A short note on the Bernstein-von Mises theorem under infinite dimensional case

Yen-Chi Chen
University of Washington
July 1, 2020

The main reference of this short note is

> [F1999] Freedman, D. (1999). Wald Lecture: On the Bernstein-von Mises theorem with infinite-dimensional parameters. The Annals of Statistics, 27(4), 1119-1141.

Consider a simple normal means model on an infinite basis coefficient problems where we observe an infinite sequence

$$\mathbb{Y}_n = \{Y_1, Y_2, \cdots\}$$

such that each observation is drawn independently from the following model:

$$Y_\ell = \theta_\ell + \varepsilon_\ell/\sqrt{n}$$

such that $\varepsilon_1, \varepsilon_2, \cdots$ are IID $N(0,1)$. Note that the object $\mathbb{Y}_n$ is an infinite length vector; the index n describes the variance of the noise. This model occurs in a nonparametric regression or density estimation with orthonormal basis or wavelet estimators. In the usual (Frequentist) scenario, we assume that the coefficients $\theta_1, \theta_2, \cdots$ are some fixed and unknown parameters.

A Bayesian approach to this model is to place priors on the coefficients. Here we consider a very simple model where we assume each coefficient are independently from the following prior:

$$\theta_\ell \sim N(0, \lambda_\ell).$$

Namely, $\mathsf{Var}(\theta_\ell) = \lambda_\ell$ and $\lambda_1, \lambda_2, \cdots$ is a decreasing sequence.

One can easily verify that when we observe $\mathbb{Y}_n$, the posterior of each coefficient becomes

$$\begin{aligned} \theta_\ell | \mathbb{Y}_n &\sim N\left(W_{n,\ell} Y_\ell, \frac{\lambda_\ell}{1+n\lambda_\ell}\right), \\ W_{n,\ell} &= \frac{n\lambda_\ell}{1+n\lambda_\ell}. \end{aligned} \tag{1}$$

Suppose we use the posterior mean as a Bayesian estimator of each coefficient, then we obtain

$$\widehat{\theta}_\ell = W_{n,\ell} Y_\ell = \frac{n\lambda_\ell}{1+n\lambda_\ell} Y_\ell. \tag{2}$$

This estimator can be viewed as a particular shrinkage estimator because the usual Frequentist estimator will be $\widehat{\theta}_{\ell,\mathsf{freq}} = Y_\ell$.

In this note, we will focus on the L_2 error of estimating all coefficients. Namely,

$$R_n = \sum_{\ell=1}^{\infty} |\widehat{\theta}_\ell - \theta_\ell|^2.$$

We will consider two decay rate of the sequence $\{\lambda_1, \lambda_2, \cdots\}$: a polynomial rate and an exponential rate. For simplicity, the polynomial rate we consider is

$$\lambda_\ell = \ell^{-\alpha} \tag{3}$$

and the exponential rate is

$$\lambda_\ell = \exp(-\beta\ell) \tag{4}$$

for each $\ell = 1, 2, 3, \cdots$.

We will show some interesting results:

- **L_2 errors R_n (Section 1).** When the goal is to infer R_n, the credible interval will not have the Frequentist coverage under the polynomial decay rate in general when the true parameter is sampled from the prior. However, the credible interval will have the Frequentist coverage when the decay rate of prior is exponential (and the true parameter is from the prior).
- **Faster decay rate (Section 2).** Suppose that the true parameter is drawn from a distribution that has a faster decay rate then the prior, the credible interval has the Frequentist coverage.

1 Analysis on the L_2 error

1.1 Polynomial decay

We first consider a polynomial decay, i.e., $\lambda_\ell = \ell^{-\alpha}$ as equation (3).

