Nonparametric prior for adaptive sparsity

Vikas C. Raykar
Siemens Healthcare
Malvern PA, USA
vikas.raykar@siemens.com

Linda H. Zhao
University of Pennsylvania
Philadelphia PA, USA
lzhao@wharton.upenn.edu

Abstract

For high-dimensional problems various parametric priors have been proposed to promote sparse solutions. While parametric priors has shown considerable success they are not very robust in adapting to varying degrees of sparsity. In this work we propose a discrete mixture prior which is partially nonparametric. The right structure for the prior and the amount of sparsity is estimated directly from the data. Our experiments show that the proposed prior adapts to sparsity much better than its parametric counterparts. We apply the proposed method to classification of high dimensional microarray datasets.

1 Adaptive Sparsity

In high-dimensional prediction/estimation problems (usually referred to as the *large p, small n* ($p \gg n$) paradigm, p being the dimension of the model and n the sample size) it is desirable to obtain sparse solutions. A sparse solution generally helps in better interpretation of the model and more importantly leads to better generalization on unseen data. While sparsity can be defined technically in various ways one intuitive notion is the proportion of model parameters that are zero (or very close to zero).

For ease of exposition and also analytical tractability we will focus on the problem of estimating a high-dimensional vector from a noisy observation. Specifically we are given p scalar observations $z_1, z_2, \ldots, z_p$ satisfying

$$z_i = \beta_i + \epsilon_i, \tag{1}$$

where each ϵ_i is independent and distributed as $\epsilon_i \sim \mathcal{N}(\epsilon_i|0, \sigma_i^2)$, *i.e.*, a normal distribution with mean zero and variance σ_i^2. Based on the observation $\boldsymbol{z} = (z_1, z_2, \ldots, z_p)$ we need to find an estimate $\hat{\boldsymbol{\beta}}$ of the unknown parameters $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)$. In the high-dimensional sparse scenario a large number of β_i's are zero (or near zero) but we do not know how many and which of them are exactly zero. Hence with no information on how sparse the vector $\boldsymbol{\beta}$ is, an estimator $\hat{\boldsymbol{\beta}}$ which adapts to the degree of sparsity is desirable.

This problem occurs in a wide range of practical applications, including model/feature selection in machine learning/data mining (Guyon and Elisseeff, 2003; Dudoit et al., 2002; Tibshirani et al., 2002) (where $\boldsymbol{\beta}$ could be the regression/classification weights), smoothing/de-noising in signal processing (Johnstone and Silverman, 2004), and multiple-hypothesis testing in genomics/bio-informatics (Abramovich et al., 2006; Efron and Tibshirani, 2007).

Estimators with following properties are desirable–

1. ***Shrinkage*** The maximum-likelihood estimator of $\boldsymbol{\beta}$ is the observation $\boldsymbol{z}$ itself. For large p, $\|\boldsymbol{z}\|$ is generally over-inflated, *i.e.*, quite larger than the true value $\|\boldsymbol{\beta}\|$ (This has to do with the geometry of high-dimensional distributions.). Hence this estimator can be considerably improved by suitable *shrinkage estimators*, which shrinks each estimate z_i towards zero.

2. ***Thresholding*** For model interpretation, feature selection, and significance testing applications it is desirable that the estimate have some values to be exactly zero (or almost close to zero).

3. ***Adaptive Sparsity*** It is very crucial that both shrinkage and thresholding properties adapt to the degree of the sparsity in the signal. A very sparse vector would generally need a large amount of shrinkage while a non-sparse signal would need no shrinkage. The degree of sparsity is generally unknown and has to be estimated from the data itself.

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9.

1.1 Bayesian shrinkage

In a Bayesian setting both shrinkage and thresholding can be achieved by imposing a suitable *sparsity promoting prior* on $\boldsymbol{\beta}$ and then using a point estimate based on the posterior of $\boldsymbol{\beta}$ (either the mean or median). Two broad categories of priors commonly used are the *shrinkage priors* and the *discrete mixture priors*. Commonly used shrinkage priors include the normal (leading to the James-Stein estimator (James and Stein, 1961)), Laplace (the Lasso (Tibshirani, 1996) estimator), student-t (relevance vector machine (Tipping, 2001)), the horseshoe (Carvalho et al., 2009)–all these priors have zero mean (to encourage sparsity) and an unknown parameter which controls the amount of shrinkage/sparsity. Though a full Bayesian treatment is optimal if the presumed prior is accurate, very often the unknown parameter is considered a *hyperparameter* and is estimated either by cross-validation or via an empirical Bayesian approach by maximizing the marginal likelihood.

1.2 Discrete mixture priors

A conceptually more suitable prior is the discrete mixture prior, which consists of a mixture prior on each $\boldsymbol{\beta}_i$ with an atom of probability w at zero and a suitable prior γ for the nonzero part, *i.e.*,

$$p(\beta_i|w,\gamma) = w\delta(\beta_i) + (1-w)\gamma(\beta_i), \tag{2}$$

where $w \in [0,1]$ is the mixture parameter and the $\delta(\beta_i)$ is defined as having a probability mass of 1 at $\beta_i = 0$ and zero elsewhere. This prior captures our belief that some of the β_i, are exactly zero. The mixing parameter w is the fraction of zeros in $\boldsymbol{\beta}$ and controls the sparsity in the signal. The sparsity parameter w is treated as a hyperparameter and estimated by maximizing the marginal likelihood. Typically a *parametric prior*, either a normal or a Laplace (Johnstone and Silverman, 2004) is used for the non-zero part. While parametric priors here have shown considerable success, they are not very robust because of the specific assumption on the shape of the prior. Our simulation results show that the estimate for w is biased and depends heavily on the mismatch between the distribution of the observation and shape of the prior used.

