A Nonparametric Bayesian Approach for Sparse Sequence Estimation

Ouyang Yunbo, Feng Liang
Department of Statistics
University of Illinois at Urbana-Champaign
Champaign, IL, USA
youyang4@illinois.edu

April 23, 2018

Abstract

A novel Bayes approach is proposed for the problem of estimating a sparse sequence based on Gaussian random variables. We adopt the popular two-group prior with one component being a point mass at zero, and the other component being a mixture of Gaussian distributions. Although the Gaussian family has been shown to be suboptimal for this problem, we find that Gaussian mixtures, with a proper choice on the means and mixing weights, have the desired asymptotic behavior, e.g., the corresponding posterior concentrates on balls with the desired minimax rate. To achieve computation efficiency, we propose to obtain the posterior distribution using a deterministic variational algorithm. Empirical studies on several benchmark data sets demonstrate the superior performance of the proposed algorithm compared to other alternatives.

Keywords: Empirical Bayes; Variational; Minimax; Posterior Consistency.

1 Introduction

Consider a Gaussian sequence model

\[ X_i = \theta_i + e_i, \quad i = 1, 2, \ldots, n, \]  

where \( \theta = (\theta_1, \ldots, \theta_n) \) is the unknown mean parameter, often assumed to be sparse, and \( e_i \)'s are independent errors following a standard normal distribution. The primary interest is to reconstruct the sparse vector \( \theta \) based on the data \( X^n = (X_1, X_2, \ldots, X_n) \). Such a simple model arises in many applications, such as astronomy, signal processing and bioinformatics.
It is also a canonical model for many modern statistical methods, such as nonparametric function estimation, large-scale variable selection and hypothesis testing.

There are different ways to define the sparsity of a vector. In this paper, we focus on the common definition of sparsity, i.e., many elements of $\theta_i$’s are zero. In particular, we assume $\theta$ is from the class of nearly black vectors,

$$\Theta_0(s_n) = \{ \theta \in \mathbb{R}^n : \#\{1 \leq i \leq n : \theta_i \neq 0\} = s_n \}.$$  

The parameter $s_n$ measures the sparsity of $\theta$, which is usually assumed to be $o(n)$, but unknown. So a desired feature of any reconstruction procedure is to adapt to the unknown sparsity level.

Since $\theta$ is known to be sparse, a natural approach is to threshold $X^n$. The thresholding rule can be chosen by the principle of minimizing the empirical fitting error with penalization (Golubev, 2002) or by minimizing the False Discover Rate (Abramovich et al., 2006). In addition, a plethora of research on variable selection and prediction in the context of high dimensional linear regression models can also be applied on this problem (Fan and Li, 2010; Bühlmann and Van De Geer, 2011; Zhang and Huang, 2008).

Some thresholding procedures are motivated from a Bayesian aspect, such as George and Foster (2000), Johnstone and Silverman (2004), and Castillo et al. (2012). The idea is to model $\theta_i$’s with a two-group structure prior, where one component is a point mass at zero due to the sparsity assumption and the other component is a continuous distribution with a smooth symmetric density function $g$, namely,

$$\pi(\theta_i) = w\delta_0(\theta_i) + (1-w)g(\theta_i).$$  \hspace{1cm} (2)

Theoretical studies have indicated that $g(\cdot)$, the prior density function on the non-zero component, should have a heavy tail for the posterior distribution to have the desired asymptotic behavior. Therefore, the double exponential distribution and its scaled variants are recommended for the choice of $g$, but not Gaussian distributions. The problem with distributions like the Gaussians which have lighter tails is that they tend to over-shrink $X_i$’s for large $\theta_i$’s therefore attain a lower posterior contraction rate (Johnstone and Silverman, 2004; Castillo et al., 2012).

In this paper, we propose an Empirical Bayes approach to this problem. Despite the warning message on Gaussian distributions, we still use Gaussians in our prior specification, which takes the following form

$$\pi(\theta_i) = w\delta_0(\theta_i) + (1-w) \sum_{t=1}^{T} w_t N(m_t, \sigma^2),$$  \hspace{1cm} (3)
then use data to learn weights $w, w_1, \cdots, w_T$ and centers $m_1, \cdots, m_T$. Our prior choice can be viewed as a special case of the two-group prior (2) with $g$ being a mixture of Gaussian density functions. The Gaussian prior is appealing due to its conjugacy, which simplifies our analysis and also enables tractable computation. As revealed by our asymptotic analysis, the suboptimal behavior of Gaussian distributions as mentioned in Johnstone and Silverman (2004) and Castillo et al. (2012) is on Gaussian distributions with mean zero, which can be avoided by a proper choice on weights $w_1, \cdots, w_T$ and centers $m_1, \cdots, m_T$.

The remainder of this paper is organized as follows. We present our main results in Section 2. We show that with some mild conditions on $m_i$’s and $w_i$’s, the corresponding posterior will have desired asymptotic behavior: it concentrates around the true parameter $\theta^*$ at the minimax rate, its effective dimension adapts to the unknown sparsity $s_n$, and the corresponding the posterior mean is asymptotic minimax. A variational implementation of our approach, as well as empirical studies, is presented in Section 3. All the proofs are given at the end after the conclusion.

We will start the remaining of this paper by briefly discussing some related work.

- This paper is motivated by a recent work by Martin and Walker (2014), where $g(\theta) = g_i(\theta)$ in (2) is set to be a Gaussian distribution centered exactly at the data point $X_i$. In Martin and Walker (2014), the resulting estimate of $\theta_i$ is still a shrinkage estimate toward zero, while in our approach, the estimate of $\theta_i$ is adaptively shrunk toward the mean of nearby data points. This explains why the empirical performance of our approach is better than the one from Martin and Walker (2014).

- Estimating $\theta$ for the Gaussian sequence model (1) can be also viewed as a compound decision problem (Robbins, 1951), where $\theta_i$’s are assumed to be i.i.d. random variables from a common but unknown distribution $G$. The aforementioned two-group prior (2) can be viewed as a special form of $G$, but in general $G$ can take any form. A number of nonparametric approaches have been proposed to estimate $G$ and then in return provide an estimate for $\theta$, such as the Maximum likelihood approach by Jiang and Zhang (2009), the nonparametric empirical Bayes approach by Brown and Greenshtein (2009), and a convex optimization based approach by Koenker (2014). In particular, Gaussian mixtures are used to estimate $G$ by Jiang and Zhang (2009) and Brown and Greenshtein (2009). However, their asymptotic results do not cover the nearly black class $\Theta_0(s_n)$. In their simulation study, Jiang and Zhang (2009) indeed consider the sparse situation and suggest to add a Gaussian component with
mean zero. In our asymptotic study, we have found that to achieve the desired asymptotic properties, a point mass zero, instead of a Gaussian with mean zero, seems necessary for the nearly black class.

2 Main Results

First we introduce some key notations. Data vector is denoted as $X^n = (X_1, X_2, \cdots, X_n)$; $\Pi_{X^n}(\cdot)$ is used to denote the prior indicating dependence with data; $\theta^* = (\theta^*_1, \cdots, \theta^*_n)$ is the unknown true mean of $X^n$; $p_\theta(X^n)$ is the likelihood function of $\theta$ given $X^n$. $a_n \succeq b_n$ is used to represent $\lim_{n} b_n/a_n = 0$.

Consider the following hierarchical prior $\Pi_{X^n}(\cdot \mid w_{1:T}, m_{1:T}, \sigma^2, \alpha)$ on $\theta$ as follows

$$w \sim \text{Beta}(\alpha n, 1), \quad (4)$$

$$\theta_i \mid w \sim w \delta_0 + (1-w) \sum_{t=1}^{T} w_t N(m_t, \sigma^2), i = 1, \cdots, n \quad (5)$$

where $\sigma^2, \alpha, w_{1:T}$ and $m_{1:T}$ are fixed parameters. $T$ is independent of the dimension $n$.

