_+}{w_+ + (1 - w_- - w_+)/(\frac{g_+(y)}{\varphi(y)}) + w_- \frac{g_-(y)}{\varphi(y)}/(\frac{g_+(y)}{\varphi(y)})},$$

$$\tilde{w}_-(y; w_-, w_+) = \frac{w_-}{w_- + (1 - w_- - w_+)/(\frac{g_-(y)}{\varphi(y)}) + w_+ \frac{g_+(y)}{\varphi(y)}/(\frac{g_-(y)}{\varphi(y)})}.$$

Therefore, Lemma 4.1 leads to the following lemma.

Lemma 4.2. *Assume $(w_-, w_+) \in \mathcal{S}(1)$. With the same $\varphi(\cdot)$ and $\gamma(\cdot)$ as in Lemma 4.1, (i) $\tilde{w}_+(y; w_-, w_+)$ is increasing in $y \in \mathbb{R}$; (ii) $\tilde{w}_-(y; w_-, w_+)$ is decreasing in $y \in \mathbb{R}$.*

Now we will use Lemma 4.2 to identify the conditions for our estimator to have the same sign as the observed data.

Proposition 4.3. *With the same $\varphi(\cdot)$ and $\gamma(\cdot)$ as in Lemma 4.1, the Bayesian estimator $\hat{\mu}(y; w_-, w_+)$ has the following properties, (i) $\hat{\mu}(y; w_-, w_+) \geq 0$, $\forall y \geq 0$, if and only if $w_- + (2a-1)w_+ \leq a$; (ii) $\hat{\mu}(y; w_-, w_+) \leq 0$, $\forall y \leq 0$, if and only if $(2a-1)w_- + w_+ \leq a$.*

Proof. Note that $g_+(0) = g_-(0)$. This proposition follows from Lemma 4.2 and the fact that

(i) $\hat{\mu}(y; w_-, w_+) \geq 0$, $\forall y \geq 0$, if and only if $\tilde{w}_-(0; w_-, w_+) \leq \frac{1}{2}$;

(ii) $\hat{\mu}(y; w_-, w_+) \leq 0$, $\forall y \leq 0$, if and only if $\tilde{w}_+(0; w_-, w_+) \leq \frac{1}{2}$.

□

Remarks: (1) If $g_+(0) \to 0$, then $\varphi(0)/\{g_+(0)+\varphi(0)\} \to 1$, and both necessary and sufficient conditions in the above proposition reduce to $w_- + w_+ \leq 1$, i.e., $(w_-, w_+) \in \mathcal{S}(1)$; (2) Since $\tilde{w}_+(y; w_-, 0) \equiv 0$ and $\tilde{w}_-(y; 0, w_+) \equiv 0$, $\hat{\mu}(y; w_-, 0) \leq 0$ for all $y \in \mathbb{R}$ and $\hat{\mu}(y; 0, w_+) \geq 0$ for all $y \in \mathbb{R}$; (3) Since $\tilde{w}_-(0; a, 0) = 0.5$ and $\tilde{w}_+(0; 0, a) = 0.5$, we have $\hat{\mu}(y; a, 0) < 0$ for all $y < 0$ and $\hat{\mu}(y; 0, a) > 0$ for all $y > 0$.

Proposition 4.4. *Assume $(w_-, w_+) \in \mathcal{S}(a)$. With the same $\varphi(\cdot)$ and $\gamma(\cdot)$ as in Lemma 4.1, we have $|\hat{\mu}(y; w_-, w_+)| \leq |y|$ for all y.*

Proof. With Proposition 4.3, it suffices to prove that, for any $y > 0$, $\hat{\mu}(y; w_-, w_+) \leq y$, or equivalently,

$$p(\mu > y | Y = y; w_-, w_+) \leq \frac{1}{2}, \forall y > 0.$$

Note that, for $y > 0$,

$$p(\mu > y | Y = y; w_-, w_+) = \frac{w_+ \int_y^\infty \varphi(y-\mu)\gamma_+(\mu)d\mu}{(1 - w_- - w_+)\varphi(y) + w_+ g_+(y) + w_- g_-(y)},$$

