Empirical Bayesian thresholding for sparse signals using mixture loss functions

Vikas C. Raykar
CAD and Knowledge Solutions
Siemens Healthcare, Malvern, PA, 19355, USA
vikas.raykar@siemens.com
Linda H. Zhao
Department of Statistics
University of Pennsylvania, Philadelphia, PA, 19104, USA
lzhao@wharton.upenn.edu

Abstract

We develop an empirical Bayesian thresholding rule for the normal mean problem that adapts well to the sparsity of the signal. An key element is the use of a mixture loss function that combines both the L_p loss and the $0-1$ loss function. The Bayes procedures under this loss are explicitly given as thresholding rules and are easy to compute. The prior on each mean is a mixture of an atom of probability at zero, and a Laplace or normal density for the nonzero part. The mixing probability as well as the spread of the non-zero part are hyperparameters that are estimated by the empirical Bayes procedure. Our simulation experiments demonstrate that the proposed method performs better than the other competing methods for a wide range of scenarios. We also apply our proposed method for feature selection to four data sets.

Keywords:

Empirical Bayes; Mixture loss; Mixture prior; Normal means; Sparsity; Thresholding

1 Introduction

We are given n scalar observations $x_1, x_2, \ldots, x_n$ satisfying

$$x_i = \mu_i + \epsilon_i, \tag{1}$$

where each ϵ_i is independent and identically distributed as $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, a normal distribution with mean zero and a known variance σ^2. Based on the observation $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ we need a desirable estimate $\widehat{\boldsymbol{\mu}}$ of the unknown parameter $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_n)$. This is generally referred to as the multivariate normal mean problem.

Very often we encounter scenarios that involve *sparsity*; a large number of μ_i's are zero but we do not know how many of them are zero. With no information on how sparse the vector $\boldsymbol{\mu}$ is, an estimator $\widehat{\boldsymbol{\mu}}$ that adapts to the degree of sparsity is desirable.

The normal mean problem occurs in a wide range of practical applications. Some examples include model selection in machine learning/data mining (George and Foster (2000)), smoothing in signal processing, de-noising in astronomical image processing (Johnstone and Silverman (2004)), wavelet approaches to non-parametric regression (Johnstone and Silverman (2005)), and significance testing in genomics and bio-informatics (Efron and Tibshirani (2007)). Situations involving Poisson or binomial observations, such as baseball batting averages, can be transformed and efficiently treated within the normal means context (Brown (2008)).

For sparse situations the desired and the natural estimator is an explicit thresholding rule of the form

$$\hat{\mu}_i = \begin{cases} 0 & \text{if } |x_i| < t(\boldsymbol{x}) \\ \text{some estimate} & \text{otherwise,} \end{cases} \tag{2}$$

where t is some threshold that can depend on $\boldsymbol{x}$. As a result of this the estimate has some values exactly zero. However it is crucial that the threshold t adapt to the degree of the sparsity in the signal, which is unknown. We propose an explicit thresholding rule that adapts to the sparsity of the signal automatically using an empirical Bayesian approach. The proposed approach has the following three components:

1. To incorporate the possibility of sparsity we use a mixture prior on each mean with an atom of probability at zero and either a Laplace or normal density for the nonzero part. This form of the mixture prior has been earlier used by Johnstone and Silverman (2004, 2005), Abramovich, Sapatinas, and Silverman (1998), Clyde, Parmigiani, and Vidakovic (1998), and Chipman, Kolaczyk, and McCulloch (1997) in the context of thresholding the wavlet coefficients.

2. The mixing probability as well as the spread of the non-zero part are hyperparameters which are estimated by an empirical Bayes procedure.

3. The novel key element is the use of a mixture loss function combining L_p-loss ($p = 1, 2$) and a $0 - 1$ loss function–more precisely we take

$$L_{p,K}(\mu_i, \widehat{\mu}_i) = K\mathbf{1}_{\{\hat{\mu}_i \neq \mu_i\}} + |\widehat{\mu}_i - \mu_i|^p. \tag{3}$$

   Here the constant K controls the amount of penalty for incorrectly estimating the exact true value of μ_i, and $K = 0$ corresponds to the usual L_p loss.

A nice property of our method is that the resulting Bayes procedures are explicitly given as thresholding rules, *i.e*, a particular parameter estimate $\widehat{\mu}_i$ is set to zero if it is less than some threshold. The resulting estimator is adaptive to the amount of sparsity, and is computed automatically based on the entire observation $\boldsymbol{x}$. To be more precise, in our procedure the hyperparameters are adaptively estimated based on the entire observation $\boldsymbol{x}$.

Without the mixture loss, the Bayes procedure for the L_2 loss is the posterior mean, which has a shrinkage property but no thresholding property at all, *i.e.*, all estimates are non-zero. The posterior mean and median are just special cases of the proposed estimator (with $K = 0$). In our simulation results (see Table 1), the proposed estimators are better than the posterior mean and median in terms of the total squared error, in addition to the fact that they adapt to the sparsity in the signal. We also study the effect of different choices of K, and empirically propose a universal value that depends only on the adaptive estimate of the hyperparameters. Because of the mixture loss, the proposed procedure turns out to be robust to mis-specification of the non-zero component of the mixture prior.

Johnstone and Silverman (2004, 2005) propose an estimator that is closely related to ours and study its theoretical properties. Their estimator is the posterior median based on the same prior as we use. This happens to be a special case of our proposed estimator–it corresponds to our loss function (3) with $p = 1$ and $K = 0$. Our results show that by using a non-zero K the mean squared error can be much lower and the sparsity is captured much more accurately. Johnstone and Silverman (2004) use the posterior median (L_1-loss) for its thresholding property. By using a non-zero K our proposed estimator is always a thresholding rule for both L_1 and L_2 loss.

There are related approaches. For example, the SURE approach of Donoho and Johnstone (1995) is based on minimizing Stein's unbiased risk estimate for the mean squared error of soft thresholding. The FDR approach of Abramovich, Benjamini, Donoho, and Johnstone (2006) is derived from the principle of controlling the false discovery rate in simultaneous hypothesis testing. Brown and Greenstein (2009) also propose a non-parametric empirical Bayes estimator; their estimator is not a thresholding estimator, but does adapt and perform well in moderately sparse or non-sparse settings.

The rest of the paper is organized as follows. In Section 2 we describe the mixture prior used to promote sparsity. We subsequently describe an empirical Bayes procedure to estimate the hyperparameters by maximizing the marginal likelihood; the estimated hyperparameters are then plugged in to derive the posterior. The mixture loss function is introduced in Section 3, and the corresponding Bayes rule is derived. Simulation results, along with the choice of K, are discussed in Section 4. In Section 5 we use the proposed procedure to select relevant features for classification on four data sets. The optimal number of features selected through our methods agrees with those selected using cross-validation.

2 Adapting to unknown sparsity

Without loss of generality we assume that the x_i are scaled such that $\sigma^2 = 1$. If σ is unknown we estimate it using a robust estimator. One good choice is the median absolute value of x_i. Since we assume that μ is sparse, the median absolute value is not strongly affected by the nonzero μ_i.

From (1), and assuming $\epsilon_i \sim \mathcal{N}(0, 1)$, we have $p(x_i|\mu_i) = \mathcal{N}(x_i|\mu_i, 1)$. Since the ϵ_i are independent, the likelihood of the parameters $\boldsymbol{\mu}$ given the observations $\boldsymbol{x}$ can be factored as

$$p(\boldsymbol{x}|\boldsymbol{\mu}) = \prod_{i=1}^{n} p(x_i|\mu_i) = \prod_{i=1}^{n} \mathcal{N}(x_i|\mu_i, 1). \tag{4}$$

Note that the maximum-likelihood estimator $\widehat{\boldsymbol{\mu}}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\mu}} p(\boldsymbol{x}|\boldsymbol{\mu})$ is the observation $\boldsymbol{x}$ itself. It is well known that this estimator can be considerably improved by such shrinkage estimators as the James-Stein estimators (see Berger (1985)).

We impose a prior on $\boldsymbol{\mu}$ and then find the Bayes solution under a suitable loss function. In order to promote sparsity, we assume that each of the parameters μ_i comes from a mixture of a delta function mass at zero and a fixed symmetric density,

$$p(\mu_i|w, a) = w\delta(\mu_i) + (1 - w)\gamma_a(\mu_i), \tag{5}$$

where $w \in [0, 1]$ is the mixture parameter and the δ puts probability mass of 1 at 0, and zero elsewhere. For the nonzero part of the prior γ_a we consider two possibilities.

1. A zero mean normal with variance a^2.

$$\gamma_a(\mu_i) = \mathcal{N}(\mu_i|0, a^2) = (2\pi a^2)^{-1/2} \exp(-\mu_i^2/2a^2). \tag{6}$$

2. A double exponential (Laplace) with scale parameter a.

$$\gamma_a(\mu_i) = 0.5a \exp(-a|\mu_i|). \tag{7}$$

The Laplace prior has a heavier tail than the normal. We consider w and a as hyper-parameters and use an empirical Bayesian approach to estimate them by maximizing the marginal likelihood. Given the hyper-parameters w and a, the posterior of $\boldsymbol{\mu}$ given the data $\boldsymbol{x}$ can be written as

$$p(\boldsymbol{\mu}|\boldsymbol{x}, w, a) = \frac{\prod_{i=1}^n p(x_i|\mu_i)p(\mu_i|w, a)}{m(\boldsymbol{x}|w, a)}, \quad \text{where} \tag{8}$$

$$m(\boldsymbol{x}|w, a) = \prod_{i=1}^{n} \int p(x_i|\mu_i)p(\mu_i|w, a)d\mu_i \tag{9}$$

is the marginal of the data given the hyper-parameters. For the likelihood (4) and the mixture prior (5) we have

$$\int p(x_i|\mu_i)p(\mu_i|w, a)d\mu_i = w\mathcal{N}(x_i|0, 1) + (1 - w)g_a(x_i), \tag{10}$$

where we define $g_a(x_i) = \int \mathcal{N}(\mu_i|x_i, 1)\gamma_a(\mu_i)d\mu_i$. Hence the log-marginal likelihood can be written as

$$\log m(\boldsymbol{x}|w, a) = \sum_{i=1}^{n} \log\left[w\mathcal{N}(x_i|0, 1) + (1 - w)g_a(x_i)\right]. \tag{11}$$

We chose w and a to maximize the log-marginal likelihood numerically,

$$\{\widehat{w}, \widehat{a}\} = \arg\max_{w,a} \log m(\boldsymbol{x}|w, a). \tag{12}$$

More specifically we used an alternate optimization technique, for a fixed w, find the a which maximizes the log-marginal likelihood; given the best a, find the best w; repeat to convergence.