In this work we propose to use an unspecified distribution for the non-zero part of the mixture. We show that the non-zero part of the prior is only involved through the marginal density of the observations. Specifically under the mixture prior (2) our final estimator is of the form (§ 4)

$$\hat{\beta}_i = (1-\hat{p}_i)\left[z_i + \frac{\hat{g}'(z_i)}{\hat{g}(z_i)}\right], \tag{3}$$

where $\hat{g}(z_i)$ is an estimate of the marginal density $g(z_i) = \int \mathcal{N}(\mu_i|z_i,1)\gamma(\mu_i)d\mu_i$ corresponding to the non-zero means, $\hat{g}'(z_i)$ is the derivative of this estimate and $\hat{p}_i$ is the estimated posterior probability of β_i being zero. Notice that γ, the prior for the non-zero part is only involved through the marginal g. In our approach both $\hat{g}(z_i)$ and $\hat{p}_i$ are directly constructed from the observations $\boldsymbol{z}$. We use a weighted nonparametric kernel density to estimate the marginal and its derivative. We do not have to specify any prior or particular form of prior for the non-zero part, nor do we need to directly construct an estimate for that prior.

The hyperparameter w is estimated by maximizing the log marginal likelihood. Conditional on γ the marginal likelihood for w depends directly on the marginal distribution of z under γ , and not otherwise on the prior γ . Thus, given the estimate $\hat{g}$ of the marginal distribution of z under γ we estimate w by using this estimated marginal distribution calculated at $\hat{g}$. Finally, the weights in the weighted kernel density estimator depend on the posterior probability that the corresponding observation comes from the non-zero part of the prior. For given w and γ these weights depend on w and g, and we estimate these weights from the corresponding formulas evaluated at $\hat{w}$ and $\hat{g}$. The estimation procedure can be fully explained by a pattern of logic that resembles the logic in the familiar Expectation-Maximization(EM) algorithm (Dempster et al., 1977), but with the kernel density estimator of g used in one of the M-steps rather than a true maximum likelihood estimator.

The rest of the paper is organized as follows. In § 2 we describe the nonparametric mixture prior used and derive the posterior. We then describe an EM algorithm (§ 3) to estimate the hyperparameter w and the marginal g jointly. The estimated hyperparameters are then plugged in to derive the posterior mean(§ 4). Our simulation results(§ 5) show that the proposed algorithm adapts to sparsity much better than its parametric counterparts. In § 6 we apply the proposed procedure for regularization of high dimensional classification problems.

2 The Bayesian Setup

We will assume that the σ_i^2 are all equal and then assume without loss of generality that the z_i are scaled such that $\sigma_i^2 = 1$.

2.1 Likelihood

The likelihood of the parameters $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)$ given independent observations $\boldsymbol{z} = (z_1, \ldots, z_p)$ can

be factored as

$$p(\boldsymbol{z}|\boldsymbol{\beta}) = \prod_{i=1}^{p} p(z_i|\beta_i) = \prod_{i=1}^{p} \mathcal{N}(z_i|\beta_i, 1). \quad (4)$$

The maximum-likelihood estimator of $\boldsymbol{\beta}$ is the observation $\boldsymbol{z}$ itself. As noted, this is not an effective estimator in the context of our intended applications.

2.2 Mixture Prior

We will assume that each β_i comes independently from a mixture of a delta function with mass at zero and a completely unspecified nonparametric density γ, *i.e.*,

$$p(\beta_i|w, \gamma) = w\delta(\beta_i) + (1-w)\gamma(\beta_i), \quad (5)$$

where $w \in [0, 1]$ describes the prior probability that each $\beta_i = 0$. This prior thus captures our belief that some of the $\beta_i = 0$ are zero, and w describes the expected fraction of zeros. In model selection applications this corresponds to the number of irrelevant parameters. We treat w as a hyperparameter and estimate it using an empirical Bayes approach by maximizing marginal likelihood.

Very often we do not have any prior information about the distribution of the true non-zero scores. The choice of γ also plays a very important role in how well the hyperparameter w can be estimated. The normal and the Laplace prior have been previously used for γ (Johnstone and Silverman, 2004) and our simulations show that the estimate of w is then biased due to the prior mis-specification. *A novel aspect of our prior is that we will leave the nonzero part of the prior γ completely unspecified*. We will later show that the non-zero part of the prior is only involved through the marginal density of the observations and propose an algorithm to estimate it directly from the data.