By Lemma 4.1, we have, for $y > 0$,

$$\frac{g_-(y)}{\varphi(y)} \leq \frac{g_-(0)}{\varphi(0)} \Rightarrow \varphi(y) \geq \frac{\varphi(0)}{g_-(0)} g_-(y).$$

Therefore,

$$\begin{aligned} p(\mu > y|Y = y; w_-, w_+) &\leq \frac{w_+ \int_y^\infty \varphi(y-\mu)\gamma_+(\mu)d\mu}{(1-w_- - w_+)\frac{\varphi(0)g_-(y)}{g_-(0)} + w_+ g_+(y) + w_- g_-(y)} \\ &= \frac{\int_y^\infty \varphi(y-\mu)\gamma_+(\mu)d\mu}{g_+(y) + g_-(y) + \frac{g_-(y)}{(1-a)w_+}\left[a - w_+ - (2a-1)w_-\right]} \\ &\leq \frac{\int_y^\infty \varphi(y-\mu)\gamma_+(\mu)d\mu}{g_+(y) + g_-(y)}, \end{aligned}$$

where the last inequality holds because $w_+ + (2a-1)w_- \leq a \leq 1$.

We will prove that,

$$\frac{\int_y^\infty \varphi(y-\mu)\gamma_+(\mu)d\mu}{g_+(y) + g_-(y)} \leq \frac{1}{2}, \forall y > 0 \tag{4.1}$$

which concludes the proof of the proposition. Since $\gamma(\mu) = \frac{1}{2}[\gamma_+(\mu) + \gamma_-(\mu)]$ is unimodal and symmetric, then for $y > 0, t \geq 0$,

$$\begin{aligned} &\gamma_+(y-t) + \gamma_-(y-t) \geq \gamma_+(y+t) + \gamma_-(y+t) \\ \Rightarrow &\int_0^\infty \varphi(t)[\gamma_+(y-t) + \gamma_-(y-t)]dt \geq \int_0^\infty \varphi(t)[\gamma_+(y+t) + \gamma_-(y+t)]dt. \end{aligned}$$

Note that

$$\begin{aligned} \int_0^\infty \varphi(t)[\gamma_+(y-t) + \gamma_-(y-t)]dt &= \int_{-\infty}^y \varphi(y-u)[\gamma_+(u) + \gamma_-(u)]du, \\ \int_0^\infty \varphi(t)[\gamma_-(y+t) + \gamma_+(y+t)]dt &= \int_y^\infty \varphi(y-u)[\gamma_+(u) + \gamma_-(u)]du, \end{aligned}$$

and

$$g_+(y) + g_-(y) = \int_{-\infty}^\infty \varphi(y-u)[\gamma_+(u) + \gamma_-(u)]du.$$

Therefore,

$$\begin{aligned} g_+(y) + g_-(y) &\geq 2\int_y^\infty \varphi(y-u)[\gamma_+(u) + \gamma_-(u)]du \\ &= 2\int_y^\infty \varphi(y-u)\gamma_+(u)du. \end{aligned}$$

Then, we have the inequality (4.1).

□

With the proof by Johnstone and Silverman (2004, p.1619), we can easily establish the following proposition.

Proposition 4.5. *If $\varphi(\cdot)$ is symmetric and log-concave, then $\hat{\mu}(y; w_-, w_+)$ is increasing in $y \in \mathbb{R}$.*

The combination of Proposition 4.4 and Proposition 4.5 proves the first part of Theorem 2.1, i.e., $\hat{\mu}(y; w_-, w_+)$ is a generalized shrinkage estimator. The second part of Theorem 2.1, i.e., $\hat{\mu}(y; w_-, w_+)$ is a generalized thresholding estimator, can be proved with the following proposition, which follows directly from Proposition 4.3 and Proposition 4.5.