Estimated hyperparameters are then plugged into the posterior. For ease of later derivation we factor the posterior as

$$p(\boldsymbol{\mu}|\boldsymbol{x}, w, a) = \prod_{i=1}^{n} p(\mu_i|x_i, w, a), \quad \text{where} \tag{13}$$

$$p(\mu_i|x_i, w, a) \quad = \quad \frac{w\delta(\mu_i)\mathcal{N}(x_i|\mu_i, 1) + (1 - w)\gamma_a(\mu_i)\mathcal{N}(x_i|\mu_i, 1)}{w\mathcal{N}(x_i|0, 1) + (1 - w)g_a(x_i)}. \tag{14}$$

Define

$$\tilde{p}_i = \frac{w\mathcal{N}(x_i|0, 1)}{w\mathcal{N}(x_i|0, 1) + (1 - w)g_a(x_i)}. \tag{15}$$

Then

$$p(\mu_i|x_i, w, a) \quad = \quad \tilde{p}_i\delta(\mu_i) + (1-\tilde{p}_i)G(\mu_i), \quad \text{where} \tag{16}$$

$$G(\mu_i) = \frac{\mathcal{N}(\mu_i|x_i,1)\gamma_a(\mu_i)}{\int \mathcal{N}(\mu_i|x_i,1)\gamma_a(\mu_i)d\mu_i}. \tag{17}$$

3 Bayes thresholding rule via mixture loss function

From now on we drop the subscript i in (16) and write the posterior as

$$p(\mu|x, w, a) \quad = \quad \tilde{p}\delta(\mu) + (1-\tilde{p})G(\mu). \tag{18}$$

We can either use the mean or the median of the posterior as our estimate. These correspond to L_2 and the L_1 loss, respectively. It is known that the mean does not have the thresholding property while the median does.

3.1 Mixture loss function

We propose the following loss function which combines the $0-1$-loss and the L_p-loss:

$$L(\mu, \widehat{\mu}) = K\mathbf{1}_{\{\hat{\mu}\neq\mu\}} + |\widehat{\mu} - \mu|^p, \tag{19}$$

where, $\mathbf{1}_{\{\widehat{\mu}\neq\mu\}} = 1$ if $\hat{\mu} \neq \mu$, and 0 otherwise. K controls the amount of penalty for wrongly estimating the exact true value of μ. In our set up, we believe that a significant proportion of the μ_i are zero, so we want the resulting estimate to have a significant chance to be *exactly zero*. The resulting estimate is a thresholding rule.

We now derive the Bayes rule for this loss function. It minimizes the expected posterior loss

$$\hat{\mu}(x, w, a) = \arg min_{\hat{\mu}} \int L(\mu, \widehat{\mu})p(\mu|x, w, a)d\mu. \tag{20}$$

Although p in the loss function is simply non-negative number, we present the results only for $p = 2$ and $p = 1$.

3.2 Bayes rule when $p = 2$

Theorem 3.1 *Under the loss (19) when $p = 2$, the Bayes' rule $\hat{\mu}$ is the thresholding rule*

$$\hat{\mu} = \begin{cases} 0 & if \;\; (1-\tilde{p})^2E_G^2[\mu|x, w, a] < K\tilde{p} \\ (1-\tilde{p})E_G[\mu|x, w, a] & otherwise, \end{cases}$$

where $\tilde{p}$ and G are given in (15) and (17), respectively. The region where $\hat{\mu} = 0$ is an interval in x.

Proof: The posterior is given by

$$p(\mu|x,w,a) \;=\; \tilde{p}\delta(\mu)+(1-\tilde{p})G(\mu). \tag{21}$$

Note that

$$p(\mu=0|x,w,a)=\tilde{p} \quad \text{and} \quad p(\mu|x,w,a,\mu\neq 0)=G(\mu). \tag{22}$$

We separately consider the cases $\hat{\mu}=0$ and $\hat{\mu}\neq 0$.

1. When $\hat{\mu}=0$ the loss function is

$$L(\mu,\hat{\mu})=\begin{cases} 0 & \mu=0 \\ K+\mu^2 & \mu\neq 0. \end{cases} \tag{23}$$

The expected posterior loss is

$$\begin{aligned} E_\mu[L(\mu,\hat{\mu})|\hat{\mu}=0] &= \int L(\mu,0)p(\mu|x,w,a)d\mu=(1-\tilde{p})\int(K+\mu^2)G(\mu)d\mu \\ &= (1-\tilde{p})(K+E_G[\mu^2]). \end{aligned} \tag{24}$$

2. Similarly when $\hat{\mu}\neq 0$,

$$L(\mu,\hat{\mu})=\begin{cases} K+\hat{\mu}^2 & \mu=0 \\ K+(\hat{\mu}-\mu)^2 & \mu\neq 0. \end{cases} \tag{25}$$

The expected posterior loss is

$$\begin{aligned} E_\mu[L(\mu,\hat{\mu})|\hat{\mu}\neq 0] &= (K+\hat{\mu}^2)\tilde{p}+(1-\tilde{p})\int(K+(\hat{\mu}-\mu)^2)G(\mu)d\mu \\ &= \hat{\mu}^2-2(1-\tilde{p})E_G[\mu]\hat{\mu}+(K+(1-\tilde{p})E_G[\mu^2]). \end{aligned} \tag{26}$$

The minimum value of (26) is attained at

$$\hat{\mu}=(1-\tilde{p})E_G[\mu] \tag{27}$$

and the minimum expected posterior loss is

$$K+(1-\tilde{p})E_G[\mu^2]-(1-\tilde{p})^2E_G^2[\mu]. \tag{28}$$

When (24) < (28) the Bayes rule is $\hat{\mu}=0$. This is equivalent to

$$(1-\tilde{p})(K+E_G[\mu^2])<K+(1-\tilde{p})E_G[\mu^2]-(1-\tilde{p})^2E_G^2[\mu]. \tag{29}$$

The Bayes thresholding rule is

$$\hat{\mu}=0 \quad \text{if} \quad E_G^2[\mu]<\frac{K\tilde{p}}{(1-\tilde{p})^2}, \tag{30}$$

otherwise

$$\hat{\mu} = (1 - \tilde{p})E_G[\mu]. \tag{31}$$

Finally, every Bayes procedure corresponding to a bowl-shaped loss is monotone (Brown and Cohen (1976)), hence the region where $\hat{\mu} = 0$ is an interval in x.

The resulting estimator is an explicit thresholding rule but when $K = 0$ the resulting (shrinkage) estimator is the posterior mean that does not do explicit thresholding.

3.3 Bayes rule when $p = 1$

In this section we consider the loss function

$$L(\mu, \widehat{\mu}) = K\mathbf{1}_{\{\hat{\mu} \neq \mu\}} + |\widehat{\mu} - \mu|. \tag{32}$$

Theorem 3.2 *Under the loss (32) the Bayes' rule $\hat{\mu}$ is a thresholding rule. Let $\gamma_0 = \int_{-\infty}^{0} G(\mu)d\mu$ where G is given in (17), and $\tilde{p}$ be the posterior probability of nonzero mean calculated in (15).*

- *If $\tilde{p} > 1/2$ the Bayes rule is $\hat{\mu} = 0$.*
- *When $\tilde{p} \leq 1/2$*

  1. *if $\frac{1-2\tilde{p}}{2(1-\tilde{p})} < \gamma_0 < \frac{1}{2(1-\tilde{p})}$ then the Bayes rule is also $\hat{\mu} = 0$.*
  2. *if $\gamma_0 > \frac{1}{2(1-\tilde{p})}$, $\hat{\mu}_{min}$ is the unique negative solution to*

  $$\int_{-\infty}^{\hat{\mu}_{min}} G(\mu)d\mu = \frac{1}{2(1-\tilde{p})}. \tag{33}$$

  3. *if $\gamma_0 < \frac{1-2\tilde{p}}{2(1-\tilde{p})}$, $\hat{\mu}_{min}$ is the unique positive solution to*

  $$\int_{-\infty}^{\hat{\mu}_{min}} G(\mu)d\mu = \frac{1-2\tilde{p}}{2(1-\tilde{p})}. \tag{34}$$

  4. *In either (2) or (3) the Bayes rule is $\hat{\mu} = 0$ if $(1 - \tilde{p})(K + E_G[|\mu|]) < \pi_{min}$, where $\pi_{min} = K + |\hat{\mu}_{min}|\tilde{p} + (1 - \tilde{p}) \int |\hat{\mu}_{min} - \mu|G(\mu)d\mu$. Otherwise the Bayes rule is $\hat{\mu} = \hat{\mu}_{min}$ defined in (33) or (34) respectively.*

Proof: As before we separately consider the cases $\hat{\mu} = 0$ and $\hat{\mu} \neq 0$.

