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

Lecture 6: March 7

Lecturer: Anirban Bhattacharya & Debdeep Pati *Scribes: Brian Kidd & Yabo Niu*

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

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

6.1 Three sufficient conditions

Let $\mathcal{P}_n \subseteq \mathcal{P}$ denote the sieves and $N(\epsilon_n, \mathcal{P}_n, d)$ denote the covering number for those sieves. Recall that there were three conditions for the theorem proving posterior contraction ($\exists c_1, c_2 > 0$ constants):

$$\log N(\epsilon_n, \mathcal{P}_n, d) \leq c_1 n\epsilon_n^2 \tag{6.1}$$

$$\pi(p : D(p_0||p) \leq \epsilon_n^2, V(p_0||p) \leq \epsilon_n^2) \geq e^{-c_2 n\epsilon_n^2} \tag{6.2}$$

$$\pi(\mathcal{P}_n^C) \leq e^{-(c_2+4)n\epsilon_n^2} \tag{6.3}$$

If these three conditions hold, we can show that $\mathbb{E}_0\Pi[d(p, p_0) > M\epsilon_n \mid x_1, \ldots, x_n] \to 0$ as $n \to \infty$.

6.2 Historical references

- Ghoshal, Ghosh, & van der Vaart: Convergence Rates of Posterior Distributions (2000, Annals of Statistics). This paper has some other useful variants, including the idea of using "shells" for quantifying prior probabilities on increasingly large circles, see Figure 6.1.

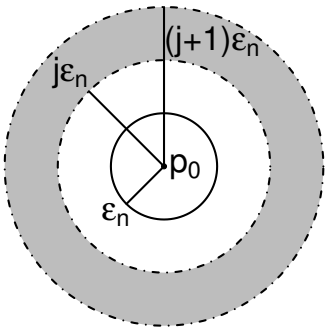

Figure 6.1: Breaking up the complement into shells.

- Shen & Wasserman: Rates of Convergence of Posterior Distributions (1999, Annals of Statistics)

- Ghoshal, Ghosh, & Ramamoorthi: Posterior consistency of Dirichlet mixtures in density estimation (1999, Annals of Statistics)
- Barron, Schervish, & Wasserman: The consistency of posterior distributions in nonparametric problems (1999, Annals of Statistics)
- Ghoshal & van der Vaart: Convergence rates of posterior distributions for noniid observations (2007, Annals of Statistics). This paper extends the ideas to non-iid (INID) cases such as regression. Some of the definition extensions include

$$p_0^{(n)} = \prod_{i=1}^{n} p_{0i}, \quad p = \prod_{i=1}^{n} p_i, \quad d(p_0^{(n)}, p^{(n)}) = \frac{1}{n}\sum_{i=1}^{n} d(p_{0i}, p_i).$$

6.3 Density Space & Parameter Space

In most practical situations, $p \equiv p_\theta$.

$$\Theta \longrightarrow \mathscr{P}$$
$$\text{mapping: } \theta \longmapsto p_\theta$$

This map above doesn't have to be 1-1. Typically we place a prior distribution Π on Θ which induces a prior Π on $\mathscr{P}$.

Example 1: DP mixture model $x_1, \ldots, x_n \mid p \overset{iid}{\sim} p$.

$$p(x) = \sum_{h=1}^{\infty} \pi_h \cdot \frac{1}{\sigma_h}\phi\Big(\frac{x-\mu_h}{\sigma_h}\Big), \quad \text{location-scale mixtures of Gaussians}$$
$$\pi_h \geq 0, \quad \sum_{h=1}^{\infty} \pi_h = 1, \quad \text{a.s.}$$
$$\pi_h = \gamma_h \cdot \Pi_{l<h}(1-\gamma_l), \quad \text{stick-breaking representation}$$
$$\gamma_l \overset{iid}{\sim} \text{Beta}(1,\alpha), \quad \mu_h \overset{iid}{\sim} \Pi_\mu, \quad \sigma_h \overset{iid}{\sim} \Pi_\sigma, \quad \text{priors}$$
$$\Pi = (\Pi_h), \quad \mu = (\mu_h), \quad \sigma = (\sigma_h), \quad p = p_\theta, \quad \theta = (\Pi, \mu, \sigma).$$

Example 2: (Regression)

$$y_i = f(x_i) + \epsilon_i, \qquad \epsilon_i \sim N(0, \sigma^2).$$

The data is of the form $(y_i, x_i)_{i=1}^n$. Two common constructions for the function f are:

- (Possibly high-dimensional) linear regression: $f(x_i) = x_i'\beta$.
- Basis function regression (with bases ψ_j): $f(x_i) = \sum_{j=1}^{K} \theta_j \psi_j(x_i)$.