The posterior distribution is proportional to $p_\theta(X^n)\Pi_{X^n}(\theta)$. Strong assumptions are usually needed to prove Bayesian consistency. However, following [Walker and Hjort 2001] and [Martin and Walker 2014], we can obtain our posterior distribution with fractional likelihood, which can weaken the conditions required for Bayesian consistency. Given a fractional parameter $\kappa$, the corresponding posterior distribution is then proportional to $p_\kappa^*(X^n)\Pi_{X^n}(\theta)$. In implementation, $\kappa$ is set to be a large number close to 1 to capture most information from data. In all of the simulation studies in Section 3, we set $\kappa = 0.99$. In particular, the posterior measure of a Borel set $A \subset \mathbb{R}^n$ involving $\kappa$, denoted by $Q_n(A)$, can be expressed as

$$Q_n(A) = \frac{\int_A \{p_\kappa(X^n)/p_\kappa^*(X^n)\}^\kappa \Pi_{X^n}(d\theta)}{\int_{\mathbb{R}^n} \{p_\kappa(X^n)/p_\kappa^*(X^n)\}^\kappa \Pi_{X^n}(d\theta)}. \quad (6)$$

$\theta^*$ is assumed to be sparse. Denote $S^*$ as the support of $\theta^*$ and $s_n = |S^*|$ is the cardinality. Minimax rate established in [Donoho et al. 1992] for estimating a vector in $\Theta_0(s_n)$ is $\varepsilon_n = s_n \log(n/s_n)$. The aim is to prove posterior mean estimator based on the above prior attains asymptotically minimax $L^2$ error rate in $\Theta_0(s_n)$. To accomplish this, first we prove our posterior measure $Q_n$ has a desired concentration rate and then prove MSE is bounded.

The key conditions are imposed on $w_1, \cdots, w_T$ and $m_1, \cdots, m_T$. To find suitable $m_1, \cdots, m_T$, we apply suitable clustering algorithms on nonzero mean set $\theta_{S^n} = \{\theta^*_i \mid i \in S^*\}$
and retrieve cluster centers as \( m_1, \cdots, m_T \). Condition 1 essentially puts an upper bound on within-cluster sum of squares of \( \theta_{S^*} \) if we apply some clustering algorithm. Denote \( m_{t_i} \) as the corresponding cluster center of \( \theta^*_i \in \theta_{S^*} \), we assume

**Condition 1.** \( \sum_{i \in S^*} (\theta^*_i - m_{t_i})^2 = o(\varepsilon_n) \).

Since the summation in the left hand side has \( s_n \) terms, if \( \max_{i \in S^*} |\theta^*_i - m_{t_i}| = o(\log(n/s_n)) \), that is, the distance between \( \theta^*_i \) and the corresponding cluster centers grows slower than \( \log(n/s_n) \), our condition could be satisfied. Therefore Condition 1 is not strict since \( s_n = o(n) \). In implementation to specify \( m_1, \cdots, m_T \), since we do not know \( \theta_{S^*} \), we use a novel clustering algorithm to identify the cluster centered at 0 and estimate \( T \) nonzero cluster centers. The details will be provided in Section 3.

The second condition is on the weight \( w_1, w_2, \cdots, w_T \):

**Condition 2.** \( \min_{1 \leq t \leq T} w_t \geq Cs_n/n > 0 \).

Since \( T \) is independent of \( n \). This condition could be easily satisfied if we simply set \( w_t = 1/T \) for each \( t \). If each cluster size is bounded below, we could also plug in cluster weights.

Given the above 2 conditions, first we introduce Lemma 1, an analogue of Lemma 1 in Martin and Walker (2014).

**Lemma 1.** Let \( D_n \) be the denominator of posterior measure \( Q_n \). If Condition 1 and Condition 2 hold, then \( D_n > \frac{\alpha}{1+\alpha} \exp\{-2\varepsilon_n - o(\varepsilon_n)\} \) with \( P_{\theta^*} \)-probability 1.

After establishing a lower bound for the denominator of \( Q_n \), in order to prove posterior concentration, we’ll establish an upper bound for that numerator of \( Q_n \) measuring the complement of a ball centered at \( \theta^* \). Specifically, we show that posterior probability measure \( Q_n \) concentrates asymptotically on a ball centered at the truth \( \theta^* \) with square radius proportional to \( \varepsilon_n \), namely,

\[
A_{M\varepsilon_n} = \{ \theta \in \mathbb{R}^n : \|\theta - \theta^*\|^2 > M\varepsilon_n \}.
\]

**Theorem 1.** If Condition 1 and Condition 2 hold, then there exists \( M > 0 \) such that \( Q_n(A_{M\varepsilon_n}) \to 0 \) with \( P_{\theta^*} \)-probability 1 as \( n, s_n \to \infty \) with \( s_n = o(n) \).

Theorem 1 implies that our posterior distribution concentrates around the right place at the right rate, so it ought to produce an estimator of \( \theta \) with good properties. Next we show that the posterior mean \( \hat{\theta} \) is a minimax estimator.
Theorem 2. If Condition 1 and Condition 2 hold, there exists a universal constant $M' > 0$, such that $\mathbb{E}_{\theta^*} \|\hat{\theta} - \theta^*\|^2 \leq M' \varepsilon_n$ for all large $n$.

The proof is the same as the one by [Martin and Walker (2014)](#), provided that we have proved Lemma 1 and Theorem 1.

As posterior concentrates around $\theta^*$ at the minimax rate, we could conclude the majority of the posterior mass concentrates on $s_n$-dimensional subspaces of $\mathbb{R}^n$. The sparsity level of posterior is measured by the posterior distribution of $w$. Theorem 3 shows posterior distribution of $w$ put large probability mass above $1 - s_n n^{-1}$.

Theorem 3. If Condition 1 and Condition 2 hold, let $\delta_n = K \varepsilon_n n^{-1}$, and $K > 0$ is a suitably large constant. Then $\mathbb{E}_{\theta^*} \{ P(1 - w > \delta_n | X^n) \} \to 0$ as $n \to \infty$.

Notice that $\delta_n = K s_n / n \log(n/s_n) \approx K s_n / n$, therefore the posterior distribution of $w$ concentrates around $1 - s_n n^{-1}$. That is, the effective dimension of our posterior distribution adapts to the unknown sparsity level $s_n$. Our posterior mean estimator could detect true sparsity pattern with large probability. The proof is the same as the one by [Martin and Walker (2014)](#) as long as we have proved Lemma 1 and Theorem 1.

### 3 Implementation

In Section 2 we have shown using good prior with parameters $(w_{1:T}, m_{1:T}, \sigma^2, \alpha)$ could result in posterior concentration and minimax rate of Bayes posterior mean estimator. In implementation, we first specify $\sigma^2$ and $\alpha$, then estimate $(w_{1:T}, m_{1:T})$ using $X^n$. In Condition 1 $\sum_{i \in S^*} (\theta^*_i - m_{ti})^2$ needs to be bounded, however, in practice $\theta_{S^*}$ is unknown so that we cannot apply clustering algorithms on it. One remedy is to apply the clustering algorithm on $X^n$ since $\sum_{i \in S^*} (\theta^*_i - m_{ti})^2 \leq 2 \sum_{i \in S^*} (X_i - m_{ti})^2 + 2 \sum_{i \in S^*} (X_i - \theta^*_i)^2$. $\sum_{i \in S^*} (X_i - \theta^*_i)^2 = o_p(\varepsilon_n)$, therefore $\sum_{i \in S^*} (\theta^*_i - m_{ti})^2 = o(\varepsilon_n)$ if $\sum_{i \in S^*} (X_i - \theta^*_i)^2 = o_p(\varepsilon_n)$.

