Unit 8: Estimation of Normalizing Constants

8.1 Problem Set-up

Let π_1, π_2 be two distributions with support $\mathcal{X}_1, \mathcal{X}_2$ respectively, where $\mathcal{X}_1 \subset \mathcal{X}_2$. Assume that both distributions have a density function with respect to the dominating measure μ, which can be expressed by

$$\frac{\mathrm{d}\pi_1}{\mathrm{d}\mu}(x) = \frac{p_1(x)}{C_1}, \quad \frac{\mathrm{d}\pi_2}{\mathrm{d}\mu}(x) = \frac{p_2(x)}{C_2},$$

where p_1, p_2 are known functions, and $C_1, C_2 \in (0, \infty)$ are unknown normalizing constants. Our goal is to estimate the ratio

$$r = \frac{C_1}{C_2} = \frac{\int_{\mathcal{X}_1} p_1(x)\mu(\mathrm{d}x)}{\int_{\mathcal{X}_2} p_2(x)\mu(\mathrm{d}x)}. \tag{1}$$

This is a very common computational problem in statistics and data science. If one is only interested in a single normalizing constant C_1 (which actually is rare in practice), one can pick a reference distribution π_2 for which the normalizing constant is known and again consider the problem of estimating r in (1).

Example 8.1. In simulated tempering, one is interested in estimating r for $p_1(x) = f(x)^{1/\tau_1}$ and $p_2(x) = f(x)^{1/\tau_2}$ (assume $\mathcal{X}_1 = \mathcal{X}_2$), where f is a known function and $\tau_1, \tau_2 > 0$ are temperatures. An accurate estimation of this ratio of normalizing constants can help one choose the auxiliary constants involved in the joint target distribution of simulated tempering [5].

Example 8.2. Consider a Bayesian hypothesis testing problem where we have two competing nested models m_1, m_2 for explaining the data D. Let the parameter space of m_1 be $\mathcal{X}_1$, which is assumed to be a subset of $\mathcal{X}_2$, the parameter space of m_2. Then, the standard Bayesian approach is to compute the Bayes factor

$$\mathrm{BF} = \frac{\int_{\mathcal{X}_1} f(D \,|\, m_1, x_1) p(x_1 \,|\, m_1)\mu(\mathrm{d}x_1)}{\int_{\mathcal{X}_2} f(D \,|\, m_2, x_2) p(x_2 \,|\, m_2)\mu(\mathrm{d}x_2)},$$

where f denotes the data likelihood, and $p(x \,|\, m)$ denotes the prior distribution of the parameter x given model m. So the Bayes factor itself is a ratio of two normalizing constants.

Example 8.3. Consider a statistical model with likelihood function $f(D \,|\, x)$, where D denotes the data and x denotes the parameter. Suppose that there is missing or censored data, and denote the observed data by D_0. To compute the likelihood of parameter x given only D_0, we need to evaluate $f(D_0 \,|\, x) = \int f(D, D_0 \,|\, x)\mathrm{d}D$, which is the normalizing constant of the complete-data likelihood (integrated over the complete data). To evaluate whether parameter x_1 or x_2 fits the data D_0 better, we need to evaluate the ratio of the two corresponding normalizing constants.

8.2 Direct Importance Sampling Methods

Example 8.4 (simple importance sampling)**.** The importance sampling methods introduced in Unit 1 can be used to estimate C_1, C_2 separately. Given i.i.d. samples $X_1, X_2, \ldots, X_n$ drawn from another distribution with density $\tilde{\pi}(x)$ and support $\mathcal{X}_2$, we can estimate C_j (for $j = 1, 2$) by

$$\hat{C}_j = \frac{1}{n} \sum_{i=1}^{n} \frac{p_j(X_i)}{\tilde{\pi}(X_i)}.$$

Then,

$$\hat{r} = \frac{\hat{C}_1}{\hat{C}_2} = \frac{\sum_{i=1}^{n} p_1(X_i)/\tilde{\pi}(X_i)}{\sum_{i=1}^{n} p_2(X_i)/\tilde{\pi}(X_i)}$$

is a consistent estimator for r. Note that (i) the assumption $\mathcal{X}_1 \subset \mathcal{X}_2$ is crucial, and (ii) to calculate $\hat{r}$, we only need to evaluate $\tilde{\pi}$ up to a normalizing constant. This method is also called ratio importance sampling [1]. Of course, $X_1, X_2, \ldots$ do not have to be independent (e.g. they can be generated from an MCMC algorithm with stationary distribution $\tilde{\pi}$), and we can also estimate C_1, C_2 using samples generated from different reference distributions.

Example 8.5 (reciprocal importance sampling)**.** Let $X_1, X_2, \ldots, X_n$ be samples generated from the distribution π_2 (e.g. by an MCMC algorithm). By using an idea known as the reciprocal importance sampling method [2], we can compute the estimator by

$$\hat{r} = n^{-1} \sum_{i=1}^{n} \frac{p_1(X_i)}{p_2(X_i)}$$

is an unbiased estimator for r. This is actually a special case of Example 8.4 with $\tilde{\pi} = \pi_2$.