6.4 Sieve construction

In most practical situations, $p \equiv p_\theta$ (the densities are parameterized). We assume there is a measureable map from $\Theta \to \mathcal{P}$. Then a prior distribution π is placed on Θ, which induces a prior π on $\mathcal{P}$. Denote the

distance metric on Θ by d_θ. For any subset $\Theta_n \subseteq \Theta$, define

$$\mathcal{P}_n = \{p_\theta : \theta \in \Theta_n\}$$
$$B(p_i, \epsilon_n) = \{p : d(p, p_i) < \epsilon_n\}.$$

We need a net on the sieves $\mathcal{P}_n$ (i.e. we need $p_1, \cdots, p_N$ s.t. $\mathcal{P}_n \subseteq \cup_{i=1}^N B(p_i, \epsilon_n)$). The idea is to find a δ_n-net in the parameter space Θ_n such that for any $\theta, \theta' \in \Theta_n$,

$$d_\theta(\theta, \theta') \leq \delta_n \implies d(p_\theta, p_{\theta'}) \leq \epsilon_n$$

Claim: Define $p_i \equiv p_{\theta_i}$. Then $\{p_1, \cdots, p_N\}$ is an ϵ_n-net of $\mathcal{P}_n$. For the complement probability, $\pi(\mathcal{P}_n^C) \leq \pi(\Theta_n^C)$.

6.5 Back to the density estimation example

$$x_1, \cdots, x_n \mid p \overset{\text{iid}}{\sim} p,$$
$$p(x) = \sum_{h=1}^{\infty} \pi_h \frac{1}{\sigma_h} \phi\left(\frac{x_i - \mu_h}{\sigma_h}\right),$$

where $\pi_h \geq 0$, $\sum \pi_h = 1$ a.s. Note that ϕ denotes the univariate normal pdf here, though the mixture model generalizes to multivariate normals and other distributions. For a prior on the (π_h), we use the common stick-breaking prior:

$$\pi_h = \nu_h \prod_{\ell<h} (1 - \nu_\ell), \qquad \nu_\ell \overset{\text{iid}}{\sim} \text{Beta}(1, \alpha).$$

To generalize to a Pitman-Yor prior, the Beta distribution should be defined with sequences α_ℓ, β_ℓ for the parameters. For the means and standard deviations, any iid prior can be used (though later we'll see the need for the prior to have exponential tails in this example). Let Π_μ, Π_σ denote these priors. Define the vectors $\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}$ to be the sequences $\{(\pi_h), (\mu_h), (\sigma_h)\}$, respectively, and

$$p \equiv p_\theta, \qquad \theta = (\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}).$$

Useful Inequalities: Let q denote a density and π denote the weights for the mixture.

$$\left\| \sum_{h=1}^{\infty} \pi_h q_h - \sum_{h=1}^{\infty} \pi'_h q_h \right\|_{TV} \leq \sum_{h=1}^{\infty} |\pi_h - \pi'_h|$$
$$\left\| \sum_{h=1}^{\infty} \pi_h q_{1h} - \sum_{h=1}^{\infty} \pi_h q_{2h} \right\|_{TV} \leq \sum_{h=1}^{\infty} \pi_h \left\| q_{1h} - q_{2h} \right\|_{TV}$$

6.5.1 Sieve construction

The construction follows Dr. Pati's paper. This construction is useful, as it generalizes much better than other distributions.

$$\Theta_n = \Big\{ (\Pi, \mu, \sigma) : \underbrace{\sum_{h=1}^{K_n} \Pi_h > (1 - \delta_n)}_{\substack{\text{first } K_n \text{ weights make up} \\ \text{most of the total mass.}}} , \mu_h \in [-m_n, m_n], \sigma_h \in [u_n, v_n], h = 1, \ldots, K_n \Big\}.$$

The first K_n weights make up most of the mass, and both the means and variances are bounded. $(m_n) \to \infty, (u_n) \to 0, (v_n) \to \infty$. Let the distance metric be the total variation metric for the density space. Let $p_\theta \equiv p_{\Pi,\mu,\sigma}$. For $\theta, \theta' \in \Theta_n$, by triangle inequality,

$$\begin{aligned}
\|p_\theta - p_{\theta'}\|_{\text{TV}} &\le \|p_{\Pi,\mu,\sigma} - p_{\Pi',\mu,\sigma}\|_{\text{TV}} + \|p_{\Pi',\mu,\sigma} - p_{\Pi',\mu',\sigma'}\|_{\text{TV}} \\
&\le \sum_{h=1}^{\infty} |\Pi_h - \Pi'_h| + \sum_{h=1}^{\infty} \Pi'_h \cdot \|N(\mu_h, \sigma_h^2) - N(\mu'_h, \sigma'^2_h)\|_{\text{TV}} \\
&\le \sum_{h=1}^{K_n} |\Pi_h - \Pi'_h| + 2\delta_n + \sum_{h=1}^{K_n} \Pi'_h \cdot \frac{|\mu_h - \mu'_h|}{\min\{\sigma_h, \sigma'_h\}} + \delta_n \\
&\le \sum_{h=1}^{K_n} |\Pi_h - \Pi'_h| + \frac{1}{u_n} \cdot \max_{h=1:K_n} \{|\mu_h - \mu'_h|\} + 3\delta_n.
\end{aligned}$$

