

A Review for Needles and Straw in a Haystack: Posterior concentration for possibly sparse sequences

Shuyu Guo

December 2024

1 Overview

To investigate the unknown possibly sparse mean vector, Castillo and Vaart[1] proposed a fully Bayesian approach. Although there has been some previous work on Bayesian inference on the sparse mean, this paper demonstrates the concentration property of posterior distribution under their wise priors setting and thus yields estimators as some functionals of posterior measures which can achieve an optimal risk rate.

In the following part of the report, the model setting and prior construction will be briefly introduced in section 2. And Section 3 will include the main results distilled from the theorems established for posterior concentration properties. Section 4 displays the simulation results and analysis. Section 5 is a conclusive discussion.

2 Construction and Main Results

2.1 Model Setting:

Suppose that we have the model:

$$X_i = \theta_i + \epsilon_i, \quad i = 1, \dots, n$$

where $\epsilon_i \stackrel{iid}{\sim} \mathcal{N}(0, 1)$ and θ is the parameter that we are interested in. In this problem, θ is assumed to be sparse, and here in the report for simplicity we define the sparsity by the Nearly Black Class $\ell_0[p_n]$,

$$\ell_0[p_n] = \{\theta \in \mathbb{R}^n : \#(1 \leq i \leq n : \theta_i \neq 0) \leq p_n\},$$

where p_n is assumed to be $o(n)$. Thus, the number of nonzero elements in θ grows much slower than the full length of θ .

2.2 Prior Construction:

Since the true mean is sparse, one might want the posterior on θ to concentrate at the regions where the dimension of nonzeros is bounded. Also, we are interested in investigating the recovery rate. These two desired aspects can be achieved by a delicate hierarchical construction of the prior on θ :

1. Exponential decrease prior on the number of nonzeros p :

$$\pi_n(p) \leq D\pi_n(p-1)$$

where D is some value less than 1 and $p > Cp_n$. By this exponential decrease prior, we assign more weights on the sparser models.

2. Given p , the set of nonzero indicators $S \subset \{1, \dots, n\}$ is uniformly distributed:

$$P(S \mid |S| = p) = \binom{n}{p}^{-1}$$

3. Given the nonzero dimension p and nonzero set S , (for simplicity) we assume a product prior for θ_S , where θ_S includes nonzero θ_i 's only:

$$g_S(\theta_S) = \prod_{i \in S} g(\theta_i)$$

where density $g(\cdot)$ has mean zero and finite second moment.

2.3 Main Results:

With above settings of prior, theorems are established for the posterior concentration properties. Here we only illustrate the main results for the simplified cases, and comprehensive discussions on more generalized cases can be found in the paper.

- **Result 1 (Theorem 2.1 Castillo and Vaart[1]):**

With the prior construction in section 2.2, $\exists M > 0$, as $p_n, n \rightarrow \infty$,

$$\sup_{\theta_0 \in \ell_0[p_n]} P_{n, \theta_0} \Pi(\theta : |S_\theta| > Mp_n | X) \rightarrow 0$$

This result illustrates that the posterior distribution concentrates at the subset where the dimension of nonzeros are at most a multiple of p_n .

- **Result 2 (Theorem 2.2 Castillo and Vaart[1]):**

This result emphasizes the investigation on the recovery rate. Since we want to prevent the posterior from concentrating on the region where θ is way sparser than true nonzero dimension, we will need to put sufficient prior mass on p_n and the prior of nonzero θ_i should have a sufficiently heavy tail.

Specifically, if we additionally assume

$$\pi_n(p_n) \gtrsim \exp(-cp_n \log(n/p_n))$$

and

$$g(\cdot) = e^h$$

where $h(\cdot)$ is uniformly Lipschitz,

$$|h(x) - h(y)| \lesssim 1 + |x - y|$$

then $\exists M > 0$, as $p_n, n \rightarrow \infty$ and $p_n/n \rightarrow 0$,

$$\sup_{\theta_0 \in \ell_0[p_n]} P_{n,\theta_0} \Pi(\theta : d_2(\theta, \theta_0) > Mr_{n,2}^* | X) \rightarrow 0,$$

where $r_{n,2}^* = p_n \log(n/p_n)$ is the square minimax rate.