Proposition 4.6. *Assume $(w_-, w_+) \in \mathcal{S}(a)$. With the same $\varphi(\cdot)$ and $\gamma(\cdot)$ as in Lemma 4.1, there exist $\tau_+(w_-, w_+) \geq 0$ and $\tau_-(w_-, w_+) \leq 0$ such that $\hat{\mu}(y; w_-, w_+) = 0$ if and only if $\tau_-(w_-, w_+) \leq y \leq \tau_+(w_-, w_+)$. Furthermore, if $w_+ = w_-$, then $\tau_+(w_-, w_+) = -\tau_-(w_-, w_+)$ and hence $\hat{\mu}(-y; w_-, w_+) = -\hat{\mu}(y; w_-, w_+)$.*

4.2 Proof of Theorem 2.2

To prove the bounded shrinkage property of the estimator $\hat{\mu}(y; w_-, w_+)$, we need the following lemmas.

Lemma 4.7. *Under the same conditions as in Theorem 2.2,*

(i) There exists $B \geq 1$ such that for all $y \in \mathbb{R}$ and $u > 0$, $\gamma_+(y - u) \leq Be^{\Lambda u}\gamma_+(y)$;

(ii) $g_+(y)/\varphi(y)$ is increasing from $g_+(0)/\varphi(0) < 1$ to $+\infty$ as $y \to \infty$;

(iii) $\limsup_{y\to\infty} \gamma_+(y)/g_+(y) < \infty$;

(iv) $\limsup_{y\to\infty} \left|\frac{d}{dy}\log g_+(y)\right| \leq \Lambda < \rho$.

Proof. (i) Since $\gamma_+(u)$ is a decreasing function on $[0, M]$, there exists $B \geq 1$ such that

$$e^{\Lambda x}\gamma_+(x) \leq Be^{\Lambda y}\gamma_+(y), \text{ for all } x, y \in [0, M].$$

Since $\frac{d}{du}\log\gamma_+(u) \geq -\Lambda$ when $u > M$, $e^{\Lambda u}\gamma_+(u)$ is an increasing function of u for $u > M$. Therefore, for all $y \in \mathbb{R}$ and $u > 0$,

$$\gamma_+(y-u) \leq Be^{\Lambda u}\gamma_+(y),$$

because $\gamma_+(y-u) = 0$ if $y < u$.

(ii) With Lemma 4.1, it suffices to prove that $\lim_{y\to\infty} g_+(y)/\varphi(y) = \infty$ and $g_+(0)/\varphi(0) < 1$.

Indeed, $g_+(0)/\varphi(0) < 1$ follows from

$$g_+(0) = \int_0^\infty \varphi(-u)\gamma_+(u)du < \int_0^\infty \varphi(0)\gamma_+(u)du = \varphi(0),$$

where the inequality holds because $\varphi(\cdot)$ is unimodal and symmetric.

Since $\frac{d}{du}\log\gamma_+(u) \geq -\Lambda$ when $u > M$, we have

$$\gamma_+(y) \geq \gamma_+(M)\exp\{\Lambda M\}\exp\{-\Lambda y\}, \forall y > M.$$

Then, for $y > \max(M, 1)$, we have

$$\begin{aligned}
g_+(y) &= \int_0^\infty \varphi(y-u)\gamma_+(u)du = \int_{-\infty}^y \varphi(v)\gamma_+(y-v)dv \\
&\geq \int_0^1 \varphi(v)\gamma_+(y-v)dv \\
&\geq \gamma_+(y)\int_0^1 \varphi(v)dv \\
&\geq \gamma_+(M)\exp\{\Lambda M\}\left(\int_0^1 \varphi(v)dv\right)\exp\{-\Lambda y\} \\
&\triangleq C_1\exp\{-\Lambda y\},
\end{aligned}$$

together with the first condition in Theorem 2.2, there exists $C_2 > 0$, such that, for sufficiently large y, $\varphi(y) \leq C_2\exp\{-\rho y\}$, thus we have

$$\frac{g_+(y)}{\varphi(y)} \geq \frac{C_1}{C_2}\exp\{(\rho-\Lambda)y\} \overset{y\to\infty}{\longrightarrow} \infty \text{ since } \Lambda < \rho.$$

(iii) From (i), we have, for $u > 0$,

$$\begin{aligned}
&\gamma_+(y) \leq Be^{\Lambda u}\gamma_+(y+u) \\
\Rightarrow\ &\gamma_+(y+u) \geq B^{-1}e^{-\Lambda u}\gamma_+(y).
\end{aligned}$$

