Frontiers of Statistics: Contraction theory for posterior distributions **Spring 2019**

Lecture 10: April 2

Lecturer: Anirban Bhattacharya & Debdeep Pati *Scribes: scribe-names*

Note: *LaTeX template courtesy of UC Berkeley EECS dept & CMU's convex optimization course taught by Ryan Tibshirani.*

Review: Main message from the risk bound for posterior fractional posterior: show the prior gives "enough" mass to an appropriate KL-neighborhood of the truth,

$$\Pi_n[B(\theta^*, \epsilon_n, \theta_0)] \geq e^{-C_n \epsilon_n^2}.$$

Next we check the prior mass condition.

10.1 Prior mass condition in the sparse context

Consider the true parameter $\theta_0 \in \ell_0[s;p]$ (nearly black vector). Let $\ell_0[s,p] = \{\theta \in \mathbb{R}^p : \#(1 \leq i \leq p : \theta_i \neq 0) \leq s\}$. Suppose $\theta \sim \Pi_n$ on $\mathbb{R}^p$, we are interested in the lower bound of $P[\|\theta - \theta_0\| < \epsilon]$.

For the Gaussian regression model we have $B_n(\theta, \epsilon_n, \theta_0) \supset \{\|\theta - \theta_0\| < \epsilon_n\}$. If $\theta_0 \in \ell_0[s,p]$, we want to show

$$P[\|\theta - \theta_0\| < \epsilon] \geq e^{-s \log p} c_\epsilon,$$

where c_ϵ denotes some term involving $\log(1/\epsilon)$.

Remark: For the sparse mean model, consider $Y \sim N(\theta, I_p)$. The minimax rate for $\ell_0[s,p]$ in Euclidean norm is $2s\log(p/s)$.

Example. Consider $\theta_j \overset{\text{i.i.d.}}{\sim} N(0,1)$. We can show that $P[\|\theta\| < \epsilon] \leq e^{-Cp\log(1/\epsilon)}$.
By Anderson inequality,

$$P(\|\theta - \theta_0\| < \epsilon) \leq P(\|\theta\| < \epsilon) \leq e^{-Cp}.$$

The inequality holds since $\|\theta\| \sim \chi_p^2$. When p is large, the whole distribution shifts to the right side of real line. As p increases, the CI can not contain the origin anymore.

Variable selection prior:

1. Pick subset $K \sim \Pi_K$ on $\{0, 1, 2, \ldots, p\}$, Π_K can be uniform
2. Pick a subset S uniformly out of the $\binom{p}{K}$ subsets of size K,
3. Set $\theta_j = 0$ for any $j \in S^c$ and $\theta_j \sim g$ for $j \in S$,

where g is a density on $\mathbb{R}$ such as $N(0,1)$, Laplace, Cauchy.
Exercise: Suppose $\theta_j \mid w \sim (1-w)\delta_0 + wg(\cdot)$ and $w \sim U(0,1)$. Find the marginal prior on θ and write it as a subset prior.
Now we state the sketch of prior concentration for subset priors.

Proof. Fix $\theta_0 \in \ell_0[s;p]$. Let S_0 denote the subset consisting of the non-zero parameters satisfying $|S_0| < s$.

$$\begin{aligned}P[\|\theta - \theta_0\| < \epsilon] &\geq P[\|\theta - \theta_0\| < \epsilon \mid K = s, S = S_0]\, P(K = s)\, P(S = S_0 \mid K = s) \\ &\geq \frac{1}{p+1}\,\frac{1}{\binom{p}{s}} \geq e^{-\log(p+1)}\, e^{-s\log(pe/s)} \geq e^{-Cs\log(p/s)}.\end{aligned}$$

The second line above holds since

$$P[\|\theta - \theta_0\| < \epsilon \mid K = s, S = S_0] \geq P(\chi_s^2 < \epsilon)\, e^{-\|\theta_0\|^2/2},$$

which does not depend on p. If $s \leq \lceil p/2 \rceil$, then $(p/s)^s \leq \binom{p}{s} \leq (pe/s)^s$. See [CV12] for more details. $\square$

Remark: Heavy tail g prior is needed to bound arbitrarily large θ_0.