1. When $\hat{\mu} = 0$ the loss function is

$$L(\mu, \hat{\mu}) = \begin{cases} 0 & \mu = 0 \\ K + |\mu| & \mu \neq 0 \end{cases} \tag{35}$$

The expected posterior loss is

$$\begin{aligned} E_\mu[L(\mu,\hat\mu)|\hat\mu=0] &= \int L(\mu,0)p(\mu|x,w,a)d\mu = (1-\tilde p)\int (K+|\mu|)G(\mu)d\mu \\ &= (1-\tilde p)(K+E_G[|\mu|]). \end{aligned} \tag{36}$$

2. Similarly when $\hat\mu \neq 0$,

$$L(\mu,\hat\mu) = \begin{cases} K+|\hat\mu| & \mu=0 \\ K+|\hat\mu-\mu| & \mu\neq 0. \end{cases} \tag{37}$$

The expected posterior loss is

$$\begin{aligned} E_\mu[L(\mu,\hat\mu)|\hat\mu\neq 0] &= (K+|\hat\mu|)\tilde p + (1-\tilde p)\int (K+|\hat\mu-\mu|)G(\mu)d\mu \\ &= K+|\hat\mu|\tilde p + (1-\tilde p)\int |\hat\mu-\mu|G(\mu)d\mu. \end{aligned} \tag{38}$$

This is minimized when the first derivative is zero. The derivative of the expected posterior loss can be written as

$$E'_\mu[L(\mu,\hat\mu)|\hat\mu\neq 0] = \mathrm{sign}(\hat\mu)\tilde p + (1-\tilde p)\left[2\int_{-\infty}^{\hat\mu} G(\mu)d\mu - 1\right]. \tag{39}$$

Hence the expected posterior loss attains a minimum at $\hat\mu_{\min}$ where

$$\int_{-\infty}^{\hat\mu_{\min}} G(\mu)d\mu = \frac{1}{2}\left[1-\frac{\tilde p}{1-\tilde p}\mathrm{sign}(\hat\mu_{\min})\right], \tag{40}$$

and the minimum value is

$$\pi_{\min} = K+|\hat\mu_{\min}|\tilde p + (1-\tilde p)\int |\hat\mu_{\min}-\mu|G(\mu)d\mu. \tag{41}$$

Since $0 \leq \int_{-\infty}^{\hat\mu_{\min}} G(\mu)d\mu \leq 1$, (40) has no solution when $\tilde p > 1/2$. Let $\gamma_0 = \int_{-\infty}^{0} G(\mu)d\mu$. It is also easy to see that there is no solution to (40) if $\frac{1-2\tilde p}{2(1-\tilde p)} < \gamma_0 < \frac{1}{2(1-\tilde p)}$. In these situations, the minimum is attained at the boundary $\{-\infty, 0, \infty\}$. The expected posterior loss is minimum when $\hat\mu_{\min}$ approaches zero, the minimum value being $K + (1-\tilde p)E_G[|\mu|]$, which is greater than the posterior loss (36) when $\hat\mu_{\min} = 0$. This means that the Bayes rule is 0.

The solution in (40) exists when $\tilde p < 1/2$ and either $\gamma_0 > \frac{1}{2(1-\tilde p)}$ or $\gamma_0 < \frac{1-2\tilde p}{2(1-\tilde p)}$. By comparing the posterior risks between (36) and (38), we conclude the Bayes rule in the theorem.

4 Simulation Studies

4.1 Experimental setup

To evaluate our proposed procedure and facilitate comparison, we followed the simulation setup specified in Johnstone and Silverman (2004). A sequence $\boldsymbol{\mu}$ of fixed length $N = 1000$ was generated with different degrees of sparsity and signal strengths. The sequence had $\mu_i = 0$ except at R randomly chosen positions, where it took a specified value V–representing the strength of the non-zero component of the signal. The observations x_i were generated by adding $\mathcal{N}(0,1)$ noise for each μ_i. The signal $\boldsymbol{\mu}$ was estimated using the proposed procedure with different mixture loss functions (0/1+L_2 or 0/1+L_1) and different non-zero component of the mixture prior (normal or Laplace). To evaluate accuracy, the total squared error (TSE) between $\boldsymbol{\mu}$ and the estimate $\hat{\boldsymbol{\mu}}$ was computed as $\sum_{i=1}^{N}(\hat{\mu}_i^2 - \mu_i)^2$. Results are reported for $R = 5$, 50, and 500, corresponding to very sparse, sparse, and dense signals. The non-zero $\mu_i's$ were set at $V = 3, 4, 5$, and 7, representing a range of the strength of the signals. For each setting, results were averaged over 100 repetitions.

4.2 Illustrative examples

Figure 1 shows an example of a sequence with $R = 5$ and $V = 7$. Figure 2(a) shows our estimate for the normal non-zero mixture prior with 0/1+L_2 mixture loss and penalty $K = 10$. Note that most of the values in the estimate are zero. Figure 2(b) shows the behavior of our estimator for this signal as a function of the observation–it is indeed a thresholding rule in which all x_i below a certain level are set to zero. The threshold depends on the sparsity and the strength of the signal, and is automatically determined. Figure 2(c) and (d) are the same but without the 0/1 loss; this leads to shrinkage but no thresholding. Figure 3 show the same results with L_1 loss. While $K = 0$ (posterior median) also results in thresholding, the TSE $K = 10$ is much smaller than with $K = 0$.

4.3 Dependence on K

The performance of the empirical Bayes procedure depends on K, the amount of penalization for a wrong estimation. Figure 4, 5, and 6, show the TSE as a function of K for $R = 5$, 50, and 500 respectively. We compare the TSE for the four different cases over a range of values of K–using two different mixture loss functions (0/1+L_2 or 0/1+L_1) and two different non-zero component of the mixture prior (normal or Laplace). In each plot the dashed horizontal line corresponds to the estimator with $K = 0$. For ease of presentation we show the plots only for $V = 3$ and $V = 7$. The following observations can be made:

1. Note that $K = 0$ shows comparable performance only for very sparse small strength signals; for the rest of the cases the proposed thresholding rule with $K > 0$ has a lower TSE.

2. Very sparse signal ($R = 5$) (see Figure 4). For a small strength signal ($V = 3$), L_2 loss with Laplace prior has lower average TSE over all values of K. For a large strength signal ($V = 7$),

L_1 loss with Laplace prior has lower average TSE over a range of values of K. However, L_2 loss with laplace prior can still achieve a lower TSE for larger K.

3. Sparse signal ($R = 50$) (see Figure 5). For a small strength signal ($V = 3$), L_2 loss with normal prior has lower average TSE over all values of K. For a large strength signal ($V = 7$), L_1 loss with normal prior has lower average TSE over a range of values of K.

4. For both Very sparse signal ($R = 5$) and Sparse signal ($R = 50$), in terms of the minimum TSE that can be achieved, there is no significant difference between all the four methods.

5. Dense signal ($R = 500$) (see Figure 6). For small strength signals ($V = 3, 4$), L_2 loss with normal prior has lower average TSE over all values of K. For a large strength signal ($V = 7$), L_1 loss with normal prior has lower average TSE over a range of values of K.

6. Note that in each plot, the dashed horizontal line for L_1 loss with laplace prior and $K = 0$ corresponds to the method proposed by Johnstone and Silverman (2004).

4.4 Choice of K

In our simulations, the optimal performance (in terms of the total squared error) of our proposed method depends on the choice of K. In practice, we need to use a suitable K in order to compare fairly with the other methods.

If the Bayesian model is correct, *i.e.*, the prior for the non-zero part is correctly specified, and if the estimator is to be judged by TSE, then the optimal K is zero. However from the simulations we saw that there was some optimal $K > 0$ which achieved a lower TSE. This happens because in practice the prior is often mis-specified and our proposed mixture loss with $K > 0$ makes the estimator more robust.

Empirically we have found that the optimal value of K that minimizes TSE depends only on the estimated hyperparameters $\hat{a}$ and $\hat{w}$. We found that choosing $K \propto 10^c/\hat{w}$ –where $c = \hat{a}$ for the normal prior and $c = \sqrt{2}/\hat{a}$ for the Laplace prior–was very close to the optimal K. The hyperparameter $\hat{a}$ determines the spread of the non-zero part of the prior and is directly related to strength of the signal V. As c increases the optimal K increases. The hyperparameter $\hat{w}$ determines the amount of sparsity in the signal. If the signal is not very sparse we need a larger K in order to achieve the minimum possible TSE.

Figures 7(a) and 8(a) show two sample plots comparing the optimal value of K (as obtained from the simulation experiments) and the value automatically chosen by our procedure ($K = C10^c/\hat{w}$, where $C = 10^{-3}$ is a constant) as functions of the estimated hyperparameters $\hat{a}$ and $\hat{w}$. The results are for $N = 1000$, $0/1 + L_2$ mixture loss, and normal non-zero prior. The plots were generated by running the simulations for different values of R and V, and are averaged over 100 repetitions. It can be seen that the proposed value of K is very close to the optimal one. This can also be seen in Figures 7(b) and 8(b) that show that there is no significant difference between the minimum TSE that can be achieved by the optimal K and that obtained by our procedure. It should be noted that for model selection type applications, the optimal K can always be chosen by a suitable cross-validation procedure.

4.5 Adapting to unknown sparsity

The hyperparameter w is directly related to the signal sparsity, *i.e*, the fraction of zeros in $\boldsymbol{\mu}$. Figure 9(a) plots the estimated w as a function of $1 - R/N$ (the fraction of zeros in the signal)– varying from 0.5 (moderately sparse setup) to 0.99 (very sparse). The results are averaged over 100 repetitions. For both the normal and the Laplace prior, as sparsity tends to one, the estimate of w becomes more accurate. Between the normal and Laplace prior, the normal gives more accurate estimates of w.

The estimator used here would converge to the true parameters w and a if the family of priors used in the estimator contains the true prior. However, in reality (and also in the simulation setup used here) this may not be true. Because of the mis-specification of the prior, the estimate of w may not be that accurate (especially for moderately sparse signals, see Figure 9(a)). While the estimated w roughly captures the amount of sparsity in the signal, for our estimator further sparsity is obtained because of the penalty term in the mixture loss function. This can be seen in Figure 9(b) which plots the actual fraction of zeros in the estimate for those methods that result in thresholding. In each of these the penalty K was chosen using the proposed heuristic rule. Note that all the proposed estimators with $K > 0$ penalty clearly adapt to the unknown sparsity in the signal.

Figures 9(c) and (d) plot the corresponding false positive rate (the fraction of zeros incorrectly labeled as non-zero) and the false negative rate (the fraction of non-zeros incorrectly labeled as zero). We see that the mixture loss function (with $K > 0$ penalty) results in a much lower false positive rate than the estimator with $K = 0$, while at the same time maintaining comparable false negative rates.

4.6 Reduction in TSE due to mixture loss

The mixture loss function also results in a reduction in the total squared error of the estimate. Figure 10(a) plots the total squared error for the same setup for different loss functions. We see that the proposed estimator with $K > 0$ results is a lower TSE than the estimator with $K = 0$ (without the mixture loss). This can be clearly seen in Figure 10(d) which plots the corresponding reduction in TSE obtained due to the mixture loss function. The normal prior with L_2 loss gives the best performance.