If we can make

$$\sum_{h=1}^{K_n} |\Pi_h - \Pi'_h| < \frac{\epsilon_n}{3}, \quad \frac{1}{u_n} \cdot \max_{h=1:K_n} |\mu_h - \mu'_h| < \frac{\epsilon_n}{3}, \quad \delta_n = \frac{\epsilon_n}{9},$$

then $\|p_\theta - p_{\theta'}\|_{\text{TV}} < \epsilon_n$.

REMARKS: If a space Θ can be written as $\Theta_a \otimes \Theta_b$ and $\theta = (\theta_a, \theta_b)$. Assume $d(\theta, \theta') \le d_a(\theta_a, \theta'_a) + d_b(\theta_b, \theta'_b)$. Let $\{\theta_{1a}, \ldots, \theta_{Na}\}$ be an $\epsilon_n/2$-net of Θ_a and $\{\theta_{1b}, \ldots, \theta_{Mb}\}$ be an $\epsilon_n/2$-net of Θ_b. Then $\{\theta_i = (\theta_{ia}, \theta_{jb}) : i = 1, \ldots, N; j = 1, \ldots, M\}$ is an ϵ-net of Θ with covering number $\le MN$.

Let's look at the complement probability. By using Bonferroni inequality,

$$\begin{aligned}
\Pi(\Theta_n^c) &= \Pi\left(\left\{\sum_{h=1}^{K_n} \Pi_h > (1-\delta_n)\right\}^c \cup \bigcup_{h=1}^{K_n} \left\{\mu_h \in [-m_n, m_n]\right\}^c \cup \bigcup_{h=1}^{K_n} \left\{\sigma_h \in [u_n, v_n]\right\}^c\right) \\
&\le \Pi\left(\left\{\sum_{h=1}^{K_n} \Pi_h > (1-\delta_n)\right\}^c\right) + \Pi\left(\bigcup_{h=1}^{K_n} \left\{\mu_h \in [-m_n, m_n]\right\}^c\right) + \Pi\left(\bigcup_{h=1}^{K_n} \left\{\sigma_h \in [u_n, v_n]\right\}^c\right) \\
&\le \Pi\left(\sum_{h=K_n+1}^{\infty} \Pi_h > \delta_n\right) + K_n \cdot \Big[\Pi\big(\mu_1 \notin [-m_n, m_n]\big) + \Pi(\sigma_1 \notin [u_n, v_n])\Big].
\end{aligned}$$

Result/Verify:

$$\begin{aligned}
\sum_{h=K+1}^{\infty} \Pi_h &= 1 - \sum_{h=1}^{K} \Pi_h \\
&= 1 - \sum_{h=1}^{K} \Big\{\gamma_h \prod_{l=1}^{h-1} (1-\gamma_l)\Big\} \\
&= (1-\gamma_1) - \gamma_2(1-\gamma_1) - \gamma_3(1-\gamma_1)(1-\gamma_2) - \ldots - \gamma_K(1-\gamma_1)\cdots(1-\gamma_{K-1}) \\
&= (1-\gamma_1)(1-\gamma_2) - \gamma_3(1-\gamma_1)(1-\gamma_2) - \ldots - \gamma_K(1-\gamma_1)\cdots(1-\gamma_{K-1}) \\
&\cdots\cdots\cdots \\
&= (1-\gamma_1)\cdots(1-\gamma_K)
\end{aligned}$$

Thus, $\sum_{h=K+1}^{\infty} \Pi_h = (1-\gamma_1)\cdots(1-\gamma_K)$.

By using the result above,

$$\begin{aligned}
\Pi\left(\sum_{h=K+1}^{\infty}\Pi_h > \delta\right) &= \Pi\left(\sum_{h=1}^{K}\underbrace{-\log(1-\gamma_h)}_{\text{Gamma}(1,\alpha)} < \log\left(\frac{1}{\delta}\right)\right) \\
&= \Pi\left(\text{Gamma}(K,\alpha) < \log\left(\frac{1}{\delta}\right)\right) \\
&= \text{exponentially small.}
\end{aligned}$$