10.1.1 Global-local continuous shrinkage priors

Consider the global-local continuous shrinkage prior,

$$\begin{aligned}\theta_j \mid \lambda_j, \tau &\sim N(0, \lambda_j^2\tau^2) \\ \lambda_j &\overset{\text{i.i.d.}}{\sim} f \\ \tau &\sim g\end{aligned}$$

Remark:

1. If $\lambda_j \sim \exp(1/2)$ it corresponds to Bayesian lasso prior, where the marginal density $p(\theta_j) \approx \exp\{-|\theta_j|/(2\tau)\}$.
2. The prior concentration of the Bayesian lasso is slightly better than the iid $N(0,1)$ priors (Bayesian shrinkage). For Dirichlet-Laplace prior and horseshoe prior the contraction rate holds.

Sketch: Suppose $\theta_0 \in \ell_0[s;p]$, let S_0 denote the sunset of non-zeros.

$$\begin{aligned}P[\|\theta - \theta_0\| \leq \epsilon] &= \int_0^\infty P[\|\theta - \theta_0\| \leq \epsilon|\tau]\, g(\tau)d\tau \\ &\geq \int_0^\infty P\left[\sum_{j\in S_0}(\theta_j - \theta_{0j})^2 < \epsilon/2 \mid \tau\right] P\left[\sum_{j\in S_0^c}\theta_j^2 < \epsilon/2 \mid \theta\right] g(\tau)d\tau \\ &\geq \int_{\tau\in[a/p,b/p]} P\left[\sum_{j\in S_0}(\theta_j - \theta_{0j})^2 < \epsilon/2 \mid \tau\right] P\left[\sum_{j\in S_0^c}\theta_j^2 < \epsilon/2 \mid \theta\right] g(\tau)d\tau,\end{aligned}$$

Since $\|\theta - \theta_0\|^2 = \sum_{j\in S_0}(\theta_j - \theta_{0j})^2 + \sum_{j\in S_0^c}\theta_j^2$.

10.1.2 Extension of the theory to variational Bayes

Recall

$$\hat{q} = \operatorname*{argmin}_{q\in\Gamma} D(q \,||\, \Pi_{n,\alpha}(\cdot|x^{(n)}))$$

where Γ denotes the variational family. Consider the mean field: $q = q_1 \times q_2 \times \cdots \times q_d$.
Question: Does $\hat{q}$ have the first order optimality (minimax rates)? (More details see [YPB17]).
How is it related to fractional?

$$\begin{aligned} D(q \,||\, \Pi_{n,\alpha}) &= -\int q(\theta) \log \frac{\Pi_{n,\alpha}(\theta)}{q(\theta)} d\theta \\ &= \int \alpha\, \gamma_n(\theta, \theta_0)\, q(\theta) d\theta + D(q||\Pi) + \log m_\alpha. \end{aligned}$$

Minimizing $D(q \,||\, \Pi_{n,\alpha})$ is equivalent to minimizing $\int \alpha\, \gamma_n(\theta, \theta_0)\, q(\theta) d\theta + D(q||\Pi)$. Since $q(\theta) \propto \Pi(\theta) \mathbb{1}_{B_n}(\theta)$, the problem is that $q(\theta)$ may not be in Γ. It cannot be written as the product of factors.

Main idea: For $\theta = (\theta_1, \theta_2)$, Find the rectangular subset of B_n such that $B_n \supseteq \mathcal{N}_1 \times \mathcal{N}_2$.

Theorem (for VB): Under certain conditions, $\int D_\alpha^{(n)}(\theta, \theta_0) \hat{q}(\theta) d\theta$ is of the order of the minimax rate, variational point estimate is minimax optimal.

References

[CV12] I. Castillo and A. van der Vaart, "Needles and straw in a haystack: Posterior concentration for possibly sparse sequences," *The Annals of Statistics*, 2012, pp. 2069–2101.

[YPB17] Y. Yang, D. Pati and A. Bhattacharya, "α-Variational Inference with Statistical Guarantees," *arXiv preprint arXiv:1710.03266*, 2017.