4.7 Loss functions and priors

The L_1 loss itself (with $K = 0$) can also result in some thresholding. However our proposed estimator, involving K in the loss, results in a more accurate threshold. From Figure 9(b) we can see that adding the penalty term in the mixture loss function accurately captures the net sparsity in the signal, but using L_1 loss alone does not. Also the total squared error is smaller with the mixture loss (See Figure 10).

4.8 Comparison with other methods

We compare our proposed method with some of the best performing methods in Table I of Johnstone and Silverman (2004).

- The EBayesThresh (Johnstone and Silverman (2004)) method is a special case of our proposed method, *i.e.*, Laplace prior with $K = 0$. Results are reported for both the posterior median and mean.

- The SURE (Donoho and Johnstone (1995)) method minimizes Stein's unbiased risk estimate for the mean squared error of soft thresholding.

- The FDR (Abramovich, Benjamini, Donoho, and Johnstone (2006)) method is derived from the principle of controlling the false discovery rate in simultaneous hypothesis testing.

- The Universal hard threshold corresponds to using a thresholded MLE with the threshold $\sqrt{2 \log N}$.

- We also compare our method with the Lasso estimator. This is a special case of our proposed estimator with the mixture parameter $w = 0$ and a Laplace prior for the non-zero part.

Table 1 shows the results of our simulations for various choices of the prior and the loss using our recommended K. We also tabulate the results for $K = 0$. The results for EBayesThresh, FDR, SURE, and Universal hard threshold are directly taken from Table I in Johnstone and Silverman (2004). The following observations can be made.

1. For very sparse signals($R = 5$) the proposed method and the EBayesThresh show similar performance.

2. For sparse($R = 50$) and dense($R = 500$) signals the proposed method is better than the other methods.

3. The posterior mean or median, *i.e.*, $K = 0$ shows good performance only for very sparse small strength signals. For the rest of the cases our proposed thresholding rule with the mixture loss function is superior.

4. For very sparse signals($R = 5$) the laplace and normal priors have similar errors.

5. For sparse($R = 50$) and dense($R = 500$) signals the normal prior in general performs better than the laplace.

6. The FDR method shows good performance for some settings; this depends on the choice of q, which varies from case to case.

7. We also ranked the different procedures by computing the average rank among all the different scenarios. Based on the table the proposed method with L_2/L_1 loss and normal prior seems to be the best performing one.

5 Feature selection for classification

In a typical two-class classification scenario we are given a training set $\mathcal{D} = \{(\boldsymbol{x}_j, y_j)\}_{j=1}^{N}$ containing N instances, where $\boldsymbol{x}_j \in \mathbf{R}^d$ is an instance (the d-dimensional feature vector) and $y_j \in \mathcal{Y} = \{0, 1\}$ is the corresponding known class label. The task is to learn a classification function $f : \mathbf{R}^d \rightarrow \mathcal{Y}$ that generalizes well on unseen datasets. During the past few decades it has become relatively easy to gather datasets with a huge number of features. In such situations very often we would like the classification function f to use as few features as possible without any appreciable decrease in predictive accuracy. Feature selection is very often beneficial for cost effectiveness and interpretability. In many situations it also increases the prediction accuracy by preventing overfitting. While many sophisticated methods have been proposed for feature selection (see Guyon and Elisseeff (2003) for a review), one of the earliest and the most widely used algorithms is feature ranking. This is a very simple and scalable method that has had considerable empirical success either as a stand-alone feature selection mechanism, or as a pre-processing step for other methods.

Essentially, for each feature, a score measuring the degree of relevance to the label is computed. The features are then ranked in the order of decreasing scores. Only the top most relevant features are used and the rest are discarded. The number of features to retain is often based on ad-hoc rules and/or domain knowledge. An important issue is how to set the threshold between the relevant and irrelevant features.

Let z_i, $i = 1, \ldots, d$, be the computed ranking criterion for the i^{th} feature. Various ranking scores have been used in different application domains (Guyon and Elisseeff (2003)). Commonly used scores are related to Fisher's criterion or the t-test criterion, but the form may vary. For example z_i can be a scaled difference between means among two classes.

$$z_i = \frac{m_i^+ - m_i^-}{\sqrt{\frac{(\sigma_i^+)^2}{N^+} + \frac{(\sigma_i^-)^2}{N^-}}}, \tag{42}$$

where m_i^+ and m_i^- are the means, $(\sigma_i^+)^2$ and $(\sigma_i^-)^2$ are the variances, and N^+ and N^- are the number of examples of the positive and negative class, respectively. Note that if a feature is irrelevant then z_i is close to zero.

We assume that each z_i is a noisy realization of some underlying μ_i, $z_i = \mu_i + \epsilon_i$, where the ϵ_i are independent and identically distributed as $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$, a normal distribution with mean zero and a known variance σ^2 (which can be well estimated when it is unknown). The normality assumption of the score z_i is valid among most commonly used scores. Even though the features are not necessarily independent, it is a reasonable assumption in the context of feature ranking methods. It has also been noticed that for datasets with a large number of features, treating features independently give us as good a classifier as others or even better ones (see for example Domingos and Pazzani (1997); Bickel and Levina (2004)).

Note that if a feature is irrelevant, then the corresponding μ_i is zero, relevant features have $\mu_i \neq 0$. Based on the observation $\boldsymbol{z} = (z_1, z_2, \ldots, z_d)$, we need to find a desirable estimate $\widehat{\boldsymbol{\mu}}$ of the unknown parameters $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_d)$, and to determine how many of them are zero. With this setup we can directly use the proposed Bayesian thresholding procedure to estimate $\widehat{\boldsymbol{\mu}}$ and select the number of relevant features.

5.1 Experimental validation

Table 2 summarizes the four publicly available datasets used in our experiments. These datasets were downloaded from `http://www.nipsfsc.ecs.soton.ac.uk/datasets/` and `http://www.agnostic.inf.ethz.ch/datasets.php`, and have been previously used for feature selection challenges.

Our proposed method of selecting the relevant features can be used in conjunction with any ranking criterion having an approximate normal distribution. For our experiments we used the two-sample t-statistic (42) for the scores. We compare dthe number of features selected by the proposed method to those selected by the cross-validated area under the ROC curve (AUC) based criterion. The AUC based criterion operates as follows: features are first sorted based on the absolute value of the two-sample t-statistic; a linear discriminant analysis (LDA) classifier is trained using the top t features. The performance of this classifier was tested on an independent validation set (see Table 2). We used the AUC as our performance metric. This was repeated for t varying from 1 to d–the number of features. The optimal number of features selected was the value of t where the AUC on the validation set was maximum.

Table 3 compares the number of features selected by the proposed method with those selected by the cross-validated AUC based criteria. The resulting AUC on the validation set was also compared. This table can be studied in conjunction with Figure 11 which plots the AUC on both the training and the validation set as a function of t. The following observations can be made.

1. From Table 3, for most datasets the number of features selected by the proposed algorithm is very close to those selected by the cross-validation based criterion; in cases where they differ the proposed method achieves very similar AUC on the test set.

2. For some datasets (see Figures 11(a) and (b)) the AUC on the validation set reaches a peak and then starts decreasing. In such cases the number of features selected by the proposed algorithm is very close to the features selected by cross-validation.

3. For some datasets (see Figures 11(c) and (d)), the AUC on the validation set saturates after which there is no further significant improvement in the AUC by including more features. In such cases the proposed method achieves very similar AUC on the validation set.

4. The proposed algorithm shows good generalization properties. The number of features selected by the proposed algorithm leads to the best AUC on the validation set.

5. The proposed method is computationally efficient. It does not require any sort of cross-validation. The cross-validated AUC based criteria is very time and memory consuming,

especially if the number of features is large.

A Appendix: Computational details

In this appendix we list the details needed for the implementation of our thresholding rules.

A.1 Quantities of interest for the Normal prior

For the normal prior $\gamma_a(\mu) = \mathcal{N}(\mu|0, a^2) = (2\pi a^2)^{-1/2} \exp(-\mu^2/2a^2)$, we have $g_a(x) = \mathcal{N}(x|0, 1 + a^2)$, $G(\mu) = \mathcal{N}(\mu|m, \sigma^2)$, where $m = \frac{a^2}{1+a^2}x$ and $\sigma = \sqrt{\frac{a^2}{1+a^2}}$. Then $E_G[\mu] = m$ and $E_G[|\mu|] = m[2\Phi(\frac{m}{\sigma}) - 1] + 2\phi(\frac{m}{\sigma})$, where $\Phi(x) = \int_{-\infty}^{x} \phi(z)dz$ is the cumulative distribution function of the standard normal $\phi(x) = \mathcal{N}(x|0, 1)$. The solution to $\int_{-\infty}^{y} G(\mu)d\mu = c$ is given by $y = \sigma\Phi^{-1}(c) + m$, and $\int_{-\infty}^{\infty} |\hat{\mu} - \mu|G(\mu)d\mu = (m - \hat{\mu})(2\Phi(\frac{m-\hat{\mu}}{\sigma}) - 1) + 2\phi(\frac{m-\hat{\mu}}{\sigma})$.