Thus result 2 indicates that the posterior of θ concentrates around a neighborhood of the true mean where the ℓ_2 distance is bounded by a multiple of minimax rate. In other words, the posterior of θ contracts to the true θ_0 at the minimax rate, uniformly over all $\theta_0 \in \ell[p_n]$.

– **Remark 1 of Result 2:**

One example of π_n that has such lower bound at p_n is beta-binomial prior:

$$p|\alpha \sim \text{Binomial}(n, \alpha) \text{ and } \alpha \sim \text{Beta}(1, n+1)$$

where we have $\pi(p-1)/\pi(p) \rightarrow 1/2$ (exponential decrease) and $\pi_n(p_n) \geq (1 - \frac{p_n+1}{n+1})^n \rightarrow \exp(-p_n)$ as $p_n/n \rightarrow 0$.

– **Remark 2 of Result 2:**

There are many choices of the $g(\cdot)$ satisfying the assumption for above theorem, including Laplace, Student distributions, densities with polynomial tails or proportional to $e^{-|x|^\alpha}$ where $0 < \alpha \leq 1$.

• **Result 3 (Corollary 2.1 Castillo and Vaart[1]):**

With the preceding prior construction in section 2.2, if we additionally assume $\pi_n(p) \lesssim \exp(-ap \log bn/p)$ for $a \geq 1$ and b larger than some constant, we have

$$\sup_{\theta_0 \in \ell_0[p_n]} P_{n,\theta_0} \|\mathbb{E}(\theta|X) - \theta_0\|^2 \lesssim r_{n,2}^*$$

This result reveals that the posterior mean is an optimal estimator, for its risk is at the minimax rate. Similar result can be established for posterior median by assuming $\pi_n(p) \propto \exp(-ap \log bn/p)$.

• **Result 4 (Theorem 2.8 Castillo and Vaart[1]):**

This result reveals the necessity of heavy tail prior on nonzero θ_i . For simplicity, now we assume that $g(y) \propto e^{-y^\alpha}$ for some $\alpha \geq 2$ and the same lower bound of $\pi_n(p_n)$ as Result 2. When $\|\theta_0^n\|^2/r_{n,2}^* \rightarrow \infty$ where θ_0^n denotes the sequence of true mean as n tends to infinity, then as $n \rightarrow \infty$, for some small enough η we have

$$P_{n,\theta_0^n} \Pi_n(\theta : d_2(\theta, \theta_0) \leq \eta \|\theta_0^n\|^2 | X^n) \rightarrow 0$$

That is saying, when the true nonzero signals are strong such that its squared ℓ_2 norm grows faster than the minimax rate, then the posterior contracts even slower than a multiple of $\|\theta_0^n\|^2$. Though the thin tail prior on nonzero mean has the impact of shrinking the posterior back to zero, which one might consider a good estimator when true mean has very weak signal. But even the zero estimator can reach a minimax rate when $\|\theta_0^n\|^2 \leq r_{n,2}^*$, which means that the estimator from the fully Bayesian approach can not even beat the zero estimator. In comparison, with the heavy tail prior in Result 2, the posterior contracts at a minimax rate uniformly over the space $\theta_0 \in \ell_0[p_n]$.