We apply some proper clustering algorithm on $X^n$ to assign clustering centers to $m_{1:T}$ and cluster weights to $w_{1:T}$. We need to pre-specify 0 as one clustering center so that the major task is to estimate the nonzero cluster centers. In Bayesian Framework, Dirichlet process mixture model is used to do clustering. We build a Dirichlet process (DP) mixture model on $X^n$ and estimate $m_{ti}$ by plugging in corresponding cluster centers. The general framework is summarized as follows:

$$\theta_i \sim G, G \sim \text{DP}(\alpha_0, G_0);$$
$$X_i \sim \text{N}(\theta_i, 1), i = 1, 2, \cdots, n;$$

$$\delta_n = K \varepsilon_n n^{-1}, K > 0;$$

$$\sum_{i \in S^*} (\theta^*_i - m_{ti})^2 \leq 2 \sum_{i \in S^*} (X_i - m_{ti})^2 + 2 \sum_{i \in S^*} (X_i - \theta^*_i)^2.$$
where \( G_0 \) is the base measure and \( \alpha_0 \) is the concentration measure. Given \( G_0 \) and \( \alpha_0 \), we have stick breaking representation of \( G \) as \( \sum_{t=1}^{\infty} \pi_t\delta_{\eta_t} (\cdot) \), where \( \eta_t \) is drawn independently and identically distributed (i.i.d.) from base measure \( G_0 \), while \( \pi_t = V_t \prod_{l=1}^{t-1} (1 - V_l) \). \( V_t \) is drawn i.i.d. from Beta(1, \( \alpha \)). From this formulation we could see that random distribution function \( G \) is almost surely discrete. Since \( \theta^* \) is a sparse vector, in order to get a random distribution \( G \) with a positive point mass at 0, \( G_0 \) ought to have a positive mass at 0. Therefore we model \( G_0 \) as a normal component with a point mass at 0:

\[
G_0 = w_0 \delta_0 + (1 - w_0)N(0, \sigma^2_0);
\]

where \( w_0 \) and \( \sigma^2_0 \) are 2 pre-specified parameters. Then \( G \) could be written as

\[
G = \sum_{t=1}^{\infty} \pi_t (\xi_t \delta_0 (\cdot) + (1 - \xi_t) \delta_{\eta_t^*} (\cdot)) = \sum_{t=1}^{\infty} \pi_t \xi_t \delta_0 (\cdot) + \sum_{t=1}^{\infty} \pi_t (1 - \xi_t) \delta_{\eta_t^*} (\cdot)
\]

\[
\equiv w \delta_0 (\cdot) + (1 - w) \sum_{t=1}^{\infty} w_t \delta_{\eta_t^*} (\cdot),
\]

where \( w = \sum_{t=1}^{\infty} \pi_t \xi_t \) and \( w_t = \frac{\pi_t (1 - \xi_t)}{1 - w} \). In this formulation \( \xi_t \) is drawn i.i.d. from Ber(\( w \)). If \( \xi_t = 0 \), then \( \eta_t^* \) is drawn from \( N(0, \sigma^2_0) \), otherwise \( \eta_t^* = 0 \). Since \( w > 0 \), \( G \) always has a positive probability mass at 0 which could induce certain level of sparsity. In the following subsection, we’ll develop a variational Algorithm to estimate \( G \).

### 3.1 Variational Algorithm for Estimating Prior

Once we have specified Dirichlet Process as the prior, we need to compute the posterior distribution of \( G \) in order to calculate Maximum A Posteriori (MAP) estimator of \( G \). The major challenge here is that we have the infinite sum in the stick breaking form of \( G \), which makes it impossible to sample from \( G \). One remedy here is to fix \( T \) as the upper bound of the number of clusters (Blei et al., 2006). Then we have the following truncated version of
stick breaking process using \( G_0 = w_0\delta_0 + (1 - w_0)N(0, \sigma_0^2) \) as the base measure.

\[
V_i |\alpha_0 \sim \text{Beta}(1, \alpha_0), t = 1, 2, \cdots, T - 1, V_T = 1; \quad (7)
\]

\[
\xi_t \sim \text{Ber}(w_0), t = 1, 2, \cdots, T; \quad (8)
\]

\[
\eta^*_t |\xi_t \sim \begin{cases} 
\delta_0 & \xi_t = 1 \\
N(0, \sigma_0^2) & \xi_t = 0
\end{cases}; t = 1, 2, \cdots, T; \quad (9)
\]

\[
\pi_t = V_t \prod_{j=1}^{t-1} (1 - V_j), t = 1, 2, \cdots, T - 1, \pi_T = \prod_{j=1}^{T} (1 - V_j); \quad (10)
\]

\[
Z_k |\{V_1, V_2, \cdots, V_{T-1}\} \sim \text{Multinomial}(\pi); \quad (11)
\]

\[
X_i |Z_i \sim N(\eta_{Z_i}^*, 1), i = 1, 2, \cdots, n. \quad (12)
\]

The observed data are \( y \) and the parameters are \( Z_1 \times n, V_1 \times (T-1), \eta^*_{1 \times T}, \xi_{1 \times T} \). \( \eta^* = (\eta_1^*, \cdots, \eta_T^*) \) contains all unique values of \( \eta = (\eta_{Z_i}^*)_{i=1}^n \).

We could use Markov Chain Monte Carlo (MCMC) (Escobar and West, 1995) to compute the posterior distribution of these parameters. However, due to large \( n \), the number of parameters is huge, making MCMC converging very slowly. In this paper, we use a variational algorithm to get posterior distribution. Variational Bayes methods are deterministic.

The essence of variational inference is to regard the computation of posterior distribution as an optimization problem. Solving this optimization problem gives an approximation to the posterior distribution. Blei et al. (2006) proposed variational algorithms for Dirichlet process mixture models for exponential family. Although normal distribution with a positive mass at 0 does not belong to exponential family, we could follow the same philosophy to deduce the corresponding algorithm.

We consider mean field variational inference and assume the following fully factorized variational distribution:

\[
q(Z, V, \eta, \xi) = q_{p,m,\tau}(\eta, \xi)q_{\gamma_1,\gamma_2}(V)q_{\Phi}(Z);
\]

Through the calculation shown in the Appendix, we could prove the optimal \( q \) must be further factorized as follows:

- \( q_{p,m,\tau}(\eta^*, \xi) = \prod_{t=1}^{T} q_{p_t, m_t, \tau_t}(\eta^*_t, \xi_t) \), where \( p = (p_1, p_2, \cdots, p_T), m = (m_1, m_2, \cdots, m_T), \tau = (\tau_1, \tau_2, \cdots, \tau_T) \), and \( q_{p_t, m_t, \tau_t}(\eta^*_t, \xi_t) = p_t \xi_t = 1 \delta_0 + (1 - p_t) \xi_t = 0 q_{m_t, \tau_t}(\eta^*_t) \), where \( q_{m_t, \tau_t}(\eta^*_t) \) is Normal density with mean \( m_t \) and variance \( \tau_t^2 \).

- \( q_{\gamma_1,\gamma_2}(V) = \prod_{t=1}^{T-1} q_{\gamma_{1t},\gamma_{2t}}(V_t) \), where \( \gamma_1 = (\gamma_{11}, \gamma_{12}, \cdots, \gamma_{1(T-1)}), \gamma_2 = (\gamma_{21}, \gamma_{22}, \cdots, \gamma_{2(T-1)}) \), \( q_{\gamma_{1t},\gamma_{2t}}(V_t) \) is Beta Distribution with parameters \( (\gamma_{1t}, \gamma_{2t}) \).
\[ q_\Phi(Z) = \prod_{i=1}^{n} q_{\Phi_i}(Z_i); \] where \( Z = (Z_1, Z_2, \cdots, Z_n) \), \( \Phi = (\phi_1, \phi_2, \cdots, \phi_n), \phi_i = (\phi_{i,1}, \phi_{i,2}, \cdots, \phi_{i,T}) \), \( \phi_{i,t} = q(Z_i = t) \), \( q_{\Phi_i}(Z_i) \) is Multinomial distribution with parameters \( \phi_i \).

The algorithm is summarized in Algorithm 1. Via iterating these steps we could update the variational parameters. After convergence of \( \Phi \), \( p \), \( m \), \( \tau \), \( \gamma_1 \) and \( \gamma_2 \), we get an approximation of the posterior by plugging in these estimated parameters. The parameters we are interested in are \( \Phi \), \( p \) and \( m \). (In the algorithm \( || \cdot ||_{\infty, \infty} \) means the element-wise maximum absolute value; \( \logit(x) = (1 + \exp(-x))^{-1} \).