Then, for $y > 0$,

$$\begin{aligned} g_+(y) &= \int_0^\infty \varphi(y-u)\gamma_+(u)du \\ &= \int_{-\infty}^{y} \varphi(v)\gamma_+(y-v)dv \\ &\geq \int_0^\infty \varphi(u)\gamma_+(y+u)du \\ &\geq B^{-1}\gamma_+(y)\int_0^\infty \varphi(u)e^{-\Lambda u}du, \end{aligned}$$

which implies that, for $y > 0$

$$\begin{aligned} & \frac{\gamma_+(y)}{g_+(y)} \leq \frac{B}{\int_0^\infty \varphi(u)e^{-\Lambda u}du} \\ \Rightarrow\ & \limsup_{y\to\infty} \frac{\gamma_+(y)}{g_+(y)} < \infty. \end{aligned}$$

(iv) Let $\Lambda_\infty = \sup_{u>0} |\frac{d}{du}\log\gamma_+(u)|$. Together with the second condition in Theorem 2.2, we have,

$$\begin{cases} |\frac{d}{du}\gamma_+(u)| \geq \Lambda\gamma_+(u), & u > M; \\ |\frac{d}{du}\gamma_+(u)| \geq \Lambda_\infty\gamma_+(u), & u \leq M. \end{cases}$$

Therefore,

$$\begin{aligned} \left|\frac{d}{dy}\log g_+(y)\right| &= \left|\frac{d}{dy}g_+(y)\right| \Big/ g_+(y) \\ &= \left|\frac{d}{dy}\int_0^\infty \varphi(y-u)\gamma_+(u)du\right| \Big/ g_+(y) \\ &= \left|\varphi(y)\gamma_+(0) - \int_0^\infty \varphi(y-u)\frac{d}{du}\gamma_+(u)du\right| \Big/ g_+(y) \\ &\leq \gamma_+(0)\frac{\varphi(y)}{g_+(y)} + \Lambda_\infty \int_0^M \varphi(y-u)\gamma_+(u)du \Big/ g_+(y) \\ &\quad +\Lambda\int_M^\infty \varphi(y-u)\gamma_+(u)du \Big/ g_+(y) \\ &\leq \gamma_+(0)\frac{\varphi(y)}{g_+(y)} + \Lambda_\infty h(y) + \Lambda, \end{aligned}$$

where

$$\begin{aligned}
h(y) &= \int_0^M \varphi(y-u)\gamma_+(u)du \Big/ g_+(y) \\
&= \int_{y-M}^{y} \varphi(v)\gamma_+(y-v)dv \Big/ g_+(y) \\
&\leq B\int_{y-M}^{y} \varphi(v)e^{\Lambda v}dv \frac{\gamma_+(y)}{g_+(y)} \\
&\rightarrow 0, \text{ as } y \rightarrow \infty.
\end{aligned}$$

The result follows from (iii) and the fact that $\int_{-\infty}^{\infty} \varphi(v)\exp\{\Lambda v\}dv < \infty$.

□

Combining Lemma 4.7 and Lemma 4.2, we will be able to define positive $\tau_+(w_-, w_+)$ and negative $\tau_-(w_-, w_+)$ such that $\tilde{w}_+(\tau_+(w_-, w_+); w_-, w_+) = 0.5$ and $\tilde{w}_-(\tau_-(w_-, w_+); w_-, w_+) = 0.5$, respectively. Apparently, $\hat{\mu}(y; w_-, w_+) = 0$ if and only if $y \in [\tau_-(w_-, w_+), \tau_+(w_-, w_+)]$. For $y > 0$, $\tilde{w}_+(\tau_+(w_-, w_+); w_-, w_+)$ is an increasing function in both w_- and w_+, then $\tau_+(w_-, w_+)$ is a decreasing function in both w_- and w_+. Similarly, $\tau_-(w_-, w_+)$ is an increasing function in both w_- and w_+. Therefore, we have $0 \leq \tau_+(w_-, w_+) \leq \tau_+(0, w_+)$ and $\tau_-(w_-, 0) \leq \tau_-(w_-, w_+) \leq 0$.