3 Estimators: Posterior Functional

To estimate the true mean θ_0 , we now consider the posterior functionals such as posterior mean and median as the point estimator for θ_0 . Specifically, we will need to compute the coordinate-wise posterior distribution and hence get its posterior functionals. Since the joint posterior can be obtained by

$$\theta, S, p | X \propto \frac{\pi_n(p)}{\binom{n}{p}} \prod_{i \in S} g(\theta_i) \phi(x_i - \theta_i) \prod_{i \notin S} \phi(x_i)$$

where the normalizing constant

$$Q_n = \sum_{p=0}^n \frac{\pi_n(p)}{\binom{n}{p}} \sum_{|S|=p} \prod_{i \in S} \psi(x_i) \prod_{i \notin S} \phi(x_i)$$

$\psi(x_i)$ is the convolution of $g(\theta_i)$ and $\phi(x_i - \theta_i)$. According to calculation, we can obtain

$$\Pi(\theta_1 | X) = (1 - q_{1,n})\delta_0 + q_{1,n} \frac{\phi(x_1 - \theta_1)g(\theta_1)}{\psi(x_1)}$$

which shows that the marginal posterior of θ_1 is a mixture of point mass at zero and a smooth function over the space excluding the origin. $1 - q_{1,n}$, the posterior probability at zero, takes the following form

$$1 - q_{1,n} = \frac{1}{Q_n} \sum_{p=0}^n \frac{\pi_n(p)}{\binom{n}{p}} \sum_{|S|=p, 1 \notin S} \prod_{i \in S} \psi(x_i) \prod_{i \notin S} \phi(x_i)$$

Thus the posterior functionals:

- coordinate-wise posterior mean

$$\hat{\theta}_1^{PM} = \frac{q_{1,n}}{\psi(x_1)} \int \theta_1 \phi(x_1 - \theta_1) g(\theta_1) d\theta_1$$

- To compute coordinate-wise posterior median, we first investigate the cdf:

$$\Pi(\theta_1 \leq u | X) = (1 - q_{1,n})\mathbb{I}(u \geq 0) + H_{1,n}(u)$$

where $H_{1,n}(u) = \frac{q_{1,n}}{\psi(x_1)} \int_{-\infty}^u \phi(x_1 - \theta_1) g(\theta_1) d\theta_1$

- When $\Pi(\theta_1 < 0 | X) = H_{1,n}(0) > 0.5$, solve $H_{1,n}(u) = 0.5$ for $\hat{\theta}_1^{PMd}$ ($\hat{\theta}_1^{PMd} < 0$)
- When $\Pi(\theta_1 \leq 0 | X) = 1 - q_{1,n} + H_{1,n}(0) < 0.5$, solve $H_{1,n}(u) + 1 - q_{1,n} = 0.5$ for $\hat{\theta}_1^{PMd}$ ($\hat{\theta}_1^{PMd} > 0$)
- Otherwise, we have $\hat{\theta}_1^{PMd} = 0$

4 Simulation Results

To verify the usefulness of the fully Bayesian procedure for estimating possibly sparse θ , we now design a simulation study and compare the results with the Empirical Bayes procedure proposed by Johnstone and Silverman[2]. Since the fully Bayesian procedure is computationally expensive, here we only implement with data size $n = 100$. We set the $p_n \in \{5, 10, 20\}$ and nonzero mean has absolute value 2.5 or 5. For instance, when we take $p_n = 5$ and signal equals 5, then the true mean $\theta = (5, 5, 5, 5, 5, 0, \dots, 0)$. For each combination of p_n and signal, we replicate 100 iterations and show the results including MSE ($\mathbb{E}(\|\hat{\theta} - \theta\|^2)$), dimension of nonzeros in estimator, true positive(TP: $\#(\hat{\theta}_i \neq 0 \ \& \ \theta_i \neq 0)$) and false positive(FP: $\#(\hat{\theta}_i \neq 0 \ \& \ \theta_i = 0)$).

For the nonzero dimension p , we consider two choices: $\pi_n^{(1)}(p) \propto \binom{2n-p}{n}$ and $\pi_n^{(2)}(p) \propto \binom{2n-p}{n}^{0.1}$.