**Algorithm 1** Variational Bayes Algorithm for Dirichlet process mixture model with \( G_0 \)

**Input:** \( y, \alpha_0, \sigma_0, w_0, T \)

**Initialize:** \( \Phi^{(1)} \) and \( \Phi^{(0)} \);

**While** \( ||\Phi^{(1)} - \Phi^{(0)}||_{\infty, \infty} > \epsilon \) do

\[ m_t \leftarrow \frac{\sigma_0^2}{\sigma_0^2 \sum_{i=1}^{n} \phi_{i,t}^{(0)} + 1}, t = 1, 2, \cdots, T; \]

\[ \tau_t^2 \leftarrow \frac{\sigma_0^2}{\sigma_0^2 \sum_{i=1}^{n} \phi_{i,t}^{(0)} + 1}, t = 1, 2, \cdots, T; \]

\[ p_t \leftarrow \logit^{-1}((\log(w_0) - \log(1 - w_0) + \log(\sigma_0^2 \cdot \sum_{i=1}^{n} \phi_{i,t}^{(0)} + 1))/2 - \frac{\sigma_0^2 \cdot (\sum_{i=1}^{n} \phi_{i,t}^{(0)} \cdot X_t)^2}{2(\sigma_0^2 \sigma_0^2 \sum_{i=1}^{n} \phi_{i,t}^{(0)} + 1)}), t = 1, 2, \cdots, T; \]

\[ \gamma_{t,1} \leftarrow 1 + \sum_{i=1}^{n} \phi_{i,t}^{(0)}, t = 1, 2, \cdots, T - 1; \]

\[ \gamma_{t,2} \leftarrow \alpha_0 + \sum_{i=1}^{n} \sum_{j=t+1}^{T} \phi_{i,j}^{(0)}, t = 1, 2, \cdots, T - 1; \]

\[ S_{i,t} \leftarrow E_q \log V_t + \sum_{i=1}^{t-1} E_q \log(1 - V_i) + (1 - p_t)m_t X_t - \frac{1}{2}(1 - p_t)(m_t^2 + \tau_t^2), t = 1, 2, \cdots, T, i = 1, 2, \cdots, n; \]

\[ \phi_{i,t}^{(1)} \propto \exp(S_{i,t}), t = 1, 2, \cdots, T, i = 1, 2, \cdots, n; \]

**End while**

**Output:** \( p, m, \Phi \)

3.2 Posterior Computation

Given approximate posterior estimates \( \hat{p}, \hat{m}, \hat{\Phi} \), we construct a MAP estimator of \( G \). Remind that \( \hat{m}_t \) is the nonzero cluster center; \( \hat{p}_t \) is the probability mass of zero of component indexed by \( t \); each entry \( \hat{\phi}_{it} \) of \( \hat{\Phi} \) is the posterior probability of \( Z_i \) belonging to the cluster \( t \). The approximate posterior distribution of \( \eta_{Z_i}^t \) is

\[
\sum_{t=1}^{T} \hat{\phi}_{it} \hat{p}_t \delta_0(\cdot) + \sum_{t=1}^{T} \hat{\phi}_{it} (1 - \hat{p}_t) \delta_{m_t}(\cdot).
\]
The most probable posterior assignment of $\eta^*_Z$, based on above posterior is denoted as $\hat{\eta}^*_Z$. MAP estimate of cluster weights including zero clusters is $\bar{w}_t = \#\{k : \hat{\eta}^*_Z = \hat{m}_t\}/n$ and $\bar{w}_0 = \#\{ k : \hat{\eta}^*_Z = 0\}/n$. Then the estimated prior is $$\hat{G}_{MAP} = \bar{w}_0 \cdot \delta_0(\cdot) + \sum_{t=1}^{T} \bar{w}_t \delta_{\hat{m}_t}(\cdot).$$

Plugging in $\hat{G}_{MAP}$, data generation process we are considering could be written as (N$^\kappa$ denotes normal likelihood to the power of $\kappa$):

$$\theta_i \sim \hat{G}_{MAP}, \hat{G}_{MAP} = \bar{w}_0 \cdot \delta_0(\cdot) + \sum_{t=1}^{T} \bar{w}_t \delta_{\hat{m}_t}(\cdot)$$

$$X_i \sim N^\kappa(\theta_i, 1), i = 1, 2, \ldots, n.$$ 

This is different from the original data generating process but could be regarded as an approximation of it by setting $\sigma^2 = 0$. The reason why we use this approximation is that instead of using MCMC, we could compute a closed-form posterior distribution of $\theta$ given $X^n$ using Bayes Formula: since $X_i \sim N(\theta_i, 1)$ and $\theta_i \sim \hat{G}_{MAP}$, the posterior distribution of $\theta_i$ given $X^n$ has the form

$$\frac{\bar{w}_0 \exp\left(-\frac{\kappa X_i^2}{2}\right) \delta_0 + \sum_{t=1}^{T} \bar{w}_t \exp\left(-\kappa (X_i - \hat{m}_t)^2/2\right) \delta_{\hat{m}_t}}{\bar{w}_0 \exp\left(-\frac{\kappa X_i^2}{2}\right) + \sum_{t=1}^{T} \bar{w}_t \exp\left(-\kappa (X_i - \hat{m}_t)^2/2\right)} \equiv \hat{w}_0 \cdot \delta_0(\cdot) + \sum_{t=1}^{T} \hat{w}_t \delta_{\hat{m}_t}(\cdot),$$

where $\hat{w}_t$ is the posterior weight and $\sum_{t=1}^{T} \hat{w}_t = 1$. The final estimate of $\theta_i$ is the posterior mean estimator

$$\hat{\theta}_i = \sum_{t=1}^{T} \hat{w}_t \hat{m}_t.$$ 

### 3.3 Simulation Studies

We conduct the following four simulation studies; R package for our empirical Bayes estimator is available in [https://github.com/yunboouyang/VBDP](https://github.com/yunboouyang/VBDP); all the source code is summarized in [https://github.com/yunboouyang/EBestimator](https://github.com/yunboouyang/EBestimator). For our empirical Bayes estimator, we set the maximal number of clusters to be $T = 10$. We set fractional likelihood parameter $\kappa = 0.99$ throughout all the simulation study. We set concentration parameter $\alpha_0 = 1$ in all of the simulation studies. For $G_0 = w_0 \delta_0 + (1 - w_0)N(0, \sigma^2_0)$, we set $w_0 = 0.01$. For simulation study 1, we set $\sigma_0 = 4$. For other simulation studies, we set $\sigma_0 = 6$. The influence of these hyper-parameters will diminish when sample size $n$ is large.
Besides our method (denoted as DP representing Dirichlet process clustering), we also consider estimators proposed by Martin and Walker (2014) (denoted as EBMW), by Koenker and Mizera (2014) (denoted as EBKM), and by Johnstone and Silverman (2004) (denoted as EBMed, we directly use functions from the EbayesThresh package by Johnstone and Silverman (2005)), as well as the Hard thresholding estimator, soft thresholding estimator, SURE estimator (using waveThresh package) and FDR estimator with parameters $q = 0.01, 0.1, 4$.

**Experiment 1**

In the first simulation study, we take sample $X^n$ of dimension $n = 200$ from the normal mean model $X_i \sim N(\theta_i, 1)$. In this case, we consider the number of nonzero elements to be $s_n = 10, 20, 40, 80$ and the signals are fixed at values $\mu_0 = 1, 3, 5, 7$. We compute mean squared error (MSE) and mean absolute error (MAE) based on 200 replications as two measure of performance. We summarize the results in Table 1 and Table 2.

| $s_n$ | 10 | 20 | 40 | 80 |
|-------|----|----|----|----|
|       | 1  | 3  | 5  | 7  |
| EBMW  | 10 | 54 | 35 | 40 |
| EBKM  | 12 | 35 | 21 | 19 |
| EBmed | 10 | 43 | 22 | 32 |
| SURE  | 14 | 42 | 43 | 42 |

| $\mu_0$ | 1  | 3  | 5  | 7  | 1  | 3  | 5  | 7  |
|---------|----|----|----|----|----|----|----|----|
| EBMW    | 10 | 54 | 35 | 40 | 69 | 53 | 32 | 44 |
| EBKM    | 12 | 35 | 21 | 19 | 7  | 32 | 25 | 44 |
| EBmed   | 10 | 43 | 22 | 19 | 7  | 32 | 25 | 44 |
| SURE    | 14 | 42 | 43 | 42 | 42 | 42 | 42 | 42 |

| Soft Thresholding | 10  | 11  | 11  | 11  | 10  | 24  | 23  | 22  | 20  |
| Hard Thresholding | 13  | 61  | 21  | 12  | 24  | 122 | 39  | 23  | 45  |
| FDR $q = 0.01$    | 10  | 75  | 25  | 11  | 20  | 143 | 40  | 22  | 41  |
| FDR $q = 0.1$     | 11  | 55  | 24  | 20  | 23  | 93  | 39  | 37  | 44  |
| FDR $q = 0.4$     | 22  | 63  | 53  | 51  | 36  | 94  | 82  | 80  | 64  |
| DP             | 11  | 37  | 11  | 3  | 19  | 50  | 17  | 4  | 33  |