Bayesian analysis. From a Bayesian perspective, the credible interval is constructed using the distribution of $R_n|\mathbb{Y}_n$. Using the techniques in [F1999] (Theorem 1; the key idea is the Lindeberg-Feller's triangular array central limit theorem), one can show that when the decay rate of the prior λ_ℓ is either polynomial or exponential,

$$R_n|\mathbb{Y} \approx N(\mu(\mathbb{Y}_n), \sigma^2(\mathbb{Y}_n)).$$

Thus, the credible interval will be constructed using

$$C_{n,1-\alpha} = [\mu(\mathbb{Y}_n) - z_{1-\alpha/2}\sigma(\mathbb{Y}_n), \mu(\mathbb{Y}_n) + z_{1-\alpha/2}\sigma(\mathbb{Y}_n)],$$

where z_α is the α quantile of a standard normal.

Now we investigate $\mu(\mathbb{Y}_n)$ and $\sigma(\mathbb{Y}_n)$.

$$\begin{aligned}
\mu(\mathbb{Y}_n) &= \mathbb{E}[R_n|\mathbb{Y}_n] \\
&= \sum_{\ell=1}^{\infty} \mathbb{E}[|\widehat{\theta}_\ell - \theta_\ell|^2|\mathbb{Y}_n] \\
&= \sum_{\ell=1}^{\infty} \mathsf{Var}\left(\theta_\ell^2|\mathbb{Y}_n\right) \quad \text{since } \widehat{\theta}_\ell \text{ is the posterior mean} \\
&\overset{(1)}{=} \sum_{\ell=1}^{\infty} \frac{\lambda_\ell}{1+n\lambda_\ell} = \sum_{\ell=1}^{\infty} \frac{1}{\lambda_\ell^{-1}+n} \\
&= \sum_{\ell=1}^{\infty} \frac{1}{\ell^\alpha+n} \approx \int_0^\infty \frac{1}{\ell^\alpha+n} d\ell \\
&= n^{-1+1/\alpha} \int \frac{1}{s^\alpha+1} ds = n^{-1+1/\alpha} \xi_\alpha,
\end{aligned} \tag{5}$$

where ξ_α is a finite constant. The $\approx$ in the above equation is due to the change of variable $d\ell = n^{1/\alpha} ds$. Interestingly, the conditional mean actually does not depend on the observed data $\mathbb{Y}_n$.

Now we turn to the conditional variance. Note that a key trick is that the 4-th centered moment of $N(\mu, \sigma^2)$ is $3\sigma^4$. Thus,

$$\begin{aligned}
\sigma^2(\mathbb{Y}_n) &= \mathsf{Var}\left(\sum_{\ell=1}^{\infty} (\widehat{\theta}_\ell - \theta_\ell)^2|\mathbb{Y}_n\right) \\
&= \sum_{\ell=1}^{\infty} \mathsf{Var}\left((\widehat{\theta}_\ell - \theta_\ell)^2|\mathbb{Y}_n\right) \\
&= \sum_{\ell=1}^{\infty} \mathbb{E}\left((\widehat{\theta}_\ell - \theta_\ell)^4|\mathbb{Y}_n\right) - \mathbb{E}^2\left((\widehat{\theta}_\ell - \theta_\ell)^2|\mathbb{Y}_n\right) \\
&= \sum_{\ell=1}^{\infty} 2\mathsf{Var}^2\left(\theta_\ell|\mathbb{Y}_n\right) \\
&= 2\sum_{\ell=1}^{\infty} \left(\frac{\lambda_\ell}{1+n\lambda_\ell}\right)^2 \\
&\approx 2\int_0^\infty \frac{1}{(\lambda_\ell^{-1}+n)^2} d\ell.
\end{aligned} \tag{6}$$

Again, using the change of variable $d\ell = n^{1/\alpha} ds$ and the polynomial decay rate $\lambda_\ell = \ell^{-\alpha}$, we obtain

$$\sigma^2(\mathbb{Y}_n) \approx n^{-2+1/\alpha} \underbrace{2\int_0^\infty \frac{1}{(s^\alpha+1)^2} ds}_{\eta_\alpha^2}.$$

To sum up,

$$R_n|\mathbb{Y} \approx N(n^{-1+1/\alpha} \cdot \xi_\alpha,\ n^{-2+1/\alpha} \cdot \eta_\alpha^2). \tag{7}$$