We firstly choose standard Laplacian density as the prior of nonzero θ_i . In comparison, we also add the simulation where a 'wrong' prior, standard normal, is applied for the nonzero θ_i and verify its invalidity through its MSE. Finally we add the results from fitting with EBayesThresh package. In the table below,

- $\hat{\theta}_{L1}^{PM}$ and $\hat{\theta}_{L1}^{PMd}$ denote the posterior mean and median with $g(\cdot) = DE(1)$ and $\pi_n^{(1)}(p)$.
- $\hat{\theta}_{L2}^{PM}$ and $\hat{\theta}_{L2}^{PMd}$ denote the posterior mean and median with $g(\cdot) = DE(1)$ and $\pi_n^{(2)}(p)$.
- $\hat{\theta}_{G1}^{PM}$ and $\hat{\theta}_{G1}^{PMd}$ denote the posterior mean and median from fully Bayesian procedure with $g(\cdot) = N(0, 1)$ and $\pi_n^{(1)}(p)$.
- $\hat{\theta}_{G2}^{PM}$ and $\hat{\theta}_{G2}^{PMd}$ denote the posterior mean and median from fully Bayesian procedure with $g(\cdot) = N(0, 1)$ and $\pi_n^{(2)}(p)$.
- $\hat{\theta}^{EBM}$ and $\hat{\theta}^{EBMd}$ denote the EB mean and EB median from EBayesThresh which also chooses standard Laplacian as the prior for nonzero mean.

4.1 Observations

- Since $\pi_n^{(1)}(p)$ decays faster than $\pi_n^{(2)}(p)$, the former one strongly favors sparser models. We observe that when signal is weaker, the choice of $\pi_n^{(1)}(p)$ leads to challenges in detecting nonzero θ_i . But when signal is strong enough, it is helpful in controlling false discovery and MSE.
- When signal is strong, as stated in Result 4, the MSE of $\hat{\theta}_G^{PM}$ and $\hat{\theta}_G^{PMd}$ are much larger than other estimators. Especially, when we choose $\pi_n^{(2)}$ as prior for p , though the dimension of nonzeros in $\hat{\theta}_{G1}^{PMd}$ is comparable to other estimators with good performance, the MSE from $\hat{\theta}_{G1}^{PM}$ and $\hat{\theta}_{G1}^{PMd}$ are significantly larger than others.
- The EB procedure consistently performs well in the stronger signal cases regardless of the true nonzero dimension. However its MSE grows much faster than the estimators from fully Bayesian procedure which choose $\pi_n^{(2)}(\cdot)$ as prior for p .