2.3 Posterior

Given the hyper-parameter w and the prior γ the posterior of $\boldsymbol{\beta}$ given the data $\boldsymbol{z}$ can be written as

$$p(\boldsymbol{\beta}|\boldsymbol{z}, w, \gamma) = \frac{\prod_{i=1}^{p} p(z_i|\beta_i)p(\beta_i|w, \gamma)}{m(\boldsymbol{z}|w, \gamma)}, \quad \text{where} \quad (6)$$

$$m(\boldsymbol{z}|w, \gamma) = \prod_{i=1}^{p} \int p(z_i|\beta_i)p(\beta_i|w, \gamma)d\beta_i \quad (7)$$

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

$$\int p(z_i|\beta_i)p(\beta_i|w, \gamma)d\beta_i = w\mathcal{N}(z_i|0, 1) + (1-w)g(z_i),$$

where

$$g(z_i) = \int \mathcal{N}(\beta_i|z_i, 1)\gamma(\beta_i)d\beta_i. \quad (8)$$

Note g is the marginal density of z_i given that β_i is non-zero. The posterior in (6) can now be factored as follows $p(\beta|\boldsymbol{z}, w, \gamma) = \prod_{i=1}^{p} p(\beta_i|z_i, w, \gamma)$, where

$$\begin{aligned} & p(\beta_i|z_i, w, \gamma) \\ = & \frac{w\delta(\beta_i)\mathcal{N}(z_i|0, 1) + (1-w)\gamma(\beta_i)\mathcal{N}(z_i|\beta_i, 1)}{w\mathcal{N}(z_i|0, 1) + (1-w)g(z_i)} \\ = & p_i\delta(\beta_i) + (1-p_i)G(\beta_i). \end{aligned} \quad (9)$$

Here

$$p_i = p(\beta_i = 0|z_i, w, \gamma) = \frac{w\mathcal{N}(z_i|0, 1)}{w\mathcal{N}(z_i|0, 1) + (1-w)g(z_i)} \quad (10)$$

is the posterior probability of β_i being 0 and

$$G(\beta_i) = \frac{\mathcal{N}(\beta_i|z_i, 1)\gamma(\beta_i)}{\int \mathcal{N}(\beta_i|z_i, 1)\gamma(\beta_i)d\beta_i} \quad (11)$$

is the posterior density of β_i when it is not 0.

3 Adapting to unknown sparsity

The hyperparameter w is estimated by maximizing the marginal likelihood. The posterior of $\boldsymbol{\beta}$ is then computed by plugging in the estimated $\hat{w}$ and the mean of the posterior is used as a point estimate of $\boldsymbol{\beta}$.

3.1 Type II MLE for w

The hyperparameter w is the fraction of zeros in $\boldsymbol{\beta}$. We chose w to maximize the marginal likelihood–which is the likelihood integrated over the model parameters.

$$\widehat{w} = \arg\max_{w} m(\boldsymbol{z}|w, \gamma) = \arg\max_{w} \log m(\boldsymbol{z}|w, \gamma). \quad (12)$$

This is also known as the Type II maximum likelihood estimator for the hyperparameter. From (7) and (2.3) the log-marginal can be written as

$$\log m(\boldsymbol{z}|w, \gamma) = \sum_{i=1}^{p} \log\left[w\mathcal{N}(z_i|0, 1) + (1-w)g(z_i)\right], \quad (13)$$

where g is the marginal density of the non-zero $\{z_i's\}$. Note that γ, the prior for the non-zero part is only involved through the marginal $g(z_i) = \int \mathcal{N}(\beta_i|z_i, 1)\gamma(\beta_i)d\beta_i$. If we can estimate $g(z_i)$ directly then we do not have to specify any prior for the non-zero part.

3.2 Nonparametric estimate of the marginal

However g is the marginal of the non-zero part. So to estimate g we need to know whether each z_i comes from the zero part ($\beta_i = 0$) or the non-zero part

($\beta_i \neq 0$). This feature of the likelihood suggests defining latent (unobserved) random variables δ_i which describes whether $\beta_i = 0$. These variables takes the value $\delta_i = 1$ if $\beta_i = 0$, and $\delta_i = 0$ otherwise. If we know δ_i we can estimate g through a nonparametric kernel density estimate (Wand and Jones, 1995) $\hat{g}_{\boldsymbol{\delta}}$ of the following form

$$\hat{g}_{\boldsymbol{\delta}}(z) = \frac{1}{n^+ h} \sum_{j=1}^{p} (1-\delta_j) K\left(\frac{z - z_j}{h}\right), \qquad (14)$$

where K is a function satisfying $\int K(x)dx = 1$ called the kernel, h is the bandwidth of the kernel, and $n^+ = \sum_{j=1}^{p}(1-\delta_j)$ is the total number of non-zeros. A widely used kernel is a normal density of zero mean and unit variance, *i.e.*, $K(x) = \mathcal{N}(x|0,1)$. We set the bandwidth of the kernel using the normal reference rule (Wand and Jones, 1995) as $h = O(p^{-1/5})$.

3.3 EM algorithm

The estimation and maximization outlined above can be considerably simplified by proceeding in the manner of the EM algorithm (Dempster et al., 1977). The EM algorithm is an efficient iterative procedure to compute the maximum-likelihood solution in presence of missing/hidden data. We will use the unknown indicator variable δ_i as the missing data. If we know the missing data $\boldsymbol{\delta} = (\delta_1, \ldots, \delta_p)$ then the complete log marginal likelihood can be written as

$$\begin{aligned} & \log m(\boldsymbol{z}, \boldsymbol{\delta}|w, g) \\ = & \sum_{i=1}^{p} \log\left[\delta_i w \mathcal{N}(z_i|0,1) + (1-\delta_i)(1-w) g(z_i)\right]. \end{aligned}$$

Each iteration of the EM algorithm consists of two steps: an Expectation(E)-step and a Maximization(M)-step. The M-step involves maximization of a lower bound on the log-likelihood that is refined in each iteration by the E-step.