Frequentist analysis. From a Frequentist perspective, the randomness comes from the data $\mathbb{Y}_n$. Thus, a Frequentists confidence interval will be focusing on the distribution of R_n where the randomness is from $\mathbb{Y}_n$

and β is assumed to be fixed. Using the fact that $\widehat{\theta}_\ell = W_{n,\ell} Y_\ell$ and $Y_\ell = \theta_\ell + \frac{\varepsilon_\ell}{\sqrt{n}}$, we expand

$$\begin{aligned} R_n &= \sum_{\ell=1}^{\infty} |\widehat{\theta}_\ell - \theta_\ell|^2 = \sum_{\ell=1}^{\infty} |W_{n,\ell} Y_\ell - \theta_\ell|^2 \\ &= \sum_{\ell=1}^{\infty} \left| (W_{n,\ell} - 1)\theta_\ell + W_{n,\ell} \frac{\varepsilon_\ell}{\sqrt{n}} \right|^2 \\ &= \sum_{\ell=1}^{\infty} (1 - W_{n,\ell})^2 \theta_\ell^2 - 2W_{n,\ell}(1 - W_{n,\ell}) \frac{\theta_\ell \varepsilon_\ell}{\sqrt{n}} + W_{n,\ell}^2 \frac{\varepsilon_\ell^2}{n}. \end{aligned}$$

A remarkable trick that was used in [F1999] (proof of Theorem 2) was to plus and minus $(1 - W_{n,\ell})^2 \lambda_\ell$ and $W_{n,\ell}^2 / n$, leading it to

$$R_n = \sum_{\ell=1}^{\infty} \underbrace{(1 - W_{n,\ell})^2 \lambda_\ell + W_{n,\ell}^2 / n}_{A_n} + \underbrace{(1 - W_{n,\ell})^2 (\theta_\ell^2 - \lambda_\ell)}_{B_n} + \underbrace{-2W_{n,\ell}(1 - W_{n,\ell}) \frac{\theta_\ell \varepsilon_\ell}{\sqrt{n}} + W_{n,\ell}^2 \frac{\varepsilon_\ell^2 - 1}{n}}_{C_n}. \tag{8}$$

The reason of this decomposition is that term A_n can be shown to be

$$A_n = n^{-1+1/\alpha} \cdot \xi_\alpha = \mu(\mathbb{Y}_n)$$

so it corresponds to the conditional mean of the posterior.

Both term B_n and C_n depends on the true parameter $\{\theta_\ell\}$. Now we consider the scenario where *prior distribution is a correct model*, i.e., the parameter θ_ℓ is indeed generated from a normal distribution with mean 0 and variance λ_ℓ.

In this case, term C_n can be shown to be a quantity converging to a normal distribution with mean 0 and variance $n^{-2+1/\alpha} \cdot \eta_\alpha^2 = \sigma^2(\mathbb{Y}_n)$, which corresponds to the variance of the of the posterior.

Thus, what is different is the term $B_n = \sum_\ell (1 - W_{n,\ell})^2 (\theta_\ell^2 - \lambda_\ell)$, which was called *Bayes bias* in [F1999]. If this term disappears, then the credible interval has the correct Frequentist coverage. Although one can easily verify that $\mathbb{E} B_n = 0$, its variance is not:

$$\mathsf{Var}(B_n) = \sum_\ell (1 - W_{n,\ell})^4 \mathsf{Var}(\theta_\ell^2 - \lambda_\ell) = \sum_\ell (1 - W_{n,\ell})^4 \lambda_\ell^2 \mathsf{Var}(Z_\ell^2 - 1) = 2 \sum_\ell (1 - W_{n,\ell})^4 \lambda_\ell^2,$$

where $Z_1, Z_2, \cdots,$ are IID standard normal. Thus, the variance is

$$\mathsf{Var}(B_n) = 2 \sum_\ell \frac{\lambda_\ell^2}{(1 + n\lambda_\ell)^4} = 2 \int_0^\infty \frac{\ell^{2\alpha}}{(n + \ell^\alpha)^4} d\ell \approx n^{-2+\alpha} \underbrace{\int_0^\infty \frac{s^{2\alpha}}{1 + s^\alpha} ds}_{\eta_{\alpha,1}^2}.$$