Table 1: MSE of Simulation Study 1, error of the best method marked as bold
Table 2: MAE of Simulation Study 1, error of the best method marked as bold

When $\mu_0$ is larger than one, our DP method is the best one in terms of MSE since it is easier to figure out the cluster centers of $\theta_i$'s and estimate the nonzero $\theta_i$'s. Besides, when $s_n$ is large, our DP method has the best performance among all since the more nonzero $\theta_i$'s we have, the easier we could estimate the cluster centers. When both $\mu_0$ and $\theta_i$ are small, which is a tough case since high dimensional noise might make it much more difficult to detect weak signals, our method still has the comparable performance among all methods.

Our method has better performance when measured by MSE than that in MAE. Other methods such as EBKM and EBmed also have comparable good performance.

**Experiment 2**

In the second simulation study we increase the data dimension to 500. We are interested in the case when $s_n = 25, 50, 100$ and all nonzero elements are fixed at $\mu_0 = 3, 4, 5$. This is a more challenging case since the signal is not very strong. MSE and MAE are shown in Table 3 and Table 4 respectively.
In terms of MSE, our DP method is consistently the best one among all the different configurations. The performance of EBKM is comparable. One drawback of EBKM is that the computational cost is high since solving a high dimensional optimization problem is needed. However for DP, we use a computationally efficient variational inference algorithm, which dramatically saves the computational time.
Experiment 3

In the third simulation study we maintain all the features of the second experiment except that non-zero elements of the true mean vector are now generated from standard Gaussian distribution centered at the original values. All different $\theta_i$'s form a data cloud which is centered at the distinct non-zero values in simulation study 2.

| $s_n$ | 25 | 50 | 100 |
|-------|----|----|-----|
|       | 3  | 4  | 5   | 3  | 4  | 5   | 3  | 4  | 5   |
| EBMW  | 110| 91 | 62  | 186| 150| 103 | 308| 245| 176 |
| EBKM  | 77 | 68 | 49  | 121| 105| 77  | 185| 163| 124 |
| EBMed | 93 | 79 | 59  | 151| 130| 100 | 236| 212| 176 |
| SURE  | 96 | 104| 106 | 155| 165| 168 | 243| 257| 260 |
| ST    | 196| 269| 313 | 389| 534| 626 | 780| 1075|1251 |
| HT    | 135| 123| 79  | 263| 237| 155 | 527| 479| 311 |
| FDR $q = 0.01$ | 151| 129| 79  | 268| 216| 133 | 479| 368| 221 |
| FDR $q = 0.1$  | 112| 90 | 66  | 187| 148| 110 | 300| 235| 187 |
| FDR $q = 0.4$  | 138| 137| 129 | 216| 213| 201 | 320| 307| 301 |
| DP    | 76 | 68 | 53  | 120| 110| 84  | 189| 164| 123 |

Table 5: MSE of Simulation Study 3, error of the best method marked as bold

| $s_n$ | 25 | 50 | 100 |
|-------|----|----|-----|
|       | 3  | 4  | 5   | 3  | 4  | 5   | 3  | 4  | 5   |
| EBMW  | 51 | 46 | 38  | 92 | 82 | 69  | 164| 145| 126 |
| EBKM  | 75 | 63 | 52  | 108| 91 | 74  | 160| 135| 111 |
| EBMed | 44 | 38 | 31  | 79 | 70 | 59  | 147| 132| 115 |
| SURE  | 69 | 74 | 76  | 114| 122| 124 | 187| 195| 197 |
| ST    | 67 | 79 | 86  | 133| 158| 171 | 266| 317| 342 |
| HT    | 51 | 43 | 31  | 100| 84 | 61  | 200| 169| 121 |
| FDR $q = 0.01$ | 55| 44 | 31  | 101| 80 | 57  | 188| 144| 104 |
| FDR $q = 0.1$  | 45 | 37 | 31  | 81 | 67 | 58  | 143| 120| 108 |
| FDR $q = 0.4$  | 60 | 62 | 60  | 105| 107| 105 | 176| 175| 176 |
| DP    | 59 | 54 | 44  | 97 | 84 | 69  | 156| 127| 102 |

Table 6: MAE of Simulation Study 3, error of the best method marked as bold

Even though our DP method is not as good as EBKM and EBMed in some configurations, our DP method has quite similar performance even it’s more difficult to estimate the
clustering centers in Simulation Study 3 than Simulation Study 2.

**Experiment 4**

In the fourth example, we consider 1000-dimensional vector estimation, with the first 10 entries of $\theta^*$ equal 10, the next 90 entries equal $A$, and the remaining 900 entries equal 0. We consider a range of $A$, from $A = 2$ to $A = 7$. DP method will automatically estimate 3 cluster centers instead of 2. Average MSE and MAE are recorded in Table 7 and Table 8.

| $A$ | 2  | 3  | 4  | 5  | 6  | 7  |
|-----|----|----|----|----|----|----|
| EBMW| 320| 421| 291| 175| 134| 125|
| EBKM| 206| 223| 150| 79 | 46 | 35 |
| EBMed| 337| 350| 240| 173| 148| 138|
| SURE| 278| 327| 333| 337| 336| 334|
| ST  | 504| 894| 1259|1436|1475|1483|
| HT  | 375| 671| 610| 301| 134| 102|
| FDR $q = 0.01$ | 375| 633| 441| 199| 121| 111|
| FDR $q = 0.1$ | 373| 421| 259| 194| 185| 181|
| FDR $q = 0.4$ | 440| 451| 404| 399| 396|
| DP  | 204| 220| 161| 205| 151| 85 |

Table 7: MSE of Simulation 4, error of the best method is marked as bold

| $A$ | 2  | 3  | 4  | 5  | 6  | 7  |
|-----|----|----|----|----|----|----|
| EBMW| 179| 199| 159| 130| 122| 120|
| EBKM| 225| 182| 115| 79 | 61 | 61 |
| EBMed| 176| 161| 127| 110| 101| 97 |
| SURE| 209| 241| 245| 248| 246| 247|
| ST  | 216| 294| 347| 367| 371| 372|
| HT  | 189| 242| 184| 110| 85 | 80 |
| FDR $q = 0.01$ | 189| 232| 147| 96 | 85 | 83 |
| FDR $q = 0.1$ | 188| 168| 120| 110| 109| 108|
| FDR $q = 0.4$ | 217| 214| 208| 212| 210| 209|
| DP  | 215| 161| 91 | 89 | 75 | 57 |

Table 8: MAE of Simulation 4, error of the best method marked as bold
In some configurations, DP method is the best one among all the state-of-the-art methods. Although in some cases EBKM and EBMed performs better, MSE and MAE of our DP method are still comparable.

4 Conclusions and Discussions

Empirical Bayes method introduced in Martin and Walker (2014) does not have good performance in simulation study since a prior centered at $X_i$ may introduce noise in posterior estimation, whereas our nonparametric Bayes based clustering method will decrease the noise but capture the general pattern: the estimate of $\theta_i$ is adaptively shrunk toward the mean of nearby data points. Therefore our DP method outperforms the empirical Bayes estimator based on Martin and Walker (2014).