[1] **E-step.** Given the observation $\boldsymbol{z}$ and the current estimate of w and g the conditional expectation (which is a lower bound on the true likelihood) is

$$\begin{aligned} & \mathrm{E}_{\boldsymbol{\delta}|\boldsymbol{z},\hat{w},\hat{g}}\left[\log m(\boldsymbol{z}, \boldsymbol{\delta}|\hat{w}, \hat{g})\right] \\ = & \sum_{i=1}^{p} \hat{p}_i \log w\mathcal{N}(z_i|0,1) + (1-\hat{p}_i)\log\left((1-w)\hat{g}(z_i)\right) \\ = & \sum_{i=1}^{p} \left[\hat{p}_i \log w + (1-\hat{p}_i)\log(1-w)\right] + C, \qquad (15) \end{aligned}$$

where the expectation is with respect to $p(\boldsymbol{\delta}|\boldsymbol{z}, \hat{w}, \hat{g})$, and $\hat{p}_i = p(\delta_i = 1|z_i, \hat{w}, \hat{g}) = p(\beta_i = 0|z_i, \hat{w}, \hat{g})$ and is given by (10). The constant C is irrelevant to w. We have used the fact that if δ_i are independent Bernoulli random variables with probability $\hat{p}_i$ and a_i, b_i are any set of nonnegative constants then $E[\log(a_i\delta_i + b_i(1-\delta_i))] = \hat{p}_i \log a_i + (1-\hat{p}_i)\log b_i$.

[2] **M-step** The parameter $\hat{w}$ is then re-estimated by maximizing (15)). By equating the gradient of (15) to zero we obtain the following estimate

$$\hat{w} = \frac{\sum_{i=1}^{p} \hat{p}_i}{p}. \qquad (16)$$

A new estimate of $\hat{g}$ is obtained as follows

$$\hat{g}(z_i) = \frac{1}{\tilde{p}h} \sum_{j=1}^{p} (1-\hat{p}_j) K\left(\frac{z_i - z_j}{h}\right), \qquad (17)$$

where $\tilde{p} = \sum_{j=1}^{p}(1-\hat{p}_j)$. Each point gets weighted by $1-\hat{p}_j$. Here $\hat{p}_j$ is computed through (10) with $\hat{g}(z_i)$ updated from the last iteration, *i.e.*,

$$\hat{p}_i = \frac{\hat{w}\mathcal{N}(z_i|0,1)}{\hat{w}\mathcal{N}(z_i|0,1) + (1-\hat{w})\hat{g}(z_i)}. \qquad (18)$$

The final algorithm consists of the following two steps which are repeated till convergence.

1. Compute $\hat{p}_i$ using the current estimate $\hat{w}$ and $\hat{g}$ as follows
$$\hat{p}_i = \frac{\hat{w}\mathcal{N}(z_i|0,1)}{\hat{w}\mathcal{N}(z_i|0,1) + (1-\hat{w})\hat{g}(z_i)}. \qquad (19)$$
2. Re-estimate $\hat{w}$ and $\hat{g}(z_i)$ using the current estimator $\hat{p}_i$ as follows
$$\hat{w} = \frac{1}{p}\sum_{i=1}^{p} \hat{p}_i, \qquad (20)$$
$$\hat{g}(z_i) = \frac{1}{\tilde{p}h} \sum_{j=1}^{p} (1-\hat{p}_j) K\left(\frac{z_i - z_j}{h}\right). \qquad (21)$$

4 Posterior mean

The estimated hyperparameter $\hat{w}$ can be plugged into the posterior, and we use the posterior mean as a point estimate for β.

$$\hat{\beta}_i = (1-\hat{p}_i) E_G[\beta_i]. \qquad (22)$$

The mean of the posterior also depends only on the marginal and its derivative. Notice that

$$E_G[\beta_i] = \frac{\int \beta_i \mathcal{N}(\beta_i|z_i, 1)\gamma(\beta_i)d\beta_i}{\int \mathcal{N}(\beta_i|z_i,1)\gamma(\beta_i)d\beta_i} = z_i + \frac{\hat{g}'(z_i)}{\hat{g}(z_i)}. \qquad (23)$$

The marginal g and its derivative g' are both estimated through a weighted kernel density estimate. The kernel density derivative estimate is given by

$$\hat{g}'(z) = \frac{1}{\tilde{p}h^2} \sum_{j=1}^{p} (1-\hat{p}_j) K'\left(\frac{z - z_j}{h}\right). \qquad (24)$$

5 Experimental validation

In order to validate our proposed procedure we design the following simulation setup. A sequence $\boldsymbol{\beta}$ of length $p = 500$ is generated with different degree of sparsity and non-zero distribution. The sequence has $\beta_i = 0$ at wp randomly chosen positions, where the parameter $0 < w < 1$ controls the sparsity and is equal to the fraction of zeros in the sequence. The non-zero values in $\boldsymbol{\beta}$ are sampled from different distributions a few of which are illustrated in Figure 1. For each of these distributions there is parameter V which controls the strength of the signal. The observation z_i is generated by adding $\mathcal{N}(0, 1)$ noise for each β_i. Given $\boldsymbol{z}$ we are interested in estimating the hyperparameter w and recovering the true signal $\boldsymbol{\beta}$.