A.2 Quantities of interest for the Laplace prior

For the Laplace prior $\gamma_a(\mu) = \frac{a}{2} \exp(-a|\mu|)$ we have the following,

$$g_a(x) = \frac{a}{2} \exp(a^2/2) \left[\exp(-ax)\Phi(x - a) + \exp(ax)(1 - \Phi(x + a))\right]. \tag{43}$$

$$G(\mu) = \frac{\exp(-\mathrm{sgn}(\mu)ax)\mathcal{N}(\mu|x - \mathrm{sgn}(\mu)a, 1)}{\exp(-ax)\Phi(x - a) + \exp(ax)(1 - \Phi(x + a))}, \tag{44}$$

where $\mathrm{sgn}(z) = 1$ if $z \geq 0$ and $\mathrm{sgn}(z) = -1$ if $z < 0$.

$$E_G[\mu] = \frac{\exp(-ax)\left((x - a)\Phi(x - a) + \phi(x - a)\right) + \exp(ax)\left((x + a)(1 - \Phi(x + a)) - \phi(x + a)\right)}{\exp(-ax)\Phi(x - a) + \exp(ax)(1 - \Phi(x + a))}. \tag{45}$$

$$E_G[|\mu|] = \frac{\exp(-ax)\left((x - a)\Phi(x - a) + \phi(x - a)\right) - \exp(ax)\left((x + a)(1 - \Phi(x + a)) - \phi(x + a)\right)}{\exp(-ax)\Phi(x - a) + \exp(ax)(1 - \Phi(x + a))}. \tag{46}$$

$$\int_{-\infty}^{y} G(\mu)d\mu = \begin{cases} \frac{e^{ax}\Phi(y-(x+a))}{e^{-ax}\Phi(x-a)+e^{ax}(1-\Phi(x+a))} & \text{if } \ y < 0 \\ \frac{e^{ax}\Phi(-(x+a))+e^{-ax}(\Phi(y-(x-a))-\Phi(-(x-a)))}{e^{-ax}\Phi(x-a)+e^{ax}(1-\Phi(x+a))} & \text{if } \ y \geq 0 \end{cases}. \tag{47}$$

The solution to $\int_{-\infty}^{y} G(\mu)d\mu = c$ is given by

$$y = \begin{cases} x + a + \Phi^{-1}\{c(e^{-2ax}\Phi(x - a) + 1 - \Phi(x + a))\} & \text{if } \ y < 0 \\ x - a + \Phi^{-1}\{1 + x - a + (c - 1)(e^{2ax} + \Phi(x - a) - e^{2ax}\Phi(x + a))\} & \text{if } \ y \geq 0 \end{cases}. \tag{48}$$

$$\begin{aligned} &\int_{-\infty}^{\infty} |\mu - \hat{\mu}|G(\mu)d\mu \\ = &\begin{cases} \frac{e^{ax}\{2(\hat{\mu}-x-a)\Phi(\hat{\mu}-x-a)-(\hat{\mu}-x-a)\Phi(-(x+a))-\phi(x+a)\}+e^{-ax}\{\phi(a-x)-(\hat{\mu}-x+a)\Phi(x-a)\}}{e^{-ax}\Phi(x-a)+e^{ax}(1-\Phi(x+a))} & \text{if } \ \hat{\mu} < 0 \\ \frac{e^{ax}\{(\hat{\mu}-x-a)\Phi(-(x+a))+\phi(x+a)\}e^{-ax}\{2(\hat{\mu}-x+a)\Phi(\hat{\mu}-x+a)-(\hat{\mu}-x+a)\Phi(a-x)-(\hat{\mu}-x+a)+2\phi(\hat{\mu}-x+a)-\phi(a-x)\}}{e^{-ax}\Phi(x-a)+e^{ax}(1-\Phi(x+a))} & \text{if } \ \hat{\mu} \geq 0 \end{cases} \end{aligned} \tag{49}$$

References

Abramovich, F., Benjamini, Y., Donoho, D. L., and Johnstone, I. M. (2006). Adapting to unknown sparsity by controlling the false discovery rate. *The Annals of Statistics 34*, 584–653.

Abramovich, F., Sapatinas, T., and Silverman, B. W. (1998). Wavelet thresholding via a Bayesian approach. *Journal of Royal Statistical Society B 60*, 725–749.

Berger, J. (1985). Statistical Decision Theory and Bayesian Analysis. *Springer-Verlag, New York.*

Bickel, P. and Levina, E. (2004). Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli 10*, 989–1010.

Brown, L. D. (2008). In-season prediction of batting averages: a field test of empirical bayes and bayes methodologies. *The Annals of Applied Statistics 2*, 113–152.

Brown, L. D. and Cohen, A. (1976). A complete class theorem for strict monotone likelihood ratio. *The Annals of Statistics 4*, 712–722.

Brown, L. D. and Greenstein, E. (2009). Nonparametric empirical Bayes and compound decision approaches to estimation of a high dimensional vector of normal means. *The Annals of Statistics 37*, 1685-1704.

Chipman, H. A., Kolaczyk, E. D., and McCulloch, R. E. (1997). Adaptive Bayesian wavelet shrinkage. *Journal of the American Statistical Association 92*, 1413–1421.

Clyde, M., Parmigiani, G., and Vidakovic, B. (1998). Multiple shrinkage and subset selection in wavelets. *Biometrika 85*, 391–401.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unkown smoothness via wavelet shrinkage. *Journal of American Statistical Association 90*, 1200–1224.

Domingos, P. and Pazzani, M. J. (1997). On the optimality of the simple bayesian classifier under zero-one loss. *Machine Learning 29*, 103–130.

Efron, B. and Tibshirani, R. (2007). On testing the significance of sets of genes. *The Annals of Applied Statistics 1*, 107–129.

George, E. I. and Foster, D. P. (2000). Calibration and empirical Bayes variable selection. *Biometrika 87*, 731–747.

Guyon, I. and Elisseeff, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research 3*, 1157–1182.

Johnstone, I. M. and Silverman, B. W. (2004). Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics 32*, 1594–1649.

Johnstone, I. M. and Silverman, B. W. (2005). Empirical Bayes selection of wavelet thresholds. *The Annals of Statistics 33*, 1700–1752.

Table 1: Average of the total squared error for various signals and different settings of our proposed procedure. See Section 4 for details of the simulation setup. Some of the methods with the minimum error are in bold, the best method is underlined. The results for EBayesThresh, FDR, SURE, and Universal hard threshold are directly taken from Table I in Johnstone and Silverman (2004). The left most column ranks the different procedures by computing the average rank among the different scenarios.

	R (Number nonzero)	5	5	5	5	50	50	50	50	500	500	500	500
Rank	V (Value nonzero)	3	4	5	7	3	4	5	7	3	4	5	7
	Proposed method												
7	L2 loss normal prior K=0	**34**	**30**	**18**	12	**194**	158	112	80	857	819	752	665
1	**L2 loss normal prior**	**34**	**30**	**17**	10	**193**	**151**	**96**	**54**	**807**	**722**	**585**	**502**
9	L2 loss laplace prior K=0	**33**	**31**	**19**	10	200	167	120	86	862	889	832	711
3	L2 loss laplace prior	**33**	**30**	**18**	**6**	198	159	**102**	60	848	774	649	572
4	L1 loss normal prior K=0	**36**	**31**	**15**	**8**	213	149	**96**	70	819	778	697	620
2	**L1 loss normal prior**	**36**	**32**	**15**	**5**	214	**155**	**96**	**57**	**790**	**698**	**614**	**516**
6	L1 loss laplace prior K=0	**35**	**31**	**15**	**8**	213	**154**	**100**	74	859	874	790	661
5	L1 loss laplace prior	**35**	**32**	**15**	**6**	215	159	**102**	67	855	777	751	673
15	L1 loss lasso prior	50	69	78	88	238	272	290	332	857	876	876	889
	EBayesThresh												
8	Laplace prior pos. median	**36**	**32**	**17**	**8**	214	**156**	**101**	73	857	873	783	658
10	Laplace prior pos. mean	**34**	**32**	21	11	201	169	122	85	860	888	826	708
	FDR												
12	q=0.01	43	51	26	**5**	392	299	125	**55**	2568	1332	656	524
11	q=0.1	40	35	**19**	13	280	175	113	102	1149	744	651	644
16	q=0.4	58	58	53	52	298	265	256	254	919	866	860	860
14	SURE	38	42	42	43	202	209	210	210	829	835	835	835
13	Universal hard threshold	39	37	**18**	**7**	370	340	163	**52**	3672	3355	1578	**505**

Table 2: The four data sets used in our feature selection experiments.

Dataset	Training examples	Validation examples	Number of features	Domain
madelon	2000	600	500	Synthetic
gina	3153	315	970	Handwriting
ada	4147	415	48	Marketing
sylva	13086	1309	216	Ecology

Table 3: The number of features selected by the proposed method (with normal prior and L_2 loss) and those selected by the cross-validated AUC based criteria for the different datasets. The resulting AUC on the validation set is also shown.

Dataset	Features	Proposed procedure		Cross-validated AUC criterion	
		Features selected	Validation set AUC	Features	Validation set AUC
madelon	500	13	0.634	15	0.645
gina	970	292	0.943	280	0.946
ada	48	20	0.861	31	0.867
sylva	216	18	0.987	11	0.997