Proof of Theorem 2.2. For any constant b_1, we have,

$$\begin{aligned}
&P(\mu > y - b_1 | Y = y) \\
\geq\ &P(\mu > y - b_1 | Y = y, \mu > 0)\tilde{w}_+(y; w_-, w_+).
\end{aligned}$$

Let us first look at the term $P(\mu > y - b_1|Y = y, \mu > 0)$, $y > 0$. From Lemma 4.7, we have, for all $b_2 > M$,

$$\begin{aligned}
&\int_0^{b_2} \gamma_+(u)\varphi(y-u)du \\
\leq\ &\int_0^{b_2} B\exp\{\Lambda(b_2 - u)\}\gamma_+(b_2)\varphi(u-y)du \\
=\ &Be^{\Lambda b_2}\gamma_+(b_2)\int_0^{b_2} e^{-\Lambda u}\varphi(u-y)du,
\end{aligned}$$

and

$$\begin{aligned}
&\int_{b_2}^{\infty} \gamma_+(u)\varphi(y-u)du \\
\geq\ & \int_{b_2}^{\infty} \exp\{\Lambda(b_2-u)\}\gamma_+(b_2)\varphi(u-y)du \\
=\ & e^{\Lambda b_2}\gamma_+(b_2)\int_{b_2}^{\infty} e^{-\Lambda u}\varphi(u-y)du.
\end{aligned}$$

Therefore, for $b_2 > M$,

$$\begin{aligned}
& Odds(\mu > b_2 | Y = y, \mu > 0) \\
=\ & \frac{\int_{b_2}^{\infty} \gamma_+(u)\varphi(y-u)du}{\int_0^{b_2} \gamma_+(u)\varphi(y-u)du} \\
\geq\ & \frac{\int_{b_2}^{\infty} e^{-\Lambda u}\varphi(u-y)du}{B\int_0^{b_2} e^{-\Lambda u}\varphi(u-y)du}.
\end{aligned}$$

Since $\Lambda < \rho$, we have $\int_{-\infty}^{\infty} e^{-\Lambda u}\varphi(u)du < \infty$. So, there exists a value b_1 such that

$$\int_{-b_1}^{\infty} e^{-\Lambda u}\varphi(u)du > 3B\int_{-\infty}^{-b_1} e^{-\Lambda u}\varphi(u)du.$$

As long as $y > b_1 + M$, we will then have

$$\begin{aligned}
& Odds(\mu > y - b_1 | Y = y, \mu > 0) \\
\geq\ & \frac{\int_{y-b_1}^{\infty} e^{-\Lambda u}\varphi(u-y)du}{B\int_0^{y-b_1} e^{-\Lambda u}\varphi(u-y)du} \\
=\ & \frac{\int_{-b_1}^{\infty} e^{-\Lambda v}\varphi(v)dv}{B\int_{-y}^{-b_1} e^{-\Lambda v}\varphi(v)dv} \\
\geq\ & \frac{\int_{-b_1}^{\infty} e^{-\Lambda v}\varphi(v)dv}{B\int_{-\infty}^{-b_1} e^{-\Lambda v}\varphi(v)dv} \\
>\ & 3,
\end{aligned}$$

so that

$$P(\mu > y - b_1 | Y = y, \mu > 0) > 3/4, \quad y > b_1 + M.$$

Then, let us consider the term $\tilde{w}_+(y; w_-, w_+)$. When $y > 0$, we have

$$\tilde{w}_+(y; w_-, w_+) \geq \tilde{w}_+(y; 0, w_+).$$

With both conditions given in the theorem, and the result in Lemma 4.7, we can choose $\tau_0 \geq M$ large enough such that, for $u \geq \tau_0$, we have

$$\frac{d}{du} \log g_+(u) \leq -\Lambda - \frac{\rho - \Lambda}{2}, \qquad \frac{d}{du} \log \varphi(u) \leq -\rho.$$

Then choose w_0 such that $\tau_+(w_-, w_0) = \tau_0$, and let $b_3 = 2\log 2/(\rho - \Lambda)$.