As a result, the frequentist distribution of R_n is

$$R_n = A_n + B_n + C_n,$$

where

$$\begin{aligned} A_n &= n^{-1+1/\alpha} \cdot \xi_\alpha \\ B_n &= n^{-2+1/\alpha} \cdot \eta_{\alpha,1}^2 Z_1 \\ C_n &= n^{-2+1/\alpha} \cdot \eta_\alpha^2 Z_2, \end{aligned}$$

where Z_1 and Z_2 are independent standard normal. Thus, the credible interval in equation (7) is offset by the amount of B_n, a random offset. So the credible interval does not have the Frequentist coverage.

1.2 Exponential decay

Suppose that λ_ℓ follows the exponential decay in equation (4) $\lambda_\ell = \exp(-\beta\ell)$. We will make good use of the following lemma from [F1999]:

Lemma 1 (Lemma 5 of Freedman 1999) *Let $a, b, c > 0$ be positive constants and $ab > c$. Then*

- $\sum_{\ell=1}^{\infty} \frac{1}{(n+e^{a\ell})^b} \sim \frac{\log n}{an^b}$
- $\sum_{\ell=1}^{\infty} \frac{e^{c\ell}}{(n+e^{a\ell})^b} \sim n^{-b+c/a}$

Bayesian analysis. Recall that the posterior will be a normal distribution with mean $\mu(\mathbb{Y}_n)$ and variance $\sigma^2(\mathbb{Y}_n)$. We have to modify both terms due to the change of prior distribution. Using Lemma 1, we can show the mean to be

$$\begin{aligned} \mu(\mathbb{Y}_n) &= \sum_{\ell=1}^{n} \frac{1}{\lambda_\ell^{-1}+n} \\ &= \sum_{\ell=1}^{n} \frac{1}{e^{\alpha\ell}+n} \\ &\sim \frac{\log n}{n} \end{aligned}$$

and the variance will be

$$\begin{aligned} \sigma^2(\mathbb{Y}_n) &= 2\sum_{\ell=1}^{n} \left(\frac{1}{\lambda_\ell^{-1}+n}\right)^2 \\ &= \sum_{\ell=1}^{n} \frac{1}{(e^{\alpha\ell}+n})^2 \\ &\sim \frac{\log n}{n^2} \end{aligned}$$

Frequentist analysis. Now we turn to the analysis of Frequentist distribution of R_n. Note that we still have $A_n = \mu(\mathbb{Y}_n) = O_P(\frac{\log n}{n})$ and $C_n \approx \sigma(\mathbb{Y}_n) = O_P(\sqrt{\frac{\log n}{n^2}})$ so what we need to focus is the Bayes bias term B_n.

Note that

$$B_n = \sum_{\ell=1}^{\infty}(1-W_{n,\ell})^2(\theta_\ell^2 - \lambda_\ell)$$

has mean 0 and variance

$$\begin{aligned}
\mathsf{Var}(B_n) &= 2\sum_\ell \frac{\lambda_\ell^2}{(1+n\lambda_\ell)^4}\\
&= 2\sum_\ell \frac{\lambda_\ell^{-2}}{(\lambda_\ell^{-1}+n)^4}\\
&= 2\sum_\ell \frac{e^{2\alpha\ell}}{(e^{\alpha\ell}+n)^4}\\
&\overset{\text{Lemma 1}}{\sim} n^{-4+2} = n^{-2}.
\end{aligned}$$

As a result, we conclude that the Frequentist distribution

$$\begin{aligned}
&R_n = A_n + B_n + C_n\\
&A_n = \mu(\mathbb{Y}_n) = O_P\left(\frac{\log n}{n}\right)\\
&B_n = O_P(n^{-1})\\
&C_n \approx \sigma(\mathbb{Y}_n) = O_P\left(\sqrt{\frac{\log n}{n^2}}\right)
\end{aligned}$$

Since $\frac{B_n}{\sigma(\mathbb{Y}_n)} = O_P\left(\frac{1}{\sqrt{\log n}}\right) = o_P(1)$, the Bayes bias is asymptotically negligible so the credible interval has Frequentist coverage.