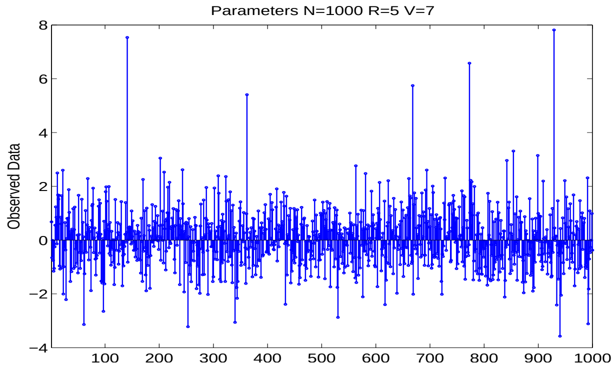
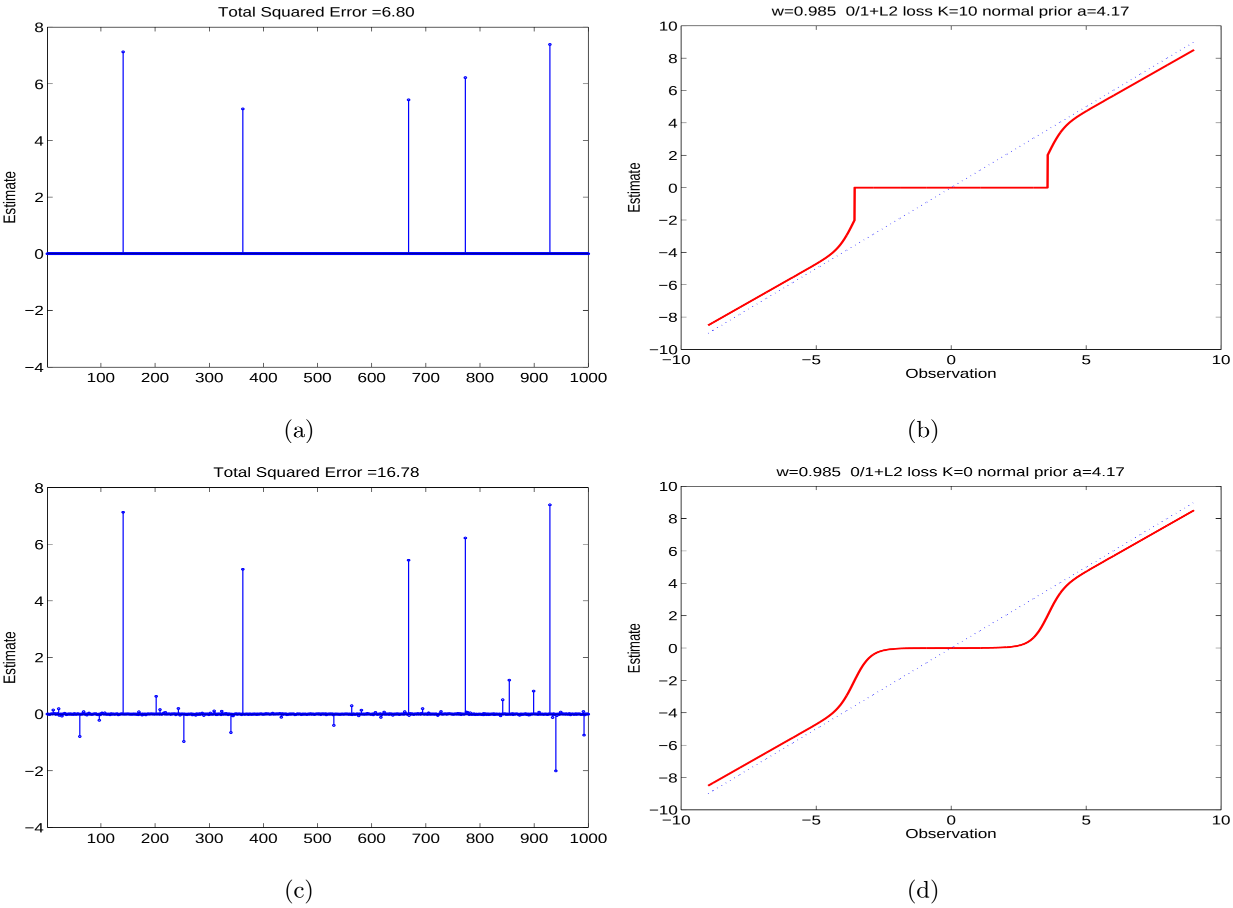


Figure 1: A sample sequence used in our simulation studies. The $N = 1000$ length sequence has $R = 5$ values that are non-zero, with signal strength $V = 7$.



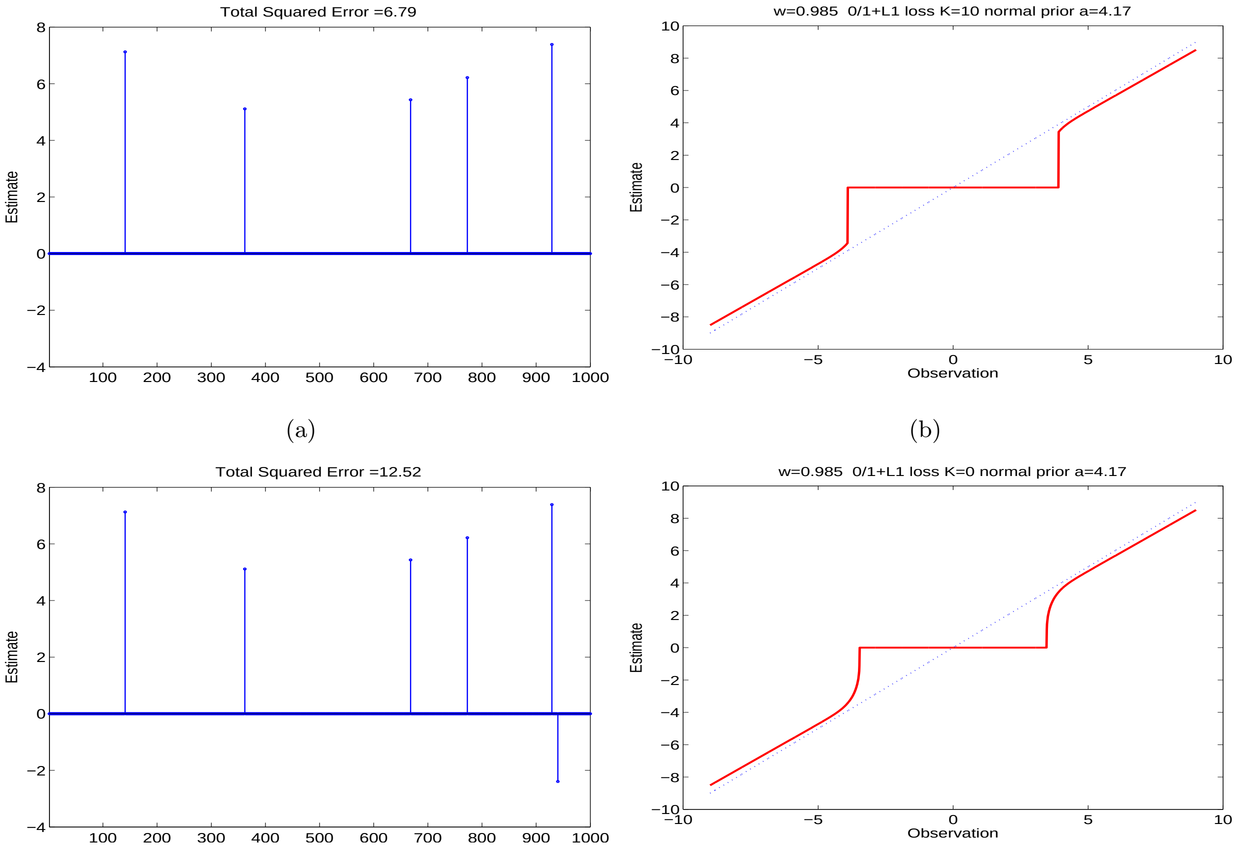
(a)

(b)

(c)

(d)

Figure 2: Illustration of our thresholding rule with $0/1{+}L_2$-mixture loss and normal non-zero prior for the sequence shown in Figure 1. (a) The estimated sequence with $K = 10$ penalty in the mixture loss. (b) The thresholding behavior of the estimator, the estimated hyperparameters a and w are also shown. (c) and (d) show the same without the 0/1 loss, *i.e*, $K = 0$. This leads to shrinkage but no thresholding.



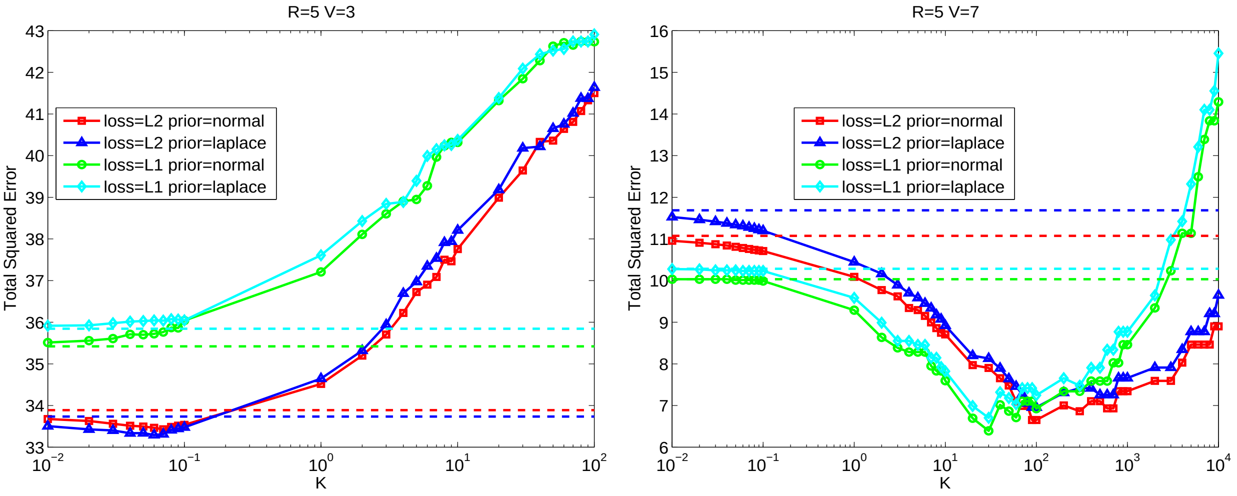
(a)

(b)

(c)

(d)

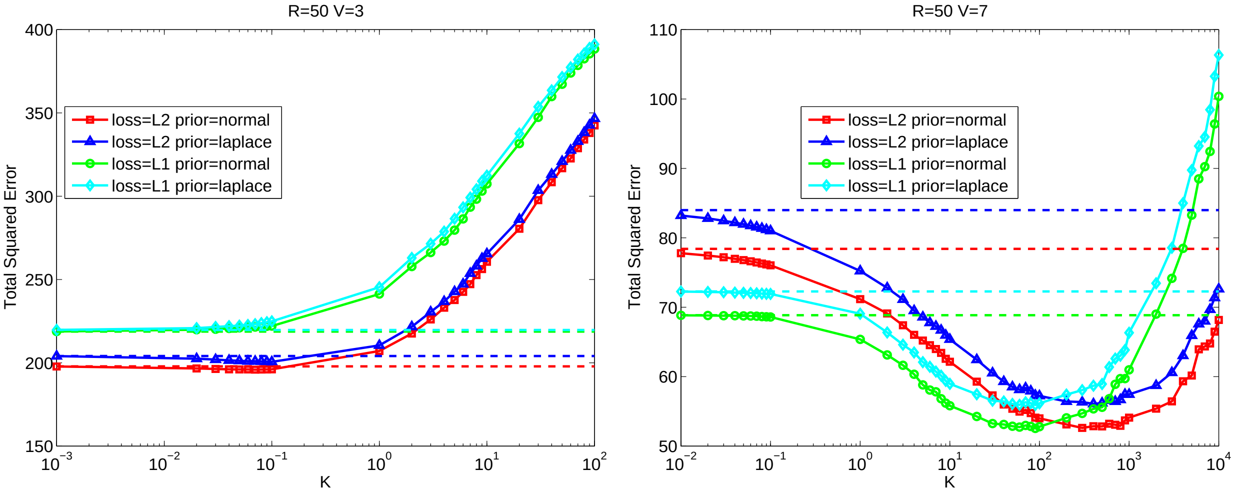
Figure 3: Illustration of our thresholding rule with $0/1{+}L_1$-loss and normal non-zero prior for the sample sequence shown in Figure 1. (a) The estimated sequence with $K = 10$ penalty in the mixture loss. (b) The thresholding behavior of the estimator, the estimated hyperparameters a and w are also shown. (c) and (d) show the same without the 0/1 loss, *i.e*, $K = 0$. While $K = 0$ also results in thresholding, the TSE with non-zero $K = 10$ is much smaller than with $K = 0$.