Figure 2(a) shows an example of a sequence $\boldsymbol{\beta}$ used in our simulations with $w = 0.9$ (a moderately sparse signal) and the non-zero samples coming from a BimodalGaussian distribution (with $V = 6$). Figure 2(b) shows the noisy observation $\boldsymbol{z}$. Figure 2(c) shows our estimate $\hat{\boldsymbol{\beta}}$ (the mean of the posterior) for the proposed method using a nonparametric prior. The proposed nonparametric method estimated w as 0.89. Figure 2(d) plots the value of the estimator as a function of the value of the observed signal. The estimates shrink towards zero. The shrinkage profile is determined by the estimated w and also the distribution of the non-zero part, which we have implicity determined from the data based on the marginal.

Figure 2(e)-(h) shows the estimated marginal of the non-zero part for four different stages of the EM algorithm. When we start the iteration we assume all observations come from the non-zero part. Hence we see a large peak at zero since most of the values are zero. It can be seen that as the EM algorithm proceeds w gets refined and marginal adapts to the non-zero part (in this case a BimodalGaussian).

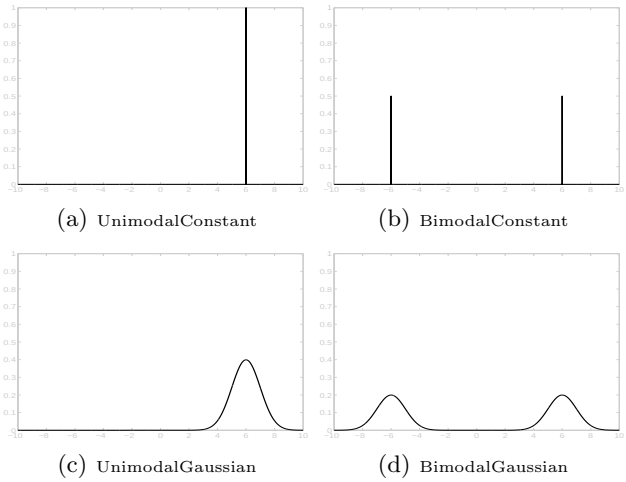
We compare our proposed method with the following different empirical Bayesian methods all of which use a mixture prior $p(\beta_i|w, \gamma) = w\delta(\beta_i) + (1 - w)\gamma(\beta_i)$.

1. **Nonparametric** The proposed procedure where the non-zero part γ is completely unspecified.

2. **Parametric normal** $\gamma = N(0, a^2)$ is a normal density where a is unknown.

3. **Parametric Laplace** (Johnstone and Silverman, 2004) γ is a Laplace density with an unknown dispersion parameter a.

4. **Nonparametric without mixing** Here γ is unspecified but there is no mixing, *i.e.*, $w = 0$. This is very similar to the approach used in Greenshtein and Park (2009).



(a) UnimodalConstant (b) BimodalConstant

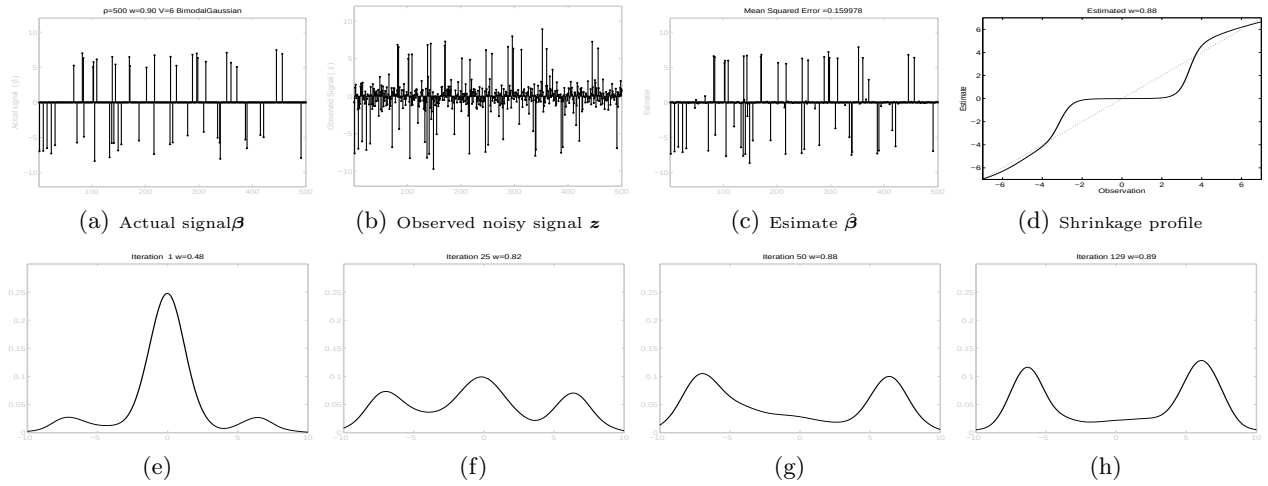
(c) UnimodalGaussian (d) BimodalGaussian

Figure 1: The different distributions used in our simulation setup for the non-zero part of the sequence.

