Nonparametric prior for adaptive sparsity

Vikas C. Raykar [1] and **Linda H. Zhao** [2]

[1] Siemens Healthcare, Malvern, PA 19355 USA
[2] University of Pennsylvania, Philadelphia, PA 19104, USA

May 14, 2010

The sparse normal mean problem

With adaptive sparsity

$\mathbf{x} = (x_1, x_2, \ldots, x_p)$ are p scalar observations satisfying

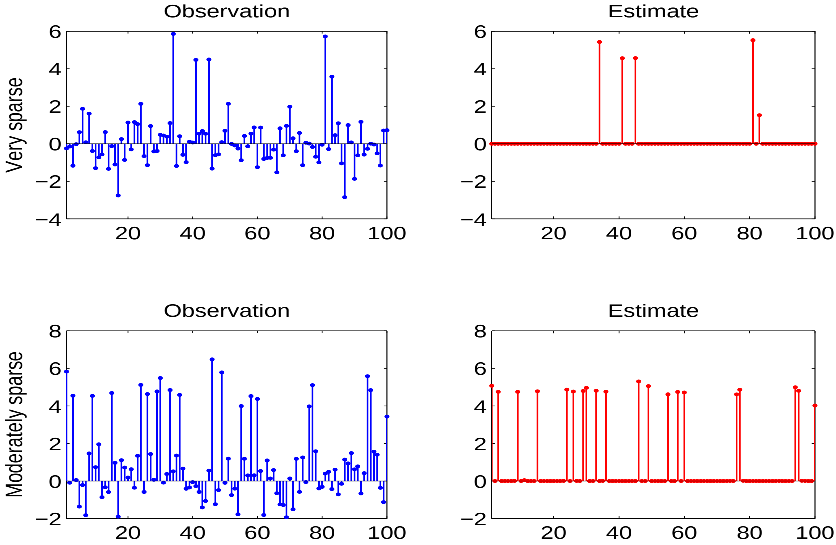
$$x_i = \mu_i + \epsilon_i,$$

where ϵ_i are independent and identically distributed as $\epsilon_i \sim \mathcal{N}(0, 1)$.

- Find a good estimate $\widehat{\boldsymbol{\mu}}$ of the unknown parameters $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_p)$.
- $\boldsymbol{\mu}$ could be *sparse*, *i.e.*, a large fraction of μ_i's are 0.
- However we do not know the amount of sparsity.
- The estimate should adapt to the sparsity.

Two examples with desired property

Estimator should adapt to the amount of sparsity



Applications

(1) Shrinkage and feature selection for high-dimensional classification.
(2) Multiple-hypothesis testing.
(3) Genomics and bio-informatics.
(4) Model selection in machine learning.
(5) Signal processing/Astronomical image processing.
(6) Wavelet smoothing.

Commonly used sparsity promoting priors

Parametric shrinkage priors

- Normal prior $\gamma_a(\mu_i) = (2\pi a^2)^{-1/2} \exp\left(-\mu_i^2/2a^2\right)$
- Laplace prior $\gamma_a(\mu_i) = 0.5a \exp\left(-a|\mu_i|\right)$
- Discrete mixture priors $w\delta(\mu_i) + (1-w)\gamma_a(\mu_i)$

The hyperparameter a (and w) controls the sparsity of the solution. Chosen by either

- Cross-validation.
- Evidence Maximization [Type II maximum-likelihood].

Type II maximum-likelihood

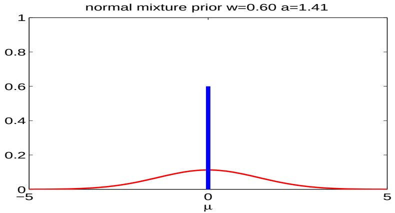
How well does the estimated hyperparameter adapt the sparsity ?
Depends on how misspecified the prior is.