(a)

(b)

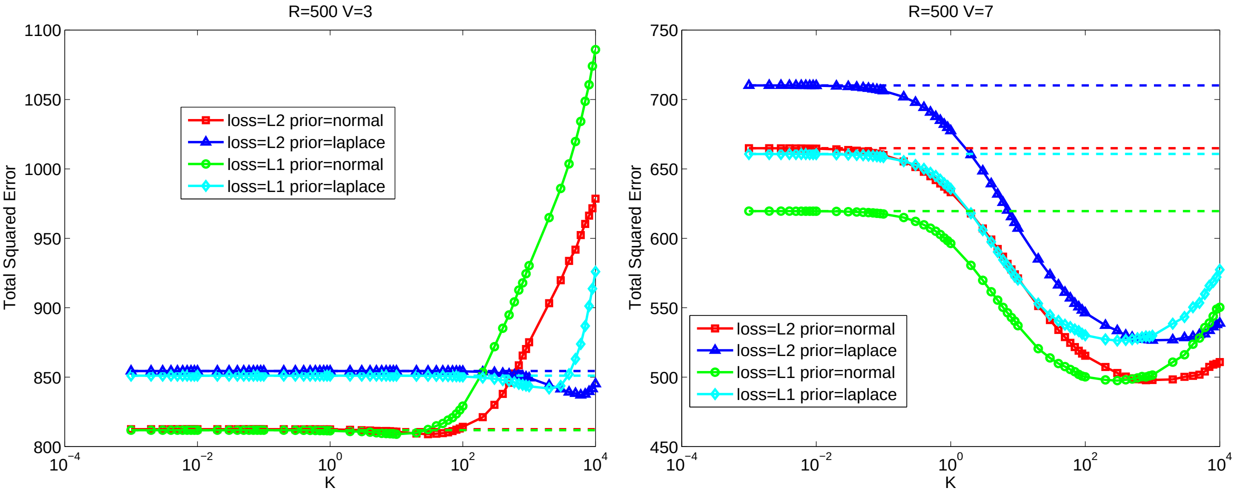
Figure 4: Very sparse signal $R = 5$. The total squared error (TSE) averaged over 100 trials as a function of K for different choices of the mixture loss functions ($0/1+L_1$ and $0/1+L_2$) and non-zero part of the mixture prior (normal and Laplace), and for signal strengths (a) $V = 3$ and (b) $V = 7$. In each plot the dashed horizontal line corresponds to the estimator with $K = 0$.



(a)

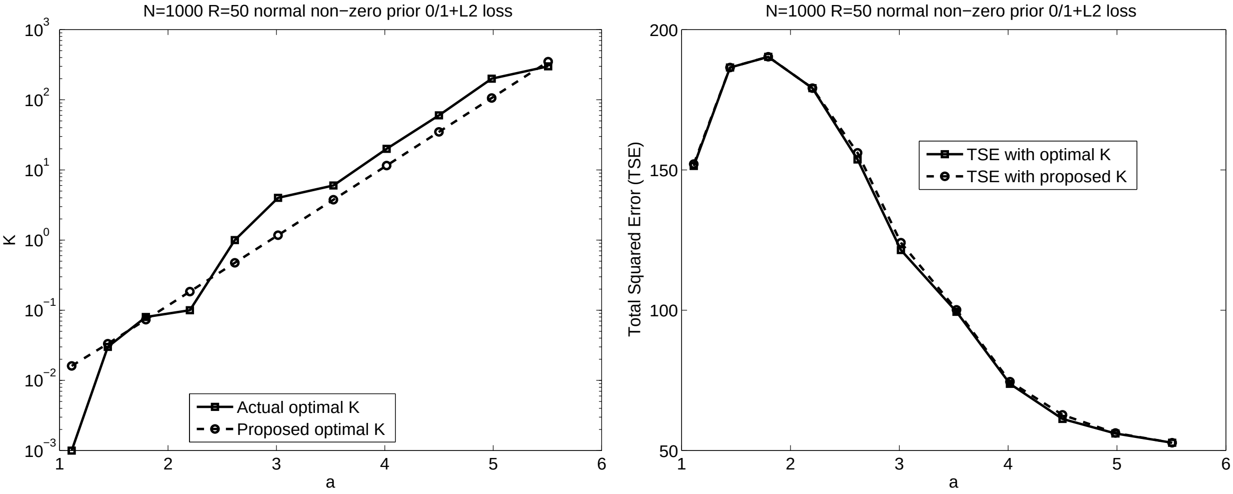
(b)

Figure 5: Sparse signal $R = 50$. The total squared error (TSE) averaged over 100 trials as a function of K for different choices of the mixture loss functions ($0/1+L_1$ and $0/1+L_2$) and non-zero part of the mixture prior (normal and Laplace), and for signal strengths (a) $V = 3$ and (b) $V = 7$. In each plot the dashed horizontal line corresponds to the estimator with $K = 0$.



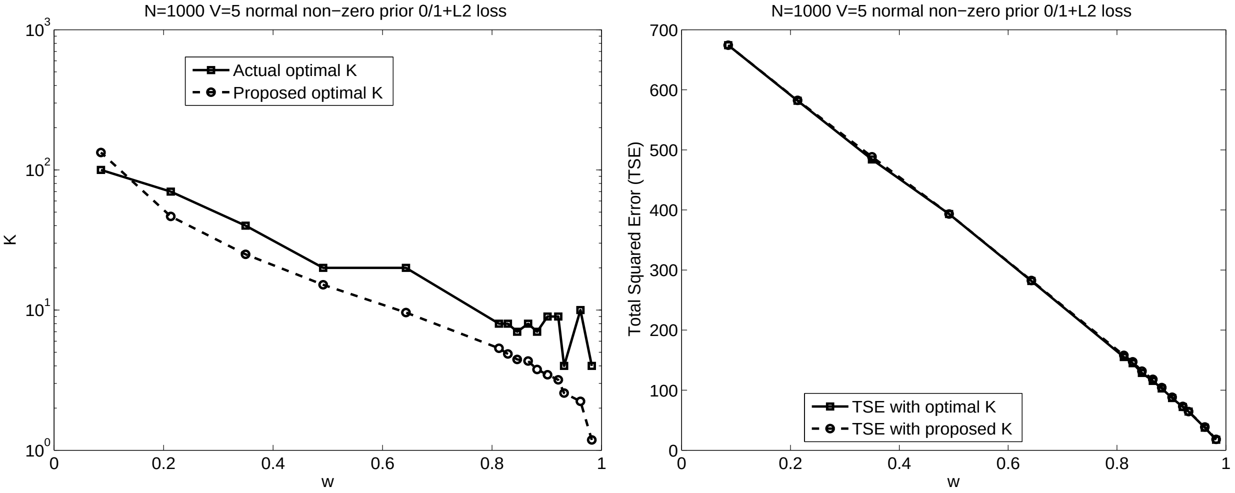
(a) (b)

Figure 6: Dense signal $R = 500$. The total squared error (TSE) averaged over 100 trials as a function of K for different choices of the mixture loss functions ($0/1+L_1$ and $0/1+L_2$) and non-zero part of the mixture prior (normal and Laplace), and for signal strengths (a) $V = 3$ and (b) $V = 7$. In each plot the dashed horizontal line corresponds to the estimator with $K = 0$.



(a) (b)