Suppose $w_+ \leq w_0$, so that $\tau_+(w_-, w_+) \geq \tau_0$. Let $\Omega_+(y; w_-, w_+) = Odds(\mu > 0|Y = y; w_-, w_+)$, then for $y > \tau_+(w_-, w_+) + b_3$,

$$\begin{aligned}
&\Omega_+(y; w_-, w_+) \\
&= \Omega_+(\tau_+(w_-, w_+); w_-, w_+) \exp\left\{\int_{\tau_+(w_-,w_+)}^{y} \frac{d}{du}[\log g_+(u) - \log \varphi(u)]du\right\} \\
&\quad \times \frac{(1 - w_- - w_+) + w_- g_-(\tau_+(w_-, w_+))/\varphi(\tau_+(w_-, w_+))}{(1 - w_- - w_+) + w_- g_-(y)/\varphi(y)} \\
&\geq \exp\left\{\int_{\tau_+(w_-,w_+)}^{y} \frac{d}{du}[\log g_+(u) - \log \varphi(u)]du\right\} \\
&\geq \exp\left\{\int_{\tau_+(w_-,w_+)}^{y} [-\Lambda - (\rho - \Lambda)/2 + \rho]du\right\} \\
&= \exp\{(\rho - \Lambda)(y - \tau_+(w_-, w_+))/2\} \geq 2,
\end{aligned}$$

where the first inequality follows from Lemma 4.1. On the other hand, if $w_+ > w_0$ we will have $\Omega_+(y; w_-, w_+) > \Omega_+(y; w_-, w_0) \geq 2$ as long as $y > \tau_0 + b_3$. In either case, we have

$$\tilde{w}_+(y; w_-, w_+) \geq 2/3.$$

In summary, we will have $P(\mu > y - b_1|Y = y) > 1/2$ whenever $y \geq \max\{b_1 + M, \tau_+(w_-, w_+) + b_3, \tau_0 + b_3\}$; otherwise, we have $y - \hat{\mu}(y; w_-, w_+) \leq y$. Hence, for all $y > 0$,

$$y - \hat{\mu}(y; w_-, w_+) \leq \max\{b_1, b_1 + M, \tau_+(w_-, w_+) + b_3, \tau_0 + b_3\} \leq \tau_+(w_-, w_+) + c_+,$$

where $c_+ = \tau_0 + b_1 \vee b_3$. Similarly, for all $y < 0$, we can get

$$\hat{\mu}(y; w_-, w_+) - y \leq -\tau_-(w_-, w_+) + c_-.$$

□

5 Discussion

We have proposed a class of Bayes procedures which lead to generalized shrinkage and thresholding estimators, useful for estimating high-dimensional parameters. These estimators generalize existing approaches, such as SURE method, FDR-controlling approach and the EB estimators, in the sense of being adaptive to not only sparse but also asymmetric parameter spaces. We have also developed generalized shrinkage and thresholding estimators by allowing negative and positive parameters to have different mixing weights w_- and w_+. When restricting $w_- = w_+$, the generalized shrinkage estimators reduce to the shrinkage estimators by Johnstone and Silverman (2004). As shown here, both classes of estimators require the error term following a PF_2 distribution, rather than the PF_3 distribution by Johnstone and Silverman (2004). It is much easier to check whether a function is PF_2 as a log-concave function suffices.

We also constructed a Bayesian estimator using the posterior median, which is essentially a generalized thresholding estimator when $(w_-, w_+) \in \mathcal{S}$. Another Bayesian estimator can be established with the posterior mean $E[\mu|y; w_-, w_+]$, which, unlike the posterior median, may only be a generalized shrinkage estimator but not a generalized thresholding estimator as it is always continuous in y. We conjecture that, under similar conditions, this estimator also has the bounded shrinkage property. Let $(\hat{w}_-, \hat{w}_+)$ maximize the marginal maximum likelihood under the restriction $(\hat{w}_-, \hat{w}_+) \in \mathcal{S}(a)$. We then can construct the generalized empirical Bayes estimators based on either the posterior median or the posterior mean. Utilizing the preliminary results in this article, we may be able to prove the intermediate lemmas in Johnstone and Silverman (2004), which essentially lead to the uniformly bounded risk property and the sparsity adaptivity properties of these two generalized empirical Bayes estimators.