The mean squared error $MSE = 1/p \sum_{i=1}^{p} (\beta_i - \hat{\beta}_i)^2$ is used to evaluate the accuracy of the estimation procedure. Note that the maximum likelihood estimator(MLE) has a MSE of one. Figure 3 plots the estimated $\hat{w}$ and the MSE as a function of the actual w–the fraction of zeros in the signal for different methods and different choices of the non-zero distribution [1]. The results are averaged over 100 repetitions. Figure 3(c) and (f) plots the estimated $\hat{w}$ and the MSE as a function of V–the signal strength for different methods and different sparsity w. The following observations can be made–

- It can be seen that the proposed nonparametric prior estimates w quite accurately and has the lowest MSE among all the competing methods.
- The prior is clearly mis-specified when we assume a normal or a Laplace and as a result there is a large bias in the estimation of the hyperparameter w–especially for less sparse signals. For less sparse signals the proposed method has much lower MSE than the estimators which use parametric priors.
- The nonparametric method adapts quite well to sparsity. The parametric priors show good performance only for very sparse signals.
- For all the methods as the sparsity of the signal increases the estimation of w is more accurate.
- For the parametric methods the normal prior does a better job at estimating w than the laplace prior and also has a lower MSE.
- The nonparametric method without mixing has a comparable MSE. By using a mixture prior the proposed procedure results in a much lower MSE.

[1] Due to lack of space we show results for two different choices of the nonparametric prior. Similar results are observed for different choices of the non-zero distribution.



(a) Actual signal $\boldsymbol{\beta}$ (b) Observed noisy signal $\boldsymbol{z}$ (c) Esimate $\hat{\boldsymbol{\beta}}$ (d) Shrinkage profile

(e) (f) (g) (h)

Figure 2: (a) A sample sequence $\boldsymbol{\beta}$ used in our simulation studies. The $p = 500$ length sequence has 5% ($w = 0.9$) of the values which are non-zero drawn from BimodalGaussian distribution with $V = 6$. (c) The observed noisy signal $\boldsymbol{z}$. (c) The estimate $\hat{\boldsymbol{\beta}}$ obtained by the proposed nonparametric method. (d) The shrinkage profile. (e),(f),(g), and (h) The marginal estimated during four different iterations of the EM algorithm

6 High dimensional classification

In a typical binary classification scenario we are given a training set $\mathcal{D} = \{(\boldsymbol{x}_j, y_j)\}_{j=1}^n$ containing n instances, where $\boldsymbol{x}_j \in \mathbf{R}^p$ is an instance and $y_j \in \mathcal{Y} = \{0, 1\}$ is the corresponding known label. The task is to learn a *classification function* $\delta : \mathbf{R}^p \to \mathcal{Y}$, which minimizes the error on the training set and generalizes well on unseen data. Very often it is convenient to learn a *discriminant function* $f : \mathbf{R}^p \to R$. The classification function can then be written as $\delta(\boldsymbol{x}) = \mathbb{I}(f(\boldsymbol{x}) > \theta)$, where $\mathbb{I}$ is the indicator function and θ determines the operating point of the classifier. The Receiver Operating Characteristic (ROC) curve is obtained as θ is swept from $-\infty$ to ∞.

Let f_0 and f_1 be the class-conditional densities of $\boldsymbol{x}$ in class 0 and 1 respectively, *i.e*, $f_0(\boldsymbol{x}) = \Pr[\boldsymbol{x}|y = 0]$ and $f_1(\boldsymbol{x}) = \Pr[\boldsymbol{x}|y = 1]$. Also let π_0 and π_1 be the prior probability of class 0 and 1 respectively. If f_0, f_1, π_0, and π_1 are known then the optimal classifier is the Bayes rule given by $\delta(\boldsymbol{x}) = \mathbb{I}\left(\log f_1(\boldsymbol{x})/f_0(\boldsymbol{x}) > \log \pi_0/\pi_1\right)$. Linear Discriminant Analysis (LDA)(also known as the Fisher's rule) assumes that f_0 and f_1 are p-variate normal distributions with means $\boldsymbol{\mu}_0$ and $\boldsymbol{\mu}_1$ respectively and the same covariance matrix $\boldsymbol{\Sigma}$. The corresponding discriminant function can be written as

$$f(\boldsymbol{x}) = \log \frac{f_1(\boldsymbol{x})}{f_0(\boldsymbol{x})} = (\boldsymbol{\mu}_1 - \boldsymbol{\mu}_0)^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu}), \quad (25)$$

where $\boldsymbol{\mu} = (\boldsymbol{\mu}_1 + \boldsymbol{\mu}_0)/2$, the overall mean of $\boldsymbol{x}$. For $p \gg n$ we cannot fit a full LDA model since the covariance matrix $\boldsymbol{\Sigma}$ will be singular–hence we need some sort of regularization. The simplest and one of the most effective form of regularization for high-dimensional data assumes that the features are independent within each class–the covariance matrix is diagonal, *i.e.*, $\boldsymbol{\Sigma} = \mathrm{diag}(\sigma_1^2, \ldots, \sigma_p^2)$. This is known as diagonal linear discriminant analysis (DLDA) (Dudoit et al., 2002) or the naive Bayes classifier. For a feature vector $\boldsymbol{x} = [x_1, \ldots, x_p]$ the corresponding discriminant function can be written as

$$f(\boldsymbol{x}) = \sum_{i=1}^{p} \beta_i \left(\frac{x_i - \mu_i}{\sigma_i} \right), \quad \beta_i = \frac{\mu_{1i} - \mu_{0i}}{\sigma_i}. \quad (26)$$

DLDA estimates $\boldsymbol{\beta}$ by plugging the empirical estimates for the population means $\boldsymbol{\mu}_1$ and $\boldsymbol{\mu}_0$, and the pooled estimate for the common population variance, *i.e.*,

$$\widehat{\beta}_i = \frac{\widehat{\mu}_{1i} - \hat{\mu}_{0i}}{\hat{\sigma}_i}. \quad (27)$$

In spite of strong independence assumptions, for high dimensional problems, DLDA works remarkably well in practice, often better than more sophisticated classifiers (Dudoit et al., 2002; Bickel and Levina, 2004).