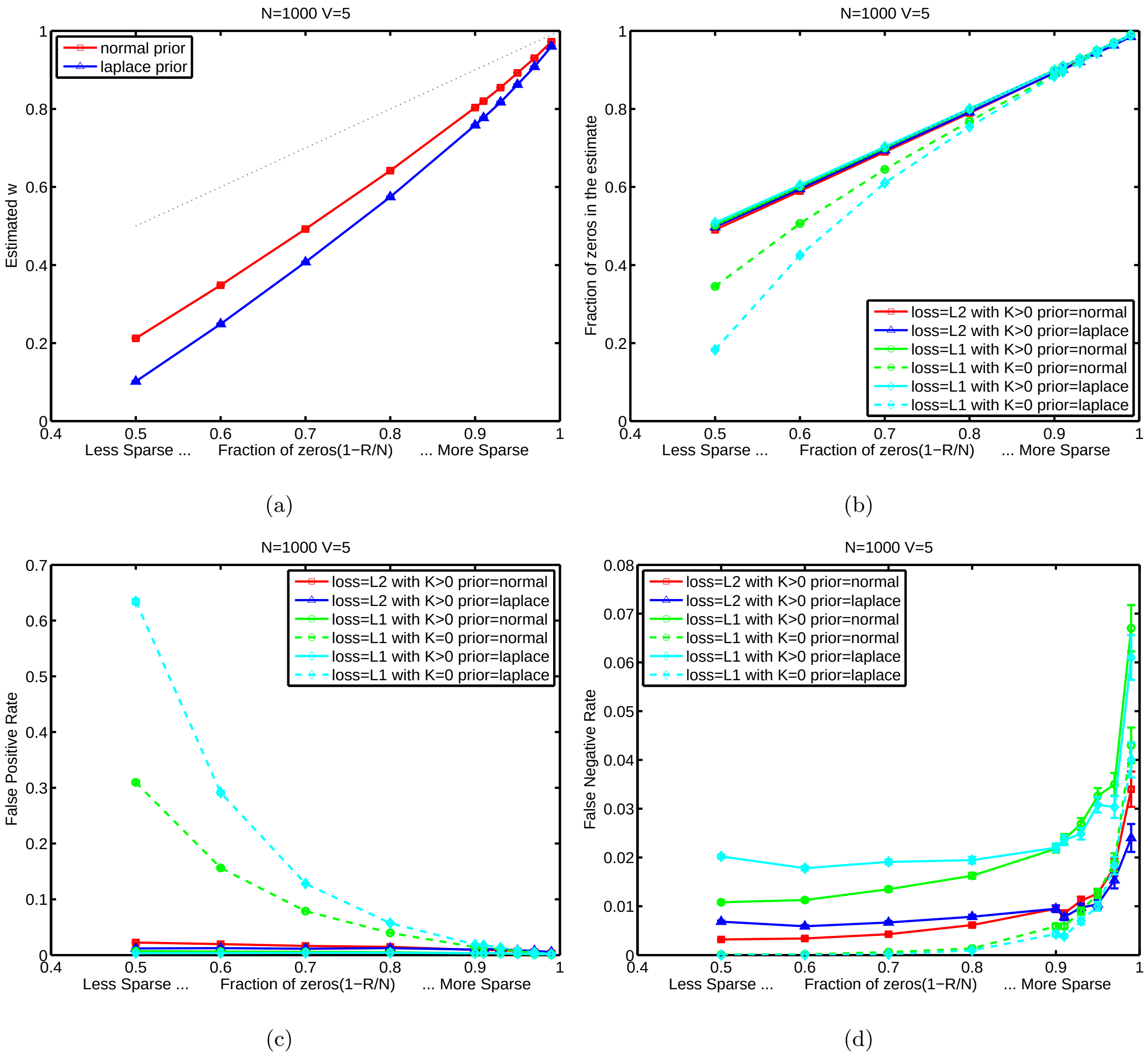
Figure 7: (a) Plots comparing the optimal value of the penalty K and the value recommended ($K = C10^c/\hat{w}$, where we have set the constant $C = 10^{-3}$) as a function of the estimated hyperparameter $\hat{a}$ for a sequence of length $N = 1000$ with $R = 50$ non-zero values; the normal non-zero prior and $0/1+L_2$ mixture loss was used. (b) Plots comparing the minimum TSE that can be achieved by the optimal K and that obtained by our procedure; plots were generated by running the simulations for different values of non-zero signal strength V and are averaged over 100 repetitions.



(a)

(b)

Figure 8: (a) Plots comparing of the optimal value of K and the value recommended ($K = C10^c/\hat{w}$, where we have set the constant $C = 10^{-3}$) as a function of the estimated hyperparameter $\hat{w}$ for a sequence of length $N = 1000$, with R non-zero values and signal strength for $V = 5$. (b) Plot comparing the minimum TSE that can be achieved by the optimal K and that obtained by our procedure; plots were generated by running the simulations for different number of non-zero values R and are averaged over 100 repetitions.



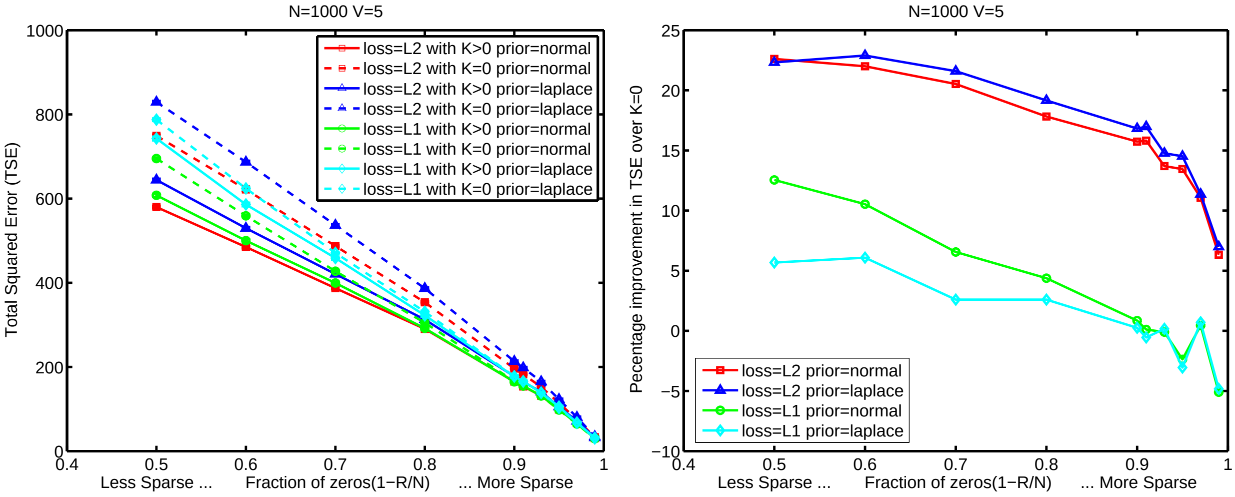
(a)

(b)

(c)

(d)

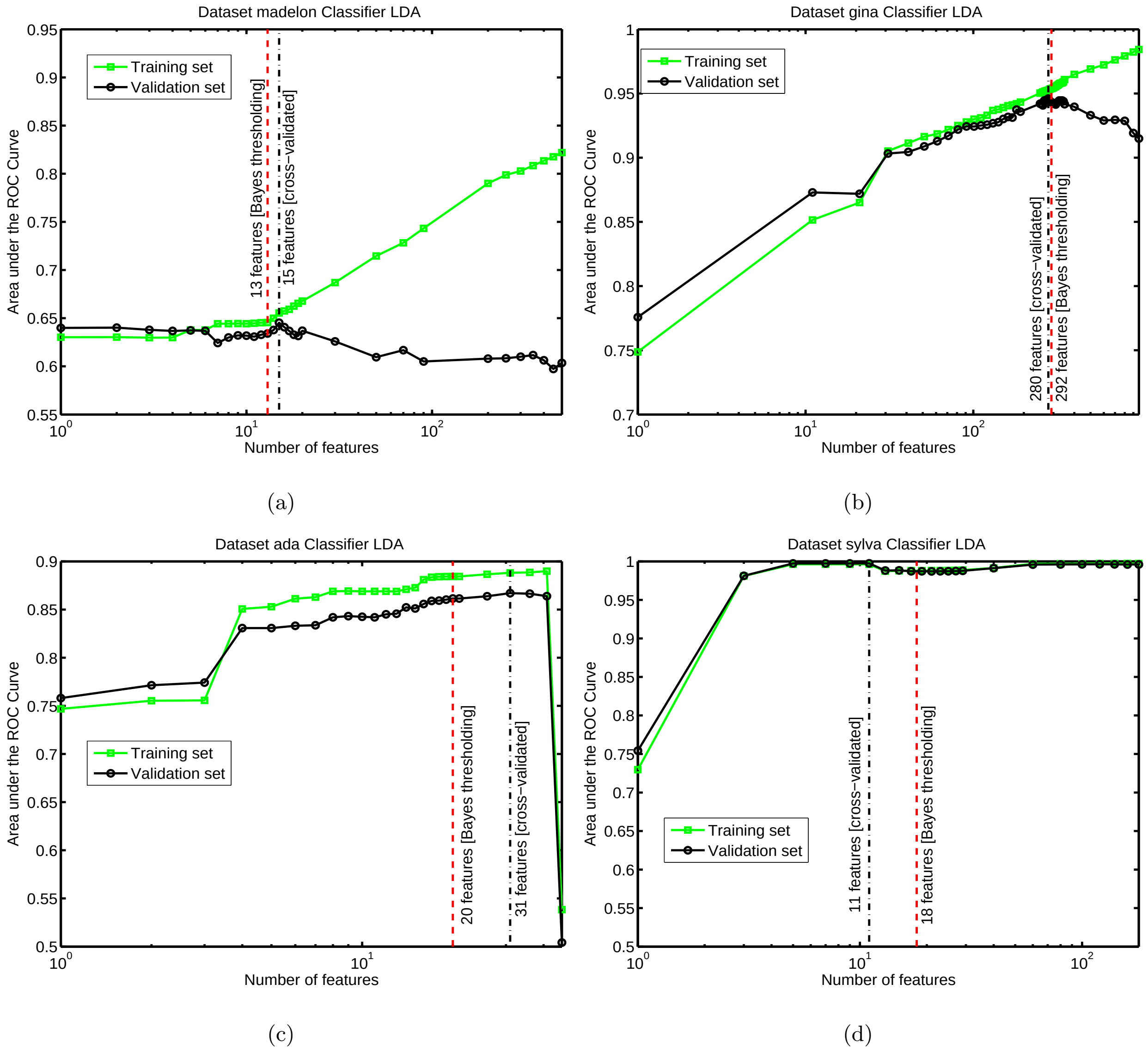
Figure 9: (a) The estimated hyperparameter w and (b) the actual fraction of zeros in the final estimate as a function of $1 - R/N$ (the fraction of zeros in the signal) for different choices of the mixture loss functions ($0/1{+}L_1$ and $0/1{+}L_2$) and non-zero part of the mixture prior (normal and laplace) and for signal strength $V = 5$. The corresponding (c) false positive rate (the fraction of zeros incorrectly labeled as non-zero) and (d) the false negative rate (the fraction of non-zeros incorrectly labeled as zero) are also shown. In each plot the dashed line corresponds to the estimator with $K = 0$.



(a)

(b)

Figure 10: (a) The total squared error (TSE) corresponding to the experiment in Figure 9. (b) The corresponding percentage improvement in TSE obtained due to the mixture loss function.



(a)

(b)

(c)

(d)

Figure 11: The Area under the ROC curve (AUC) for both the training and the validation set as a function of the number of top features used to train the LDA classifier for different datasets. The number of features selected by the proposed method (which requires no cross-validation) is marked as a red dotted line. The number of features selected by the cross-validation based AUC criterion is marked as a dotted black line.