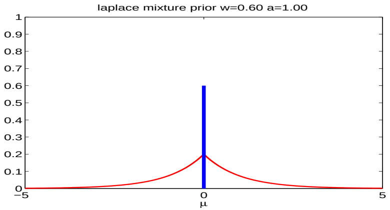
Discrete Mixture Prior

Mixture prior

$$p(\mu_i|w,\gamma) = w\delta(\mu_i) + (1-w)\gamma(\mu_i)$$

–$w \in [0,1]$ is the mixture parameter–proportion of $\{\mu_i = 0\}$.
–We consider w as a *hyperparameter*.
– γ is the non-zero part of the prior.
– For the nonzero part of the prior γ two commonly used parametric priors are normal and Laplace.





Non-parametric Mixture prior

Parametric priors are not very robust because of its 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 a completely unspecified density for the non-zero part of the mixture. The prior is completely nonparametric, i.e., there is no specific functional form.

$$p(\mu_i|w, \gamma) = w\delta(\mu_i) + (1 - w)\gamma(\mu_i)$$

(1) We do not specify any functional form for γ.

(2) We do not really need to specify any functional form for γ.

Posterior

If we know the hyperparameter w

Mixture prior

$$p(\mu_i|w,\gamma) = w\delta(\mu_i) + (1-w)\gamma(\mu_i)$$

Posterior

$$p(\mu_i|x_i,w,\gamma) = \tilde{p}_i\delta(\mu_i) + (1-\tilde{p}_i)G(\mu_i)$$

$$\tilde{p}_i = p(\mu_i = 0|x_i,w,\gamma) = \frac{w\mathcal{N}(x_i|0,1)}{w\mathcal{N}(x_i|0,1) + (1-w)g(x_i)}.$$

$$G(\mu_i) = p(\mu_i|x_i,w,\gamma,\mu_i \neq 0) = \mathcal{N}(\mu_i|x_i,1)\gamma(\mu_i)/g(x_i).$$

where

$$g(x_i) = \int \mathcal{N}(\mu_i|x_i,1)\gamma(\mu_i)d\mu_i$$

Note that g is the marginal density of the observations corresponding to those $\{\mu_i \neq 0\}$.

Posterior mean

We will use the mean of the posterior as our point estimate for $\boldsymbol{\mu}$.

So if we know the
(1) mixing parameter w
(2) the marginal g and its derivative g'
then the proposed estimate for μ is given by

$$\hat{\mu}_i = (1 - \tilde{p}_i)\left[x_i + \frac{g'(x_i)}{g(x_i)}\right]$$

where

$$\tilde{p}_i = \frac{w\mathcal{N}(x_i|0,1)}{w\mathcal{N}(x_i|0,1) + (1-w)g(x_i)} \quad g(x_i) = \int \mathcal{N}(\mu_i|x_i,1)\gamma(\mu_i)d\mu_i$$

The hyperparameter w can be estimated by maximizing the marginal likelihood and the posterior mean is then computed by plugging in the estimated $\hat{w}$.

But what about g ?

So what about g?

$$g(x_i) = \int \mathcal{N}(\mu_i | x_i, 1)\gamma(\mu_i) d\mu_i.$$

For example if we use the normal prior $\gamma(\mu_i) = \mathcal{N}(\mu_i|0, a^2)$ then $g(x_i) = \mathcal{N}(x_i|0, 1 + a^2)$. Hence

$$\hat{\mu}_i = (1 - \tilde{p}_i)\frac{a^2}{1 + a^2} x_i.$$

Both w and a are considered as hyper-parameters and we can estimate them by maximizing the marginal likelihood. We could also use a Laplace prior instead of the normal.

A crucial property of our method is that we avoid selecting a specific prior family for the nonzero part of the mixture prior.
Note that all we need is $g(x_i)$ and not $\gamma(\mu_i)$.

Estimating w–the fraction of zeros

Type II maximum likelihood

w is estimated by maximizing the log-marginal likelihood.

$$\widehat{w} = \arg\max_{w} \log m(\mathbf{x}|w).$$

The log-marginal can be written as

$$\log m(\mathbf{x}|w,\gamma) = \sum_{i=1}^{n} \log\left[w\mathcal{N}(x_i|0,1) + (1-w)g(x_i)\right]$$

But to estimate w we need to know $g(x_i)$

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

Estimating g–marginal of the non-zero part

Kernel density estimate

Non-parametric kernel density estimate

$$\hat{g}(x) = \frac{1}{\tilde{p}h} \sum_{j=1}^{p} (1 - \delta_j) K\left(\frac{x - x_j}{h}\right)$$

where

- $\delta_j = 1$ if $\mu_j = 0$ and zero otherwise.
- K is the *kernel*.
- h is the *bandwidth* of the kernel.
- $\tilde{p} = \sum_{j=1}^{p} (1 - \delta_j)$.