In the high-dimensional setting when $p \gg n$ the plug-in estimates $\hat{\boldsymbol{\beta}}_j$ are generally over-inflated (quite larger than the true value $\boldsymbol{\beta}_i$) due to the limited sample size–thus leading to poor generalization performance on unseen data. We propose further regularization by using the proposed estimator to shrink the estimates. Note that $z_i = c\hat{\boldsymbol{\beta}}_i$ is approximately normal, *i.e.*, $\sim \mathcal{N}(c\beta_i, 1)$, where $c = \sqrt{n^+ n^-/n}$ and n^+ and n^- are the number of positive and negative examples respectively. Thus we can directly apply the proposed procedure directly on z_i and then rescale the estimates.

6.1 Microarray example

We will demonstrate our procedure on a prostrate cancer microarray dataset (Singh et al., 2002). The dataset consists of $p = 6033$ genes measured on $n = 102$ subjects (50 healthy controls and 52 prostate cancer patients). Given a new microarray measuring the same 6033 genes the task is to predict whether or not the subject develops prostate cancer.

Figure 4(a) shows the leave-one-patient-out cross-validated ROC curve and also the area under the ROC curve (AUC) for the DLDA classifier and DLDA with shrinkage based on parametric (normal/Laplace) and the proposed nonparametric priors. It can be seen that shrinkage estimators show a substantial improvement over the plug-in estimates used by DLDA. The nonparametric prior has a larger AUC than the parametric counterparts.

Figure 4(b) shows the profile of the different shrinkage estimators and also the estimated w–which can be interpreted as the fraction of irrelevant features/genes. It can be seen that all shrinkage estimators achieve effective feature selection by shrinking the right proportion of the estimates to zero. The nonparametric prior does a better job of adapting to the unknown sparsity by estimating w more accurately and selects the most sparse model.

7 Conclusions

In this work we propose a nonparametric discrete mixture prior which adapts well to the unknown sparsity in the signal. The key idea is to assume that there is a prior on the parameter but to impose no structural assumptions on that prior distribution and estimate it directly from the data. An iterative EM algorithm based on weighted non-parametric kernel density estimate was developed to estimate the sparsity in the signal. Our experiments show that the proposed prior adapts to sparsity much better than its parametric counterparts. We applied this procedure for the high-dimensional classification problems.

References

F. Abramovich, Y. Benjamini, D. L. Donoho, and I. M. Johnstone. Adapting to Unknown Sparsity by Controlling the False Discovery Rate. *The Annals of Statistics*, 34(2):584–653, 2006.

P. Bickel and E. Levina. Some theory for Fisher's linear discriminant function, 'naive Bayes', and some alternatives when there are many more variables than observations. *Bernoulli* , 10(6):989–1010, 2004.

C. M. Carvalho, N. G. Polson, and J. G. Scott. Handling sparsity via the horseshoe. In *In proceedings of the 12th International Conference on Artificial Intelligence and Statistics*, pages 73–80, 2009.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society: Series B*, 39(1):1–38, 1977.

S. Dudoit, J. Fridlyand, and T. P. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97(457):77–87, 2002.

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

E. Greenshtein and J. Park. Application of non parametric empirical bayes estimation to high dimensional classification. *Journal of Machine Learning Research*, 10:1687–1704, 2009.

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

W. James and C. Stein. Estimation with quadratic loss. In *Proc. 4th Berkeley Symp. Math. Statist. Prob.*, pages 361–379, 1961.

I. M. Johnstone and B. W. Silverman. Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32(4):1594–1649, 2004.

D. Singh, P. G. Febbo, K. Ross, D. G. Jackson, J. Manola, C. Ladd, P. Tamayo, A. A. Renshaw, A. V. D'Amico, J. P. Richie, E. S. Lander, M. Loda, P. W. Kanto, T. R. Golub, and W. R. Sellers. Gene expression correlates of clinical prostate cancer behavior. *Cancer Cell*, 1(2):203–209, 2002.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. Royal. Statist. Soc B*, 58(1):267–288, 1996.

R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences*, 99(10):6567–6572, 2002.

M. E. Tipping. Sparse bayesian learning and the relevance vector machine. *Journal of Machine Learning Research*, 1:211–244, 2001.

