BvM theorem – basics

Anirban Bhattacharya

http://www.stat.tamu.edu/~anirbanb

March 8, 2018

1 Notations

For two sequences a_n, b_n, $a_n \approx b_n$ means $|a_n/b_n - 1| \to 0$ as $n \to \infty$. We write $a_n = O(b_n)$ if there is a constant $C > 0$ such that $a_n \leq C b_n$ for all large n; $a_n = o(b_n)$ means $a_n/b_n \to 0$.

2 Distances between probability measures

We recall some important distances between probability measures and their interrelations. Let P and Q be two probability measures. Let μ be a dominating measure so that $P, Q \ll \mu$; denote p and q to be the Radon–Nikodym derivatives $dP/d\mu$ and $dQ/d\mu$.

The **total variation distance** $\|\cdot\|_{TV}$ between P and Q is

$$\|P - Q\|_{TV} := \sup_A |P(A) - Q(A)| = \frac{1}{2}\int |p - q| d\mu = \int_{p>q} (p - q) d\mu.$$

The squared **Hellinger distance** $H^2(P, Q)$ is

$$H^2(P, Q) := \frac{1}{2}\int (\sqrt{p} - \sqrt{q})^2 d\mu = 1 - A(P, Q),$$

where $A(P, Q) = \int \sqrt{pq}\, d\mu$ is the **Hellinger affinity**.

The **Kullback–Leibler divergence** $D(P \,||\, Q) = \int p \log(p/q)\, d\mu$. Similarly, $D(Q \,||\, P)$.

If $P \equiv N(\mu_1, \sigma_1^2)$ and $Q \equiv N(\mu_2, \sigma_2^2)$, then

$$H^2(P, Q) = 1 - \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2}}\, \exp\left(-\frac{1}{4}\frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}\right),$$

$$D(P \,||\, Q) = \frac{1}{2}\left[\left\{\frac{\sigma_1^2}{\sigma_2^2} - \log\frac{\sigma_1^2}{\sigma_2^2} - 1\right\} + \frac{(\mu_1 - \mu_2)^2}{\sigma_2^2}\right].$$

In general, there isn't a close-form expression for the TVD. However, if $\sigma_1 = \sigma_2$, then (maybe up to a factor of 2?)

$$\|P - Q\|_{TV} = 2\Phi\left(\frac{|\mu_1 - \mu_2|}{2\sigma}\right) - 1.$$

The Hellinger distance and KL divergence are more amenable to deal with product measures:

1. $D(\otimes_{i=1}^n P_i \,||\, \otimes_{i=1}^n Q_i) = \sum_{i=1}^n D(P_i, Q_i)$.

2. $H^2(\otimes_{i=1}^n P_i, \otimes_{i=1}^n Q_i) = 1 - A(P^n, Q^n) = 1 - \prod_{i=1}^n A(P_i, Q_i) = 1 - \prod_{i=1}^n \{1 - H^2(P_i, Q_i)\} \leq \sum_{i=1}^n H^2(P_i, Q_i)$ where the inequality follows from Weirstrass's inequality.

List of inequalities:

1. $2H^2(P,Q) \leq \|P - Q\|_{TV} \leq 2H(P,Q)\sqrt{1 + A(P,Q)} \leq 2\sqrt{2}H(P,Q)$.
2. $4H^2(P,Q) \leq D(P \mid\mid Q) \leq 4H^2(P,Q)(1 + \log \|f_P/f_Q\|_\infty)$.
3. **Pinsker's inequality:** $\|P - Q\|_{TV}^2 \leq D(P \mid\mid Q)/2$.

3 Bernstein–von Mises theorem

One has to exercise some care in defining the mode of convergence in Bayesian asymptotics, as we are not dealing with a random estimator, rather random probability measures. This requires defining certain metrics on the space of probability measures.

Given two probability measures P and Q on the same probability space, the total variation distance $d_{\mathrm{TV}}(P,Q) = 2\sup_A |P(A) - Q(A)|$. In particular, if P and Q have densities f_P and f_Q w.r.t. some dominating measure μ (the Lebesgue measure in most cases for us), then $d_{\mathrm{TV}}(P,Q) = \int_{-\infty}^{\infty} |f_P(x) - f_Q(x)|\mu(dx)$.

NOTE: For the discussion that follows, we take the dimension $d = 1$, i.e., $\Theta \subset \mathbb{R}$, although everything generalizes to $d > 1$.