Remark. Although $\frac{B_n}{\sigma(\mathbb{Y}_n)} = O_P\left(\frac{1}{\sqrt{\log n}}\right) = o_P(1)$ converges to 0, it is a very very slow term. The rate is $\sqrt{\log n}$! So in practice, we need an extremely large sample size to see its convergence. For some people, a rate of this order will not be viewed as convergence since the required sample size is too huge.

2 Faster decay rate

Now we go back to the polynomial decay rate and study an interesting question: if our prior is a polynomial $\ell^{-\alpha}$ but the true parameter is drawn from a distribution with polynomial $\ell^{-\beta}$, will the credible interval has the asymptotic coverage when $\beta > \alpha$?

The answer is yes! If the true parameter is drawn from a distribution with a faster rate, we do have the correct coverage (in fact, we have overcoverage).

The main reference of this section is the following paper:

> [L2011] Leahu, H. (2011). On the Bernstein-von Mises phenomenon in the Gaussian white noise model. Electronic journal of statistics, 5, 373-404.

Consider the following two classes:

$$\mathcal{P}_\beta = \{\theta : \theta_\ell \asymp \ell^{-\beta}\}$$
$$\mathcal{S}_\beta = \left\{\theta : \sum_{\ell=1}^{\infty} \ell^{\beta-1}\theta_\ell < \infty\right\}.$$

The first one $\mathcal{P}_\beta$ is the polynomial decay rate class and the second one $\mathcal{S}_\beta$ is something related to the so-called Sobolev ball. The two classes have the following relation:

$$\mathcal{P}_\beta \cap \mathcal{S}_\beta = \emptyset, \quad \mathcal{P}_\beta \cap \mathcal{S}_{\beta+\varepsilon} = \emptyset, \quad \mathcal{P}_{\beta+\varepsilon} \subset \mathcal{S}_\beta, \quad \mathcal{P}_\beta \subset \mathcal{S}_{\beta+\varepsilon},$$

for any $\varepsilon > 0$.

Recall that we assume the prior $\lambda_\ell \asymp \ell^{-\beta}$ and we construct a credible interval of R_n using

$$C_{n,1-\alpha} = [\mu(\mathbb{Y}_n) - z_{1-\alpha/2}\sigma(\mathbb{Y}_n), \mu(\mathbb{Y}_n) + z_{1-\alpha/2}\sigma(\mathbb{Y}_n)],$$

where $\mu(\mathbb{Y}_n)$ and $\sigma(\mathbb{Y}_n)$ are defined in equation (7). Then we have the following results, which are revised from Section 3 of [L2011]:

- **Class $\mathcal{P}_\beta$.** If $\theta \in \mathcal{P}_\beta$ such that
  - $\beta > \alpha$ (a.k.a. undersmoothing), then $\inf_{\theta\in\mathcal{P}_\beta} P(\theta \in C_{n,1-\alpha}) \to 1$.
  - $\beta = \alpha$, then $\inf_{\theta\in\mathcal{P}_\beta} P(\theta \in C_{n,1-\alpha}) = 0, \sup_{\theta\in\mathcal{P}_\beta} P(\theta \in C_{n,1-\alpha}) = 1$; the actual coverage depends on the true parameter (determined by the B_n term in Section 1.1).
  - $\beta < \alpha$ (a.k.a. oversmoothing), then $\sup_{\theta\in\mathcal{P}_\beta} P(\theta \in C_{n,1-\alpha}) \to 0$.
- **Class $\mathcal{S}_\beta$.** If $\theta \in \mathcal{S}_\beta$ such that
  - $\beta \geq \alpha$ (a.k.a. undersmoothing), then $\inf_{\theta\in\mathcal{S}_\beta} P(\theta \in C_{n,1-\alpha}) \to 1$.
  - $\beta < \alpha$ (a.k.a. oversmoothing), then $\sup_{\theta\in\mathcal{S}_\beta} P(\theta \in C_{n,1-\alpha}) \to 0$.