On estimating multivariate standard normal means, a computationally tractable implementation of the generalized empirical Bayes estimators follows using the quasi-Cauchy prior (3.1). As shown in the simulation study, the GEB estimator has an excellent performance when $p_- \neq p_+$. However, when $|\mu_-| \neq \mu_+$, the GEB estimator may not gain its advantage over the EB estimator. Indeed, although the parameter space is asymmetric, the GEB estimator may perform worse than the EB estimator when $|\mu_-| \neq \mu_+$ but $p_- = p_+$, see Figure 5 for example. This is due to the fact that the prior (2.2) lacks sufficient ability to model the asymmetric parameter space with $|\mu_-| \neq \mu_+$ but $p_- = p_+$.

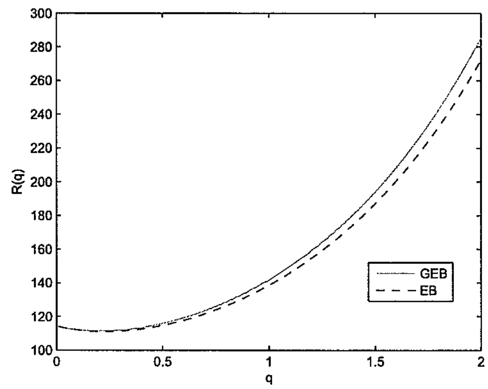

Figure 5: Risks of the EB and GEB Estimators When $(\mu_-, \mu_+) = (-6, 3)$ and $p_- = p_+ = 50$.

A possible extension is to consider the Laplace prior, i.e.,

$$\begin{cases} \gamma_+(\mu|\alpha_+) = \frac{\alpha_+}{2} \exp(-\mu\alpha_+) 1_{[0,\infty)}(\mu) \\ \gamma_-(\mu|\alpha_-) = \frac{\alpha_-}{2} \exp(\mu\alpha_-) 1_{(-\infty,0]}(\mu) \end{cases} \tag{5.1}$$

which is also investigated by Johnstone and Silverman (2004) to develop the empirical Bayes shrinkage estimator with $\alpha_+ = \alpha_-$. Employing α_+ and α_- makes it capable to model the asymmetric parameter space with $|\mu_-| \neq \mu_+$. Empirical Bayes estimators can be established by maximizing the marginal likelihood function for the hyperparameter α_+ and α_-, together with the weights w_+ and w_-. However, for any Bayesian estimator based on the prior (5.1) to be either generalized shrinkage estimator or generalized thresholding estimator, we certainly need restrictions on both (α_-, α_+) and (w_-, w_+), which is of our future research interest.

Let $\varphi(\cdot)$ be a symmetric PF_2 density function on $\mathbb{R}$, and $\mathbf{Y} = (y_1, y_2, \cdots, y_p)$ be the observed data with $y_i - \mu_i \overset{iid}{\sim} \varphi(\cdot/\sigma)$. When σ is unknown and the parameter space is sparse, we can robustly estimate σ as a trivial solution. However, it may be challenging to construct an efficient estimate of σ, which help establish the properties of the resultant estimators. Alternatively, we can estimate σ by maximizing the marginal likelihood function together with other hyperparameters, which make it even more challenging to investigate all the properties of the resultant estimators.

References

[1] Abramovich, F., Benjamini, Y., Donoho, D. L. and Johnstone, I. M. (2000). Adapting to unknown sparsity by controlling the false discovery rate. Technical Report 2000-19. Department of Statistics, Stanford University.

[2] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society Series B* 57, 289-300.

[3] Cai, T. T. and Silverman, B. W. (2001). Incorporating information on neighboring coefficients into wavelet estimation. *Sankhya* Series B 63, 127-148.

[4] Cai, T. T. (2002). On block thresholding in wavelet regression: Adaptivity, block size, and threshold level. *Statistica Sinica* 12, 1241-1273.

[5] Copas, J.B. (1969). Compound decisions and empirical Bayes. *Journal of the Royal Statistics Society Series B* 31, 397-425.

[6] Donoho, D. L. and Johnstone, I. M. (1994a). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* 81, 425-455.