Theoretical results shows with proper choice of parameters in the prior, our posterior mean estimator achieves asymptotically minimax rate. In the implementation we propose a fast variational inference method which approximates posterior distribution, which is more efficient than Empirical Bayes method proposed by Koenker (2014) in terms of computation time. To our knowledge, this is a first work connecting high dimensional sparse vector estimation with clustering. Our nonparametric Bayesian estimator could be applied to high dimensional classification, feature selection, hypothesis testing and nonparametric function estimation. Possible extension in implementation includes using other well-studied clustering methods to estimate prior and comparing their performance.

Proofs

Proof for Lemma 1

Denote $w = (w_1, w_2, \ldots, w_T)$ and $m = (m_1, m_2, \ldots, m_T)$. Write $D_n$ in terms of the conditional prior $(\theta_1, \theta_2, \ldots, \theta_n)|w \sim \Pi_{w,w,m}$ and the marginal prior $w \sim \Pi$. Therefore

$$D_n = \int_0^1 \prod_{i=1}^n \int_{\mathbb{R}} \left\{ \frac{p_\theta(X_i)}{p^*_{\theta_i}(X_i)} \right\}^* \Pi_{w,w,m}(d\theta_i) \Pi(dw);$$

where $p_\theta(X_i) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(X_i-\theta)^2}{2}\right)$. For given $w$, the inner expectation involves an average over all configurations of the indicators $(1_{\theta_1=0}, 1_{\theta_2=0}, \ldots, 1_{\theta_n=0})$. This average is clearly larger than just the case where the indicators exactly match up with the support $S^*$ of $\theta^*$,
times the probability of that configuration, that is,

\[ D_n > \int_0^1 w^{n-s_n}(1-w)^{s_n} \Pi_i (d\nu) \prod_{i \in S} \int_{\mathbb{R}} e^{2\pi i (X_i - \theta_i)^2 -(X_i - \theta_i)^2} \sum_{t=1}^T \frac{w_t}{\sqrt{2\pi \sigma^2}} e^{-\frac{i}{2\sigma^2} (m_t - \theta_i)^2} d\theta_i. \]

For each \( i \in S^* \) and each \( 1 \leq t \leq T \),

\[
\begin{align*}
\int_{\mathbb{R}} e^{2\pi i (X_i - \theta_i)^2 -(X_i - \theta_i)^2} \frac{w_t}{\sqrt{2\pi \sigma^2}} e^{-\frac{i}{2\sigma^2} (m_t - \theta_i)^2} d\theta_i &= \mathbb{E} e^{\frac{i}{\kappa \sigma^2} \sum_{t=1}^T (X_i - m_t)^2} \\
&= \mathbb{E} e^{\frac{i}{\kappa \sigma^2} \sum_{t=1}^T (X_i - m_t)^2} \\
&= \mathbb{E} \left\{ e^{\frac{i}{\kappa \sigma^2} \sum_{t=1}^T (X_i - m_t)^2} \right\} \prod_{i \in S^*} \left\{ \mathbb{E} \left\{ e^{\frac{i}{\kappa \sigma^2} \sum_{t=1}^T (X_i - m_t)^2} \right\} \right\}.
\end{align*}
\]

Therefore we have

\[
\begin{align*}
\prod_{i \in S^*} \int_{\mathbb{R}} e^{2\pi i (X_i - \theta_i)^2 -(X_i - \theta_i)^2} \frac{w_t}{\sqrt{2\pi \sigma^2}} e^{-\frac{i}{2\sigma^2} (m_t - \theta_i)^2} d\theta_i &= \prod_{i \in S^*} \left\{ e^{\frac{i}{\kappa \sigma^2} \sum_{t=1}^T (X_i - m_t)^2} \right\} \prod_{i \in S^*} \left\{ \mathbb{E} \left\{ e^{\frac{i}{\kappa \sigma^2} \sum_{t=1}^T (X_i - m_t)^2} \right\} \right\}.
\end{align*}
\]

Since

\[
\prod_{i \in S^*} \left\{ \sum_{t=1}^T w_t \exp\left(-\frac{\kappa (X_i - m_t)^2}{\kappa \sigma^2 + 1}\right) \right\} \geq \prod_{i \in S^*} \left\{ w_t \exp\left(-\sum_{t=1}^T \frac{\kappa (X_i - m_t)^2}{\kappa \sigma^2 + 1}\right) \right\}
\]

We have \( \prod_{i \in S^*} w_t \geq (C s_n/n)^{s_n} = \exp(\log C \cdot s_n - \varepsilon_n) \). Since \( \sum_{i \in S^*} (X_i - m_t)^2 \leq 2 \sum_{i \in S^*} (X_i - \theta_i)^2 + 2 \sum_{i \in S^*} (X_i - \theta_i)^2 \) and \( \sum_{i \in S^*} (X_i - \theta_i)^2 \sim \chi_s^2 \), by concentration inequality, we have

\[
P\left( \sum_{i \in S^*} (X_i - \theta_i)^2/s_n - 1 \geq \log(n/s_n) - 1 \right) \leq 2 \exp(-s_n(\log(n/s_n) - 1)^2/8).
\]

Therefore \( \sum_{i \in S^*} (X_i - \theta_i)^2 = o_p(\varepsilon_n) \). Hence

\[
\prod_{i \in S^*} \left\{ \sum_{t=1}^T w_t \exp\left(-\frac{\kappa (X_i - m_t)^2}{\kappa \sigma^2 + 1}\right) \right\} \geq \exp(\log C \cdot s_n - \varepsilon_n - \frac{\kappa}{\kappa \sigma^2 + 1} o_p(\varepsilon_n)).
\]

According to Law of Large Numbers,

\[
\prod_{i \in S^*} e^{2\pi i (X_i - \theta_i)^2} = e^{2\pi \prod_{i \in S^*} (X_i - \theta_i)^2} = e^{\frac{i}{\kappa \sigma^2} + o_p(s_n)}.
\]
Therefore
\[
\prod_{i \in S^*} \int_{\mathbb{R}} e^{\frac{w_i}{2}((X_i-\theta_i)^2-(X_i-\theta_i^*)^2)} \sum_{t=1}^{T} \frac{w_t}{\sqrt{2\pi \sigma^2}} e^{-\frac{1}{2\sigma^2}(m_t-\theta_i)^2} d\theta_i \geq (1 + \kappa \sigma^2)^{-\frac{\alpha}{2}} \exp(-\varepsilon_n - o_{p}(\varepsilon_n)).
\]

Using the same trick in the proof of Lemma 1 in [Martin and Walker (2014)], \(\int_0^1 w^{n-s_n}(1-w)^s_n \Pi(dw)\) is lower bounded by \(\frac{\alpha}{1 + \alpha} \exp[-2\varepsilon_n - \alpha s_n + o(s_n)]\) as \(n\) is sufficiently large. Putting these pieces together, we obtain
\[
D_n \geq \frac{\alpha}{1 + \alpha} \exp\{-3\varepsilon_n - o_{p}(\varepsilon_n)\}.
\]

### 4.1 Proof for Theorem

The main aim of the proof is to show the numerator, for sets \(A_n\) away from \(\theta^*\), is not too large. Let \(N_n\) be the numerator for \(Q_n(A_{M\varepsilon_n})\), i.e.,
\[
N_n = \int_0^1 \int_{A_{M\varepsilon_n}} \prod_{i=1}^{n} \left\{ \frac{p_{\theta_i}(x_i)}{p_{\theta^*_i}(x_i)} \right\}^{\kappa} \Pi_{w,w,m}(d\theta_i) \Pi(dw).
\]

Taking expectation of \(N_n\), with respect to \(P_{\theta^*}\), we get
\[
\mathbb{E}_{\theta^*}(N_n) = \int_0^1 \int_{A_{M\varepsilon_n}} \left( \prod_{i=1}^{n} \int_{\mathbb{R}} \left\{ \frac{p_{\theta_i}(x_i)}{p_{\theta^*_i}(x_i)} \right\}^{\kappa} p_{\theta^*_i}(x_i) dx_i \right) \Pi_{w,w,m}(d\theta_i) \Pi(dw).
\]