		signal = 2.5				signal = 5			
		MSE	dim	TP	FP	MSE	dim	TP	FP
$p_n = 5$	$\hat{\theta}_{L1}^{PM}$	29.47				15.51			
	$\hat{\theta}_{L1}^{PMd}$	35.02	1.36	1.3	0.06	16.09	4.76	4.74	0.02
	$\hat{\theta}_{G1}^{PM}$	35.84				44.27			
	$\hat{\theta}_{G1}^{PMd}$	40.92	0.54	0.52	0.02	45.16	4.42	4.42	0
	$\hat{\theta}_{L2}^{PM}$	18.71				15.72			
	$\hat{\theta}_{L2}^{PMd}$	20.74	4.24	3.39	0.85	12.70	6.64	4.97	1.67
	$\hat{\theta}_{G2}^{PM}$	22.56				38.20			
	$\hat{\theta}_{G2}^{PMd}$	23.68	4.58	3.42	1.16	34.63	7.81	4.97	2.84
	$\hat{\theta}^{EBM}$	23.56				12.96			
	$\hat{\theta}^{EBMd}$	29.27	3.42	2.72	0.7	11.12	5.93	4.95	0.98
		signal = 2.5				signal = 5			
		MSE	dim	TP	FP	MSE	dim	TP	FP
$p_n = 10$	$\hat{\theta}_{L1}^{PM}$	53.12				29.83			
	$\hat{\theta}_{L1}^{PMd}$	65.75	3.16	3.1	0.06	30.00	9.72	9.6	0.12
	$\hat{\theta}_{G1}^{PM}$	66.72				80.12			
	$\hat{\theta}_{G1}^{PMd}$	77.81	1.72	1.72	0	78.96	9.48	9.46	0.02
	$\hat{\theta}_{L2}^{PM}$	35.73				31.00			
	$\hat{\theta}_{L2}^{PMd}$	39.96	10.32	7.5	2.82	26.07	15.95	9.99	5.96
	$\hat{\theta}_{G2}^{PM}$	42.70				75.70			
	$\hat{\theta}_{G2}^{PMd}$	44.15	12.61	8.05	4.56	70.66	20.37	10	10.37
	$\hat{\theta}^{EBM}$	44.46				25.00			
	$\hat{\theta}^{EBMd}$	57.26	4.85	4.37	0.48	22.52	10.94	9.79	1.15
		signal = 2.5				signal = 5			
		MSE	dim	TP	FP	MSE	dim	TP	FP
$p_n = 20$	$\hat{\theta}_{L1}^{PM}$	89.78				46.56			
	$\hat{\theta}_{L1}^{PMd}$	114.02	8.72	8.6	0.12	44.64	20.2	19.78	0.42
	$\hat{\theta}_{G1}^{PM}$	115.55				142.35			
	$\hat{\theta}_{G1}^{PMd}$	138.73	6.16	6.08	0.08	138.63	19.94	19.7	0.24
	$\hat{\theta}_{L2}^{PM}$	54.68				52.94			
	$\hat{\theta}_{L2}^{PMd}$	56.90	27.7	18.02	9.68	47.26	38.64	19.98	18.66
	$\hat{\theta}_{G2}^{PM}$	69.58				142.35			
	$\hat{\theta}_{G2}^{PMd}$	68.14	35.5	19.02	16.48	138.63	50.68	19.98	30.7
	$\hat{\theta}^{EBM}$	86.52				44.56			
	$\hat{\theta}^{EBMd}$	116.42	8.7	8.48	0.22	41.14	21.14	19.82	1.32

Table 1: Simulation Results. The values which indicate the best performance is marked in red.

- When signal is weaker, $\hat{\theta}_{L2}^{PM}$ and $\hat{\theta}_{L2}^{PMd}$ exhibit the best performances in this table regardless of p_n . When signal becomes stronger, then $\hat{\theta}_{L1}^{PM}$ and $\hat{\theta}_{L1}^{PMd}$ have a comparable performance with EB mean and median. Thus, the fully Bayesian procedure is more adaptive to different settings, though it requires carefully choosing of prior on p .

5 Discussion

This fully Bayesian procedure provides a posterior distribution of the unknown mean, which enables the investigation on any function of this vector. More importantly, with delicately chosen priors the posterior can contract with an optimal rate uniformly over the sparse space of the true mean. Also, compared to Empirical Bayes Procedure, the fully Bayesian procedure avoids from overly exploiting the data.

But this method has its weakness. Firstly, since it needs the computation of summation over subset S with size p , it is computationally expensive as it requires at least $n \log^2 n$ operations in this step as authors stated in the paper. when the size of data becomes much larger than just hundreds, it is much more time consuming compared to the fast Empirical Bayes procedure. Also, as stated in the theorems and observed from simulations, this procedure requires an appropriate choice of priors to guarantee an optimal estimator for the true mean. Otherwise, dimension of nonzero in estimators can be far from the truth and the risk will be large.

References

- [1] Ismaël Castillo and Aad van der Vaart. Needles and Straw in a Haystack: Posterior concentration for possibly sparse sequences. *The Annals of Statistics*, 40(4):2069 – 2101, 2012. [1](#), [2](#), [3](#)
- [2] Iain M. Johnstone and Bernard W. Silverman. Needles and straw in haystacks: Empirical Bayes estimates of possibly sparse sequences. *The Annals of Statistics*, 32(4):1594 – 1649, 2004. [5](#)