[7] Donoho, D. L. and Johnstone, I. M. (1994b). Minimax risk over ℓ_p-balls for ℓ_q-error. *Probability Theory and Related Fields* 99, 1200-1224.

[8] Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association* 90, 1200-1224.

[9] Donoho, D. L., Johnstone, I. M., Hoch, C. and Stern, A. (1992). Maximum entropy and the nearly black object (with discussion). *Journal of Royal Statistical Society Series B* 54, 41-81.

[10] Efron, B. and Morris, C. (1972a). Empirical Bayes on vector observations: an extension of Stein's method. *Biometrika* 59, 335-347.

[11] Efron, B. and Morris, C. (1972b). Limiting the risk of Bayes and empirical Bayes estimators. II. The empirical Bayes case. *Journal of the American Statistical Association* 67, 130-139.

[12] Efron, B. and Morris, C. (1973a). Combining possibly related estimation problems (with discussion). *Journal of the Royal Statistical Society Series B* 35, 379-421.

[13] Efron, B. and Morris, C. (1973b). Stein's estimation rule and its competitors – an empirical Bayes approach. *Journal of the American Statistical Association* 68, 117-130.

[14] Fan, J. and Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association* 96, 1348-1360.

[15] Finehout, E.J., Franck, Z. and Lee, K.H. (2005). Complement protein isoforms in CSF as possible biomarkers for neurodegenerative disease. *Disease Markers* 21, 93-101.

[16] Fourdrinier, D., Strawderman W. E. and Wells, M. T. (1998). On the construction of Bayes minimax estimators. *The Annals of Statistics* 26, 660-671.

[17] James, W. and Stein, C. (1961). Estimation with quadratic loss. *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability* 1, 361-379. University of California Press, Berkeley.

[18] Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: empirical Bayes estimates of possibly sparse sequence. *The Annals of Statistics* 32, 1594-1649.

[19] Johnstone, I.M. and Silverman, B.W. (2005). EbayesThresh: R programs for empirical Bayes thresholding. *Journal of Statistical Software* 12, 1-38.

[20] Keller, A., Purvine, S., Nesvizhskii, A.I., Stolyar, S., Goodlett, D.R. and Koler, E. (2002). Experimental protein mixture for validating mass spectral analysis. *Omics* 6, 207-212.

[21] Li, Z., Van Calcar, S., Qu, C., Cavenee, W.K., Zhang, M.Q., Ren, B. (2003). A global transcriptional regulatory role for c-Myc in Burkitt's lymphoma cells. *Proceedings of the National Academy of Sciences of the United States of America* 100, 8164-8169.

[22] Robbins, H. (1951). Asymptotically subminimax solutions of compound decision problems. *Proceedings of the Second Berkeley Symposium on Mathematical Statistics and Probability* 1, 131-148. University of California Press, Berkeley.

[23] Samuel, E. (1965). On simple rules for the compound decision problem. *Journal of the Royal Statistical Society Series B* 27, 238-244.

[24] Stein, C. (1956). Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability* 1, 197-206. University of California Press, Berkeley.

[25] Stein, C. (1981). Estimation of the mean of a multivariate normal distribution. *The Annals of Statistics* 9, 1135-1151.

[26] Strawderman, W.E. (1971). Proper Bayes minimax estimators of the multivariate normal mean. *Annals of Mathematical Statistics* 22, 118-171.

[27] van de Peppel, J., Kemmeren, P., van Bakel, H., Radonjic, M., van Leenen, D. and Holstege, F.C. (2003). Monitoring global messenger RNA changes in externally controlled microarray experiments. *EMBO Reports* 4, 387-393.

[28] Xu, S. (2003). Estimating polygenic effects using markers of the entire genome. *Genetics* 163, 789-801.

[29] Zhang, D., Zhang, M. and Wells, M.T. (2006). Multiplicative background correction for spotted microarrays to improve reproducibility. *Genetical Research* 97, 195-206.

[30] Zhang, M., Montooth, K. L., Wells, M. T., Clark, A. G. and Zhang, D. (2005). Mapping multiple quantitative trait loci by Bayesian classification. *Genetics* 169, 2305-2318.