Since \(\int_{\mathbb{R}} \left\{ \frac{p_{\theta_i}(x_i)}{p_{\theta^*_i}(x_i)} \right\}^{\kappa} p_{\theta^*_i}(x_i) dx_i = \exp\{-\frac{\kappa(1-\kappa)}{2}(\theta_i - \theta^*_i)^2\}\), we have
\[
\mathbb{E}_{\theta^*}(N_n) = \int_0^1 \int_{A_{M\varepsilon_n}} \prod_{i=1}^{n} \exp\{-\frac{\kappa(1-\kappa)}{2}(\theta_i - \theta^*_i)^2\} \Pi_{w,w,m}(d\theta_i) \Pi(dw).
\]

Since \(A_{M\varepsilon_n} = \{\theta \mid \|\theta - \theta^*\|^2 > M\varepsilon_n\}\), we have
\[
\mathbb{E}_{\theta^*}(N_n) \leq \exp\{-\frac{\kappa(1-\kappa)}{2}M\varepsilon_n\}.
\]

Let \(\frac{\kappa(1-\kappa)}{2} = c\). Next, take \(M\) such that \(cM > 3\), and then take \(K \in (3,cM)\), then using Markov Inequality we get \(P_{\theta^*}(N_n > e^{-K\varepsilon_n}) \leq e^{-(cM-K)\varepsilon_n}\). This upper bound has a finite sum over \(n \geq 1\), so the Borel-Catelli lemma gives that \(N_n \leq e^{-K\varepsilon_n}\) with probability 1 for all large \(n\). Therefore we get
\[
\frac{N_n}{D_n} \leq \frac{1 + \alpha}{\alpha} e^{-(K-3)\varepsilon_n - \eta s_n + o(s_n)}.
\]

Since \(s_n = o(\varepsilon_n)\), \(Q_n(A_{M\varepsilon_n}) \to 0\) as \(n \to \infty\).
Variational Inference Algorithm Derivation

We will derive the variational inference algorithm for Dirichlet process mixture model. \( \alpha_0, T, w_0, \sigma_0^2 \) and the data vector \( X^n \) is given in advance. The data generating process is summarized in [7]. We treat \( Z \) as latent variables and \( V, \eta^*, \xi \) as parameters. Posterior distribution of all the parameters and latent variables is proportional to

\[
P(Z, V, \eta^*, \xi | X^n) \propto P(X^n | Z, V, \eta^*, \xi) P(Z | V) P(\eta^* | \xi) P(V | \alpha_0) P(\xi | w_0).
\]

Recall that under the fully factorized assumption, we have

\[
\log P \left( \sum_{t=1}^T \xi_i (1 - w_0)^T = \sum_{t=1}^T \xi_i \prod_{t=1}^{T-1} (1 - V_t)^{\alpha_0 - 1} \prod_{t: \xi_t = 1} \delta_0(\eta^*_t). \right.
\]

where \( \log p \left( \sum_{i=1}^n Z_i = 1 \right) = \frac{1}{2 \pi \sigma_0} \exp \left( -\frac{(\eta^*_t)^2}{2 \sigma_0^2} \right) \prod_{t: \xi_t = 1} \exp \left( -\frac{\sum_{i=1}^n (X_i - \eta^*_t)^2 1_{Z_i = t}}{2} \right) \].

Recall that under the fully factorized variational assumption, we have

\[
q(Z, V, \eta^*, \xi) = q_{p.n.m.}(\eta^*, \xi) q_{\gamma_1, \gamma_2}(V) q_{\Phi}(Z).
\]

Define \( P(Z_i = t) = \phi_{i,t} \). First we find the optimal form of \( q(\eta^*, \xi) \), which satisfies

\[
\log q(\eta^*, \xi) = \mathbb{E}_{V, Z}[\log(w_0 \sum_{t=1}^T \xi_t (1 - w_0)^T - \sum_{t=1}^T \xi_t \prod_{t: \xi_t = 1} \delta_0(\eta^*_t))].
\]

\[
= \sum_{t=1}^T \left[ 1_{\xi_t = 1} (\log w_0 + \log \delta_0(\eta^*_t)) + 1_{\xi_t = 0} (\log(1 - w_0) - \log \sqrt{2 \pi \sigma_0^2} - \frac{(\eta^*_t)^2}{2 \sigma_0^2}) - \sum_{i=1}^n \phi_{i,t}(X_i - \eta^*_t)^2 \right] + \text{const}
\]

\[
= \sum_{t=1}^T \left[ 1_{\xi_t = 1} (\log w_0 + \log \delta_0(\eta^*_t)) - \frac{\sum_{i=1}^n \phi_{i,t} X_i^2}{2} \right] + \text{const}
\]

\[
= \sum_{t=1}^T \log q(\xi_t, \eta^*_t);
\]

where \( \log q(\xi_t, \eta^*_t) = 1_{\xi_t = 1} (\log w_0 + \log \delta_0(\eta^*_t)) - \frac{\sum_{i=1}^n \phi_{i,t} X_i^2}{2} + 1_{\xi_t = 0} (\log(1 - w_0) - \log \sqrt{2 \pi \sigma_0^2} - \frac{(\eta^*_t)^2}{2 \sigma_0^2} - \frac{\sum_{i=1}^n \phi_{i,t} (X_i - \eta^*_t)^2}{2}) + \text{const} \). Therefore the optimal form of \( q_{p.m.}(\eta^*, \xi) \) is fully factor-
ized across different clusters:

\[ q_{p,m,\tau}(\eta^*, \xi) = \prod_{t=1}^{T} q_{p_t,m_t,\tau_t}(\eta^*_t, \xi_t). \]

In order to determine the updating formula for \( p_t, m_t, \tau_t \), we use Method of Undetermined Coefficients. Suppose \( q(\xi_t, \eta^*_t) \) is a mixture of a point mass of zero and normal distribution,

\[ q(\xi_t, \eta^*_t) = p_t \delta_{\xi_t=1} + (1-p_t)N(\eta^*_t, \xi_t), \]

therefore

\[ \log q(\xi_t, \eta^*_t) = \begin{cases} \log p_t + \log \delta_{\xi_t=1} + \log \left( \sqrt{2\pi\tau^2_t} - \frac{(\eta^*_t - m_t)^2}{2\tau^2_t} \right) + \text{const} & \text{if } \xi_t = 1 \\ \log (1-p_t) - \log \left( \sqrt{2\pi\tau^2_t} - \frac{(\eta^*_t - m_t)^2}{2\tau^2_t} \right) + \text{const} & \text{if } \xi_t = 0 \end{cases} \]

Even though there’s a normalizing constant, but the difference between multipliers of \( 1_{\xi_t=1} \) and \( 1_{\xi_t=0} \) is invariant with respect to the constant. Therefore we have the following equation:

\[ \log p_t - \log (1-p_t) + \log \left( \sqrt{2\pi\tau^2_t} - \frac{(\eta^*_t - m_t)^2}{2\tau^2_t} \right) = \]

\[ \log w - \log (1-w) - \frac{\sum_{i=1}^{n} \phi_{i,t}X^2_i}{2} + \frac{(\eta^*_t)^2}{2\sigma^2_0} + \frac{\sum_{i=1}^{n} \phi_{i,t}(X_i - \eta^*_t)^2}{2}; \]

which holds for any \( \eta^*_t \in \mathbb{R} \). The solutions are given as follows:

\[ m_t = \frac{\sigma^2_0 \cdot \sum_{i=1}^{n} \phi_{i,t}X_i}{\sigma^2_0 \cdot \sum_{i=1}^{n} \phi_{i,t} + 1}, t = 1, 2, \cdots, T \]

\[ \tau^2_t = \frac{\sigma^2_0 \cdot \sum_{i=1}^{n} \phi_{i,t}}{\sigma^2_0 \cdot \sum_{i=1}^{n} \phi_{i,t} + 1}, t = 1, 2, \cdots, T \]

\[ p_t = \frac{\exp \left( \log(w_0) - \log(1-w_0) + \log \left( \sqrt{\sigma^2_0 \cdot \sum_{i=1}^{n} \phi_{i,t}X^2_i} \right) \right)}{\exp \left( \log(w_0) - \log(1-w_0) + \log \left( \sqrt{\sigma^2_0 \cdot \sum_{i=1}^{n} \phi_{i,t}X^2_i} \right) \right) + 1}, t = 1, 2, \cdots, T. \]
Next we deal with the optimal form for \( q(V) \), which satisfies