But we do not know δ_j.

Estimating both w and g simultaneously

EM algorithm [See paper for more details]

- Compute p_i using the current estimate $\hat{w}$ and $g_e(x_i)$ as follows

$$p_i = \frac{\hat{w}\mathcal{N}(x_i|0,1)}{\hat{w}\mathcal{N}(x_i|0,1) + (1-\hat{w})g_e(x_i)}.$$

- Re-estimate $\hat{w}$ and $\hat{g}(z_i)$ using the current estimate of p_i as follows

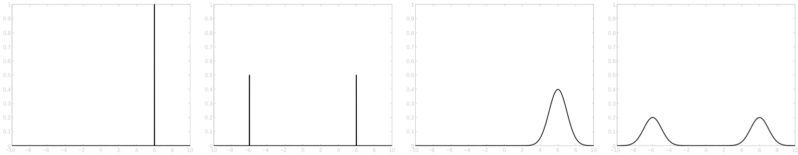
$$\hat{w} = \frac{1}{p}\sum_{i=1}^{p} p_i.$$

$$g_e(x_i) = \frac{1}{\tilde{p}h}\sum_{j=1}^{p}(1-p_j)K\left(\frac{x_i - x_j}{h}\right).$$

Simulations

Setup

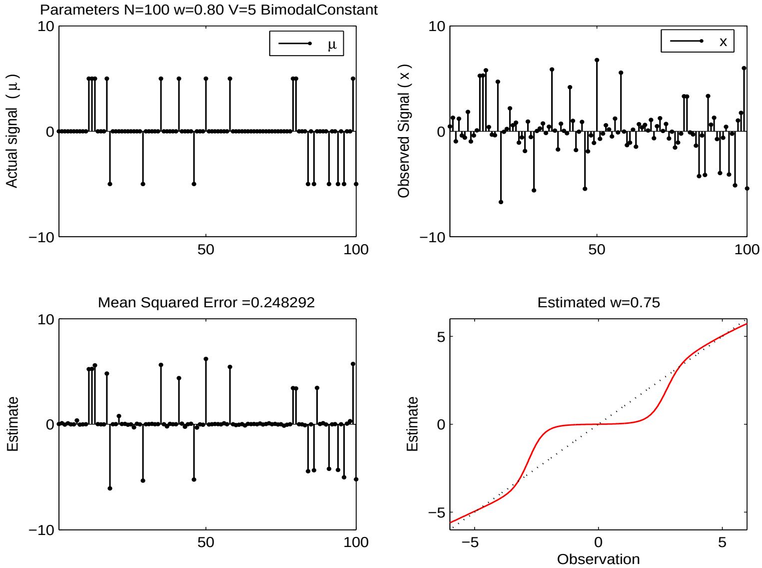
- A sequence $\boldsymbol{\mu}$ of length $p = 500$ is generated with different degree of sparsity and non-zero distribution.
  - w–sparsity parameter, the fraction of zeros in the sequence.
  - V controls the strength of the non-zero part.
  - The non-zero μ's are sampled from different distributions.
  - The observation x_i is generated from $\mathcal{N}(\mu_i, 1)$.



(a) UnimodalConstant (b) BimodalConstant (c) UnimodalGaussian (d) BimodalGaussian