Thus, as long as we are undersmoothing (i.e., the prior has a slower decay rate than the true parameter), the credible interval has the enough coverage (though it has over-coverage). An interesting note is that the two classes $\mathcal{P}_\beta$ and $\mathcal{S}_\beta$ have distinct results when $\beta = \alpha$–the polynomial decay class $\mathcal{P}_\alpha$ may not have enough coverage while the Sobolev-type class $\mathcal{S}_\beta$ still have enough coverage.

3 Remarks

- **A sad news on the choice of prior.** This analysis shows that if the true parameter is generated from our prior distribution, unfortunately, the credible interval will not have enough coverage in general.

  The analysis in Section 2 shows that if the true parameter lies in an undersmoothing class (the decay rate is faster than the prior, i.e., $\mathcal{P}_{\alpha-\varepsilon}$ for some $\varepsilon > 0$), then we do have the coverage. However, suppose we sample θ_ℓ from the prior distribution $N(0, \lambda_\ell = \ell^{-\alpha})$, the probability $P(\theta \in \mathcal{P}_{\alpha-\varepsilon}) = 0$. Namely, there is no chance that we would obtain a set of parameter that we can have an overwhelming coverage. In conclusion, I would quote the following statement from the abstract of [L2011]:

> *The overall conclusion is that, unlike in the parametric setup, positive results regarding frequentist probability coverage of credible sets can only be obtained if the prior assigns null mass to the parameter space.*

In other words, to use a Bayesian approach as a Frequentist method and suppose that we know the true decay rate is $\ell^{-\alpha}$, we have to purposely choose the prior to be slower than $\ell^{-\alpha}$, which is a bizarre situation.

- **A note on the derivation of Section 2.** Here is a note on the derivation of Section 2. Recalled from equation (8), we can decompose $R_n = A_n + B_n + C_n$. In the analysis of a faster decay rate, we will combine $A_n + B_n$ and jointly analyze them and deal with C_n separately:

$$\begin{aligned} A_n + B_n &= \sum_{\ell=1}^{\infty} \frac{W_{n,\ell}^2}{n} + (1 - W_{n,\ell})^2 \theta_\ell^2, \\ C_n &= -2W_{n,\ell}(1 - W_{n,\ell}) \frac{\theta_\ell \varepsilon_\ell}{\sqrt{n}} + W_{n,\ell}^2 \frac{\varepsilon_\ell^2 - 1}{n}. \end{aligned}$$

Using the fact that $W_{n,\ell} = \frac{n\lambda_\ell}{1+n\lambda_\ell}$ and $\lambda_\ell = \ell^{-\alpha}$ and the two terms in C_n are uncorrelated (θ_ℓ has mean 0 and is independent of ε_ℓ), we conclude that

$$\begin{aligned} A_n + B_n &= \sum_{\ell=1}^{\infty} \frac{n\lambda_\ell^2}{(1+n\lambda_\ell)^2} + \frac{\theta_\ell^2}{(1+n\lambda_\ell)^2} \\ &= \sum_{\ell=1}^{\infty} \frac{n}{(\ell^\alpha + n)^2} + \frac{\ell^{2\alpha}\theta_\ell^2}{(\ell^\alpha + n)^2} \\ \mathsf{Var}(C_n) &= \sum_{\ell=1}^{\infty} 4W_{n,\ell}^2 (1 - W_{n,\ell})^2 \frac{\theta_\ell^2}{n} + 2\frac{W_{n,\ell}^4}{n^2} \\ &= \sum_{\ell=1}^{\infty} \frac{4n\ell^{2\alpha}\theta_\ell^2}{(\ell^\alpha + n)^4} + \frac{2n^2}{(\ell^\alpha + n)^4}. \end{aligned}$$

Note: $A_n + B_n$ equals the quantity $M_n + Q_n(\theta)$ in [L2011] and $\mathsf{Var}(C_n)$ equals $\mathsf{Var}(Z_n(\theta, \varepsilon))$ in [L2011] (both are just right below Lemma 3 of [L2011]). With these equalities, one can work out the results in Section 2.