\[
\log q(V) = \mathbb{E}_\mathbf{Z}[\log(\prod_{t=1}^{T-1} (1 - V_t)^{\alpha_0 - 1} \cdot V_1^{\gamma_{11}} \cdot \prod_{i=2}^{T-1} (1 - V_i)^{\gamma_{12}}) + \text{const}]
\]

\[
= \sum_{i=1}^{n} \phi_{i,1} \cdot \log V_1 + (\alpha_0 - 1 + \sum_{t=2}^{T} \sum_{i=1}^{n} \phi_{i,t}) \log(1 - V_1) + \sum_{i=1}^{n} \phi_{i,2} \log V_2 + (\alpha_0 - 1 + \sum_{t=3}^{T} \sum_{i=1}^{n} \phi_{i,t}) \log(1 - V_2) + \cdots + \sum_{i=1}^{n} \phi_{i,T-1} \log V_{T-1}
\]

\[
+ (\alpha_0 - 1 + \sum_{i=1}^{n} \phi_{i,t}) \log(1 - V_{T-1}) + \text{const}
\]

\[
\equiv \sum_{t=1}^{T-1} \log q(V_t);
\]

where \( \log q(V_1) = \sum_{i=1}^{n} \phi_{i,1} \cdot \log V_1 + (\alpha_0 - 1 + \sum_{t=2}^{T} \sum_{i=1}^{n} \phi_{i,t}) \log(1 - V_1) + \text{const} \),

\( \log q(V_2) = \sum_{i=1}^{n} \phi_{i,2} \log V_2 + (\alpha_0 - 1 + \sum_{t=3}^{T} \sum_{i=1}^{n} \phi_{i,t}) \cdot \log(1 - V_2) + \text{const} \),

\( \cdots \),

\( \log q(V_{T-1}) = \sum_{i=1}^{n} \phi_{i,T-1} \cdot \log V_{T-1} + (\alpha_0 - 1 + \sum_{i=1}^{n} \phi_{i,T-1}) \log(1 - V_{T-1}) + \text{const} \).

Thus we proved

\[
q_{\gamma_1, \gamma_2}(V) = \prod_{t=1}^{T-1} q_{\gamma_{1t}, \gamma_{2t}}(V_t).
\]

Besides, \( V_1 \) follows Beta Distribution with parameters \( (\gamma_{11}, \gamma_{21}) = (\sum_{i=1}^{n} \phi_{k,1} + 1, \alpha_0 + \sum_{t=2}^{T} \sum_{i=1}^{n} \phi_{i,t}) \), \( V_2 \) follows Beta Distribution with parameters \( (\gamma_{12}, \gamma_{22}) = (\sum_{i=1}^{n} \phi_{k,2} + 1, \alpha_0 + \sum_{t=3}^{T} \sum_{i=1}^{n} \phi_{i,t}) \), \( \cdots \), \( V_{T-1} \) follows Beta Distribution with parameters \( (\gamma_{T-1,2}, \gamma_{T-1,2}) = (\sum_{i=1}^{n} \phi_{k,T-1} + 1, \alpha_0 + \sum_{i=1}^{n} \phi_{i,T-1}) \).

Finally, we deal with \( q(Z) \). We have the following optimal form

\[
\log q(Z) = \mathbb{E}_{\xi, \eta, \mathbf{X}}[\log(\prod_{t=1}^{T} \prod_{i=1}^{T} \pi_t^{Z_{i,t}}) \cdot \exp\left(-\frac{\sum_{i=1}^{n} \sum_{t=1}^{T} (X_i - \eta_i^p)^2 Z_{i,t}}{2}\right)] + \text{const}
\]

\[
= \sum_{i=1}^{n} \mathbb{E}_{\xi, \eta, \mathbf{X}}[\sum_{t=1}^{T} (\log \pi_t - (X_i - \eta_i^p)^2) Z_{i,t}] + \text{const}
\]

\[
= \sum_{i=1}^{n} \sum_{t=1}^{T} (\mathbb{E}_V(\log \pi_t) - \frac{\mathbb{E}_{\xi, \eta}((X_i - \eta_i^p)^2)}{2}) Z_{i,t} + \text{const}
\]

\[
= \sum_{i=1}^{n} \log q(Z_i);
\]
therefore we proved $q_{\Phi}(Z) = \prod_{i=1}^{n} q_{\phi_i}(Z_i)$. Since $E_{\xi_t,\eta_t}(X_i - \eta_t^2)^2 = X_i^2 - 2(1 - p_t)m_t X_i + (1 - p_t)(m_t^2 + \tau_t^2)$, we have

$$\log q(Z_i) = \sum_{t=1}^{T} (E_{V_t}(\log \pi_t) + (1 - p_t)m_t X_i - \frac{1}{2}(1 - p_t)(m_t^2 + \tau_t^2))1_{Z_i=t} + \text{const}$$

$$= \sum_{t=1}^{T} [E_{\gamma_1,t,\gamma_2,t}(\log V_t) + \sum_{i=1}^{t-1} E_{\gamma_1,i,\gamma_2,i}(\log(1 - V_i)) + (1 - p_t)m_t X_i - \frac{1}{2}(1 - p_t)(m_t^2 + \tau_t^2)]1_{Z_i=t} + \text{const}.$$

Therefore $q(Z_i)$ is the probability mass function of Multinomial Distribution. Once we fix $i$, $\phi_{i,t} \propto \exp(S_t)$, where $S_t = \exp[E_{\gamma_1,t,\gamma_2,t}(\log V_t) + \sum_{i=1}^{t-1} E_{\gamma_1,i,\gamma_2,i}(\log(1 - V_i)) + (1 - p_t)m_t X_i - \frac{1}{2}(1 - p_t)(m_t^2 + \tau_t^2)].$

References

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

Blei, D. M., M. I. Jordan, et al. (2006). Variational inference for dirichlet process mixtures. *Bayesian Analysis* 1(1), 121–143.

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

Bühlmann, P. and S. Van De Geer (2011). *Statistics for high-dimensional data: methods, theory and applications*. Springer Science & Business Media.

Castillo, I., A. van der Vaart, et al. (2012). Needles and straw in a haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics* 40(4), 2069–2101.

Donoho, D. L., I. M. Johnstone, J. C. Hoch, and A. S. Stern (1992). Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society. Series B (Methodological)*, 41–81.

Escobar, M. D. and M. West (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association* 90(430), 577–588.
Fan, J. and J. Lv (2010). A selective overview of variable selection in high dimensional feature space. *Statistica Sinica* 20(1), 101.

George, E. and D. P. Foster (2000). Calibration and empirical bayes variable selection. *Biometrika* 87(4), 731–747.

Golubev, G. K. (2002). Reconstruction of sparse vectors in white gaussian noise. *Problems of Information Transmission* 38(1), 65–79.

Jiang, W. and C.-H. Zhang (2009). General maximum likelihood empirical bayes estimation of normal means. *The Annals of Statistics* 37(4), 1647–1684.

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

Johnstone, I. M. and B. W. Silverman (2005). Ebayesthresh: R and s-plus programs for empirical bayes thresholding. *Journal of Statistical Software* 12, 1–38.

Koenker, R. (2014). A gaussian compound decision bakeoff. *Stat* 3(1), 12–16.

Koenker, R. and I. Mizera (2014). Convex optimization, shape constraints, compound decisions, and empirical bayes rules. *Journal of the American Statistical Association* 109(506), 674–685.

Martin, R. and S. G. Walker (2014). Asymptotically minimax empirical bayes estimation of a sparse normal mean vector. *Electronic Journal of Statistics* 8(2), 2188–2206.

Robbins, H. (1951). Asymptotically subminimax solutions of compound statistical decision problems. In *Proc. Second Berkeley Symp. Math. Statist. Probab.*, Volume 1, pp. 131–148. Univ. California Press, Berkeley.

Walker, S. and N. L. Hjort (2001). On bayesian consistency. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)* 63(4), 811–821.

Zhang, C.-H. and J. Huang (2008). The sparsity and bias of the lasso selection in high-dimensional linear regression. *The Annals of Statistics* (36), 1567–1594.