Sample Results

Moderately sparse signal



Simulations

Methods compared

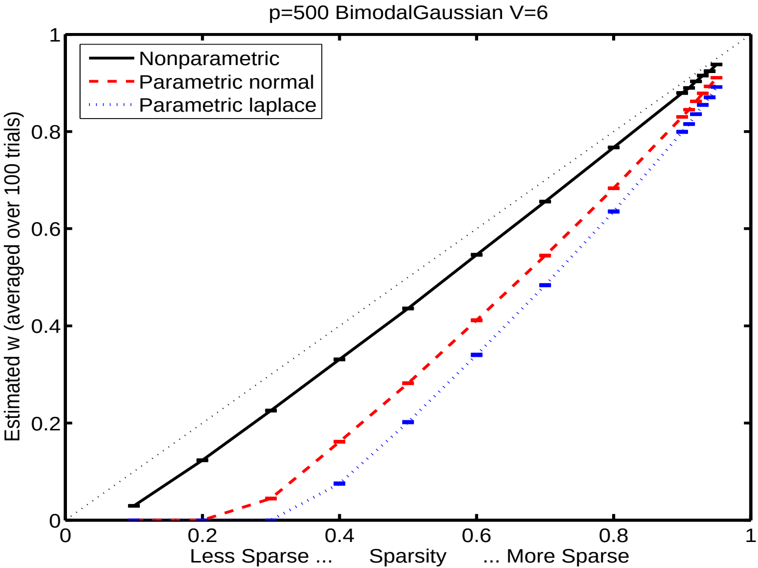
Mixture prior

$$p(\mu_i|w,\gamma) = w\delta(\mu_i) + (1-w)\gamma(\mu_i)$$

1. Non-parametric [Proposed] γ is unspecified.
2. Parametric normal [Johnstone and Silverman 2005] γ is a normal density.
3. Parametric laplace [Johnstone and Silverman 2005] γ is a Laplace density.
4. Non-parametric without mixing [similar to Eitan and Brown 2008] γ is unspecified but no mixing, *i.e.*, $w = 0$. In this case w cannot be estimated.

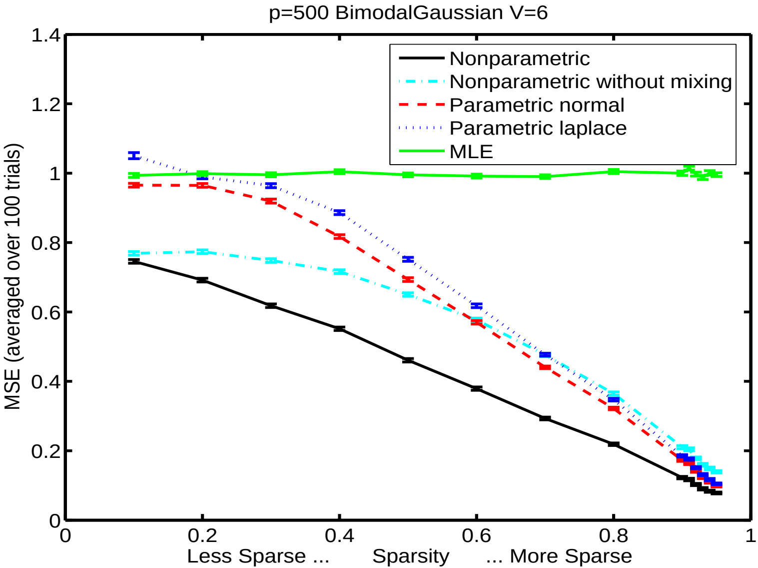
Simulation Results

Estimated $\hat{w}$



Simulation Results

Mean squared error

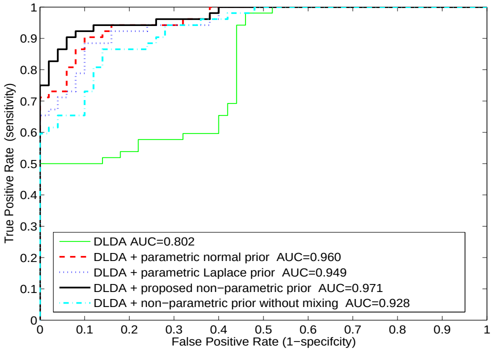


High dimensional classification

Diagonal Linear Discriminant analysis

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

Use the proposed procedure to shrink β_i.



Conclusions

(1) Non-parametric mixture prior

$$p(\mu_i|w,\gamma) = w\delta(\mu_i) + (1-w)\gamma(\mu_i)$$

(2) We impose no structural form on γ.
(2) Adaptive sparsity.
(3) Iterative EM algorithm to estimate w.
(4) Estimate of w is more accurate and the MSE much lower than parametric mixture priors.