M. P. Wand and M. C. Jones. *Kernel Smoothing*. Chapman and Hall, 1995.

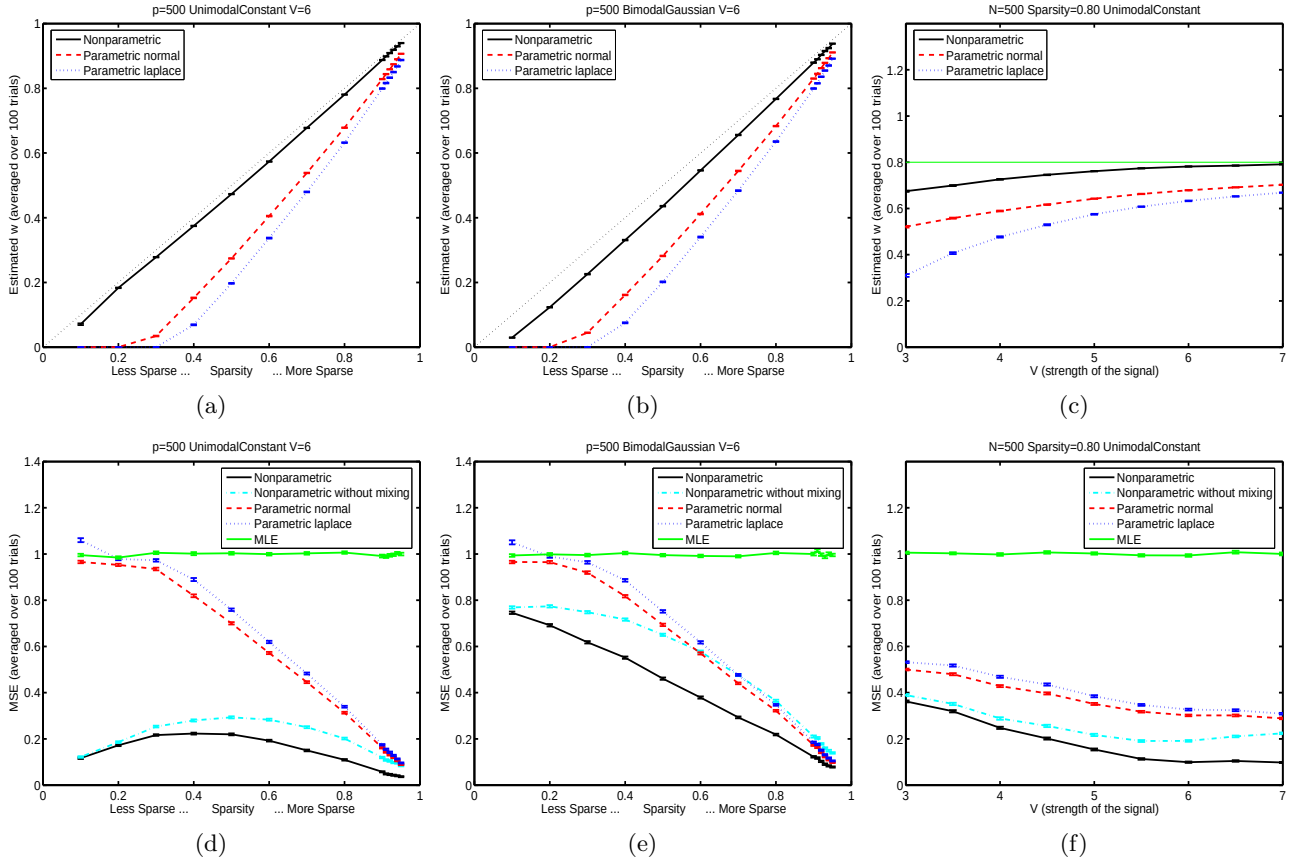


Figure 3: (a),(b),(d), and (e)–The estimated $\hat{w}$ and the mean squared error (MSE) as a function of sparsity–the fraction of zeros in the signal–for different methods and two different choices of the non-zero distribution (Unimodal constant and Bimodal Gaussian). A sequence of length $p = 500$ with signal strength parameter set to $V = 6$ was used. (c) and (e) The estimated $\hat{w}$ and the MSE as a function of V–the signal strength for different methods. All the results are averaged over 100 repetitions.

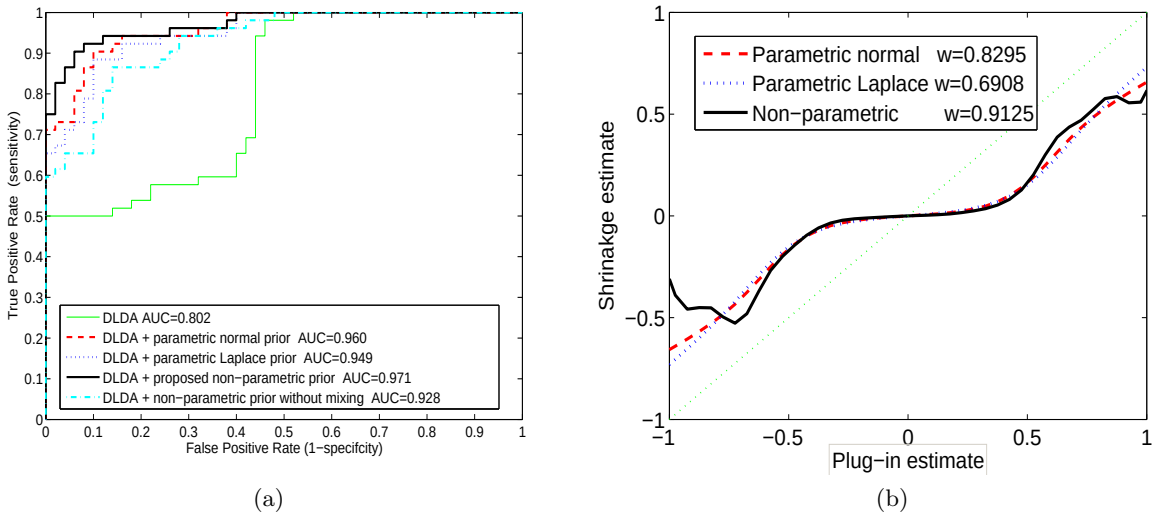


Figure 4: (a) The leave-one-patient-out cross-validated ROC curve for the prostrate cancer dataset for different classifiers. (b) The profile of the different shrinkage estimators and also the estimated w.