Fix $\theta_0 \in \Theta$; we consider a classical framework where we assume that θ_0 is the true data generating parameter. Let $\mathbb{P}_0$ denote probabilities under the sampling distribution $f(\cdot \mid \theta_0)$. [To be technically precise, we should write n-fold products of $\mathbb{P}_0$ below.] Assume usual regularity conditions on the likelihood function that are used to prove asymptotic normality of the maximum likelihood estimator (thrice differentiable etc.). Also assume the *prior distribution is continuous and positive at* θ_0. Let $t = \sqrt{n}(\theta - \hat{\theta})$ and let $\pi_n^*(t \mid \mathbf{x})$ denote the posterior distribution of t given $\mathbf{x}$. Then,

$$\lim_{n\to\infty} \int_{\mathbb{R}} \left| \pi_n^*(t \mid \mathbf{x}) - \phi(t \mid 0, \mathrm{I}(\hat{\theta})^{-1}) \right| dt = 0$$

with $\mathbb{P}_0$ probability 1, where $\phi(t \mid \mu, \sigma^2)$ denote the $N(\mu, \sigma^2)$ density evaluated at t.

Remark: Observe that both $\pi_n^*(t \mid \mathbf{x})$ and a $N(0, \mathrm{I}(\hat{\theta})^{-1})$ are random probability distributions (RPM) since they involve the random quantity $\mathbf{x}$ which is distributed according to $\mathbb{P}_0$. To say that these two RPMs are close, we first reduce the problem to a scalar random quantity, namely the total variation distance between the two RPMs:

$$T(\mathbf{x}) = \int_{\mathbb{R}} \left| \pi_n^*(t \mid \mathbf{x}) - \phi(t \mid 0, \mathrm{I}(\hat{\theta})^{-1}) \right| dt.$$

$T(\mathbf{x})$ is now a scalar random variable (function of $x_1, \ldots, x_n$) and it makes sense to study its convergence properties as $n \to \infty$, just like you study convergence properties of usual statistics like the sample mean or sample variance. The BVM theorem states that $T(\mathbf{x}) \to 0$ a.e. $[\mathbb{P}_0]$.

Remark: Under the additional assumption that π has finite expectation, the following stronger result holds:

$$\lim_{n\to\infty} \int_{\mathbb{R}} |t| \left| \pi_n^*(t \mid \mathbf{x}) - \phi(t \mid 0, \mathrm{I}(\hat{\theta})^{-1}) \right| dt = 0$$

with $\mathbb{P}_0$ probability 1. If this holds, then with $\mathbb{P}_0$ probability 1,

$$\left| \int t\pi_n^*(t \mid \mathbf{x})dt - \int t\phi(t \mid 0, \mathrm{I}(\hat{\theta})^{-1})dt \right| \to 0,$$

which further implies $\int t\pi_n^*(t \mid \mathbf{x})dt \to 0$. Now, denoting $\theta_n^* = E\theta \mid \mathbf{x} = \int \theta\pi_n(\theta \mid \mathbf{x})d\theta$, it is straightforward to see that $\sqrt{n}(\theta_n^* - \hat{\theta}) = \int t\pi_n^*(t \mid \mathbf{x})dt$. Thus, $\sqrt{n}(\theta_n^* - \hat{\theta}) \to 0$ a.e. $[\mathbb{P}_0]$, or in other words, $\theta_n^* \approx \hat{\theta} + 1/\sqrt{n}$. Therefore, the posterior mean and the m.l.e are asymptotically equivalent.

Remark: A slightly different version of the BvM theorem replaces the data dependent quantity $\mathrm{I}(\hat{\theta})^{-1})$ by $I(\theta_0)^{-1}$. That is,

$$\lim_{n\to\infty} \int_{\mathbb{R}} \left| \pi_n^*(t \mid \mathbf{x}) - \phi(t \mid 0, \mathrm{I}(\theta_0)^{-1}) \right| dt = 0$$

with $\mathbb{P}_0$ probability 1.

Remark: A BvM result implies that Bayesian credible sets have asymptotically correct frequentist coverage. This gives frequentist justification to using credible sets as confidence sets in situations where a confidence set is not readily available otherwise.

3.1 Illustration

As an illustration, consider $x_1, \ldots, x_n \mid \theta \sim N(\theta, 1)$, with $\theta \sim N(\xi, \lambda^{-1})$. We have seen previously that $\pi(\theta \mid x_{1:n}) \equiv N((n\bar{x} + \lambda\xi)/(n+\lambda), 1/(n+\lambda))$. We showed in class that

$$\begin{aligned} & d_{TV}^2\Big(N\big((n\bar{x} + \lambda\xi)/(n+\lambda), 1/(n+\lambda)\big), N(0, 1/n)\Big) \\ & \leq \left[\log(1 + \lambda/n) - \frac{\lambda}{n+\lambda} + \frac{n\lambda^2}{(n+\lambda)^2}(\bar{x} - \xi)^2\right] \end{aligned}$$

by bounding the squared TV distance by twice the KL divergence. This quantity converges to zero in probability or in expectation as $n \to \infty$.