8.3 Bridge Sampling

Let $\mathbb{E}_i$ denote the expectation with respect to π_i. Let α be a function defined on $\mathcal{X}_1$. Extend p_1, α to $\mathcal{X}_2$ by letting $p_1(x) = \alpha(x) = 0$ for $x \in \mathcal{X}_2 \setminus \mathcal{X}_1$. Observe that

$$r = \frac{C_1}{C_2} = \frac{\mathbb{E}_2[p_1(X)\alpha(X)]}{\mathbb{E}_1[p_2(X)\alpha(X)]},$$

provided that the expectations are defined and nonzero, i.e.,

$$0 < \left| \int_{\mathcal{X}_1} p_1(x) p_2(x) \alpha(x) \mu(\mathrm{d}x) \right| < \infty.$$

Hence, if we have samples $X_1, \ldots, X_{n_1}$ drawn from π_1 and $Y_1, \ldots, Y_{n_2}$ drawn from π_2, we can estimate r by

$$\hat{r}_\alpha = \frac{n_2^{-1} \sum_{i=1}^{n_2} p_1(Y_i)\alpha(Y_i)}{n_1^{-1} \sum_{i=1}^{n_1} p_2(X_i)\alpha(X_i)}. \tag{2}$$

This method is called bridge sampling [6]. By choosing $\alpha(x) = 1/p_2(x)$ for $x \in \mathcal{X}_1$, we get the estimator given in Example 8.5.

Let $\rho_i = n_i/n$. It was shown in [6] that the asymptotically optimal choice of α is

$$\alpha(x) \propto \frac{1}{\rho_1 \pi_1(x) + \rho_2 \pi_2(x)}, \quad \forall x \in \mathcal{X}_1,$$

where $\pi_i(x) = p_i(x)/C_i$ denotes the normalized density function. This choice asymptotically minimizes the relative mean-squared error $\mathrm{RE}(\alpha) = r^{-2}\mathbb{E}[(\hat{r}_\alpha - r)^2]$. Bridge sampling with this optimal choice of α coincides with the reverse logistic regression method proposed by [4].

We can also derive the bridge sampling estimator by generalizing Example 8.5. We write

$$r = \frac{B/C_2}{B/C_1}, \text{ where } B = \int_{\mathcal{X}_1} p_1(x)p_2(x)\alpha(x)\mu(\mathrm{d}x).$$

By Example 8.5, the numerator of the right-hand side of (2) is an unbiased estimator of B/C_2, and the demoniator is an unbiased estimator of B/C_1. Here, the distribution with un-normalized density $p_1(x)p_2(x)\alpha(x)$ serves as a "bridge" connecting two potentially very different distributions π_1, π_2. One can also use a sequence of bridges by writing

$$r = \frac{C_1}{C_2} = \prod_{k=1}^{L} \frac{B_{2k-1}/B_{2k}}{B_{2k-1}/B_{2k-2}}$$

with $B_0 = C_1$ and $B_{2L} = C_2$. Letting $L \to \infty$, we obtain a "path" of distributions that evolve from π_1 to π_2, which is the motivation behind the method to be introduced in the next subsection.

8.4 Path Sampling

In this subsection, we assume $\mathcal{X}_1 = \mathcal{X}_2 = \mathcal{X}$. Let $p(x \,|\, \theta)$ be a function of (x, θ) such that $p_1(x) = p(x \,|\, \theta_1)$ and $p_2(x) = p(x \,|\, \theta_2)$ for some real numbers $\theta_1 < \theta_2$. Define

$$Z(\theta) = \int_{\mathcal{X}} p(x \,|\, \theta)\mu(\mathrm{d}x).$$

So we have $C_i = Z(\theta_i)$ for $i = 1, 2$. Assume that

(i) $p(x \,|\, \theta) > 0$ for every $\theta \in [\theta_1, \theta_2]$ and $x \in \mathcal{X}$;

(ii) $\int_{\mathcal{X}} p(x \,|\, \theta)\mu(\mathrm{d}x) < \infty$ for every $\theta \in [\theta_1, \theta_2]$;

(iii) for every $\theta \in [\theta_1, \theta_2]$,

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \int_{\mathcal{X}} p(x \,|\, \theta)\mu(\mathrm{d}x) = \int_{\mathcal{X}} \frac{\partial p(x \,|\, \theta)}{\partial \theta}\mu(\mathrm{d}x),$$

where all derivatives involved are also assumed to exist.

Under the above assumptions, we have

$$\frac{\mathrm{d}}{\mathrm{d}\theta} \log Z(\theta) = \mathbb{E}_{X \sim p(x \,|\, \theta)} \left[\frac{\partial \log p(X \,|\, \theta)}{\partial \theta} \right], \tag{3}$$

where $X \sim p(x \,|\, \theta)$ indicates that X is a random variable with *un-normalized* density $p(x \,|\, \theta)$. Note that (3) is essentially the Fisher's identity frequently used in mathematical statistics.

It follows from (3) that

$$\lambda := -\log r = \log \frac{Z(\theta_2)}{Z(\theta_1)} = \int_{\theta_1}^{\theta_2} \mathbb{E}_{X \sim p(x \,|\, \theta)} \left[U(X, \theta) \right] \mathrm{d}\theta,$$

where

$$U(x, \theta) = \frac{\partial \log p(X \,|\, \theta)}{\partial \theta}.$$

Let $\nu(\theta)$ denote a "prior" probability distribution of θ with support $[\theta_1, \theta_2]$. We can further express λ by

$$\lambda = \mathbb{E} \left[\frac{U(X, \theta)}{\nu(\theta)} \right],$$

where (X, θ) is generated from the joint distribution with density proportional to $p(x \,|\, \theta)\nu(\theta)$. Hence, if we have samples $(X_i, \theta_i)_{i=1}^n$ drawn from $p(x \,|\, \theta)\nu(\theta)$, we can estimate λ by

$$\hat{\lambda} = \frac{1}{n} \sum_{i=1}^n \frac{U(X_i, \theta_i)}{v(\theta_i)}.$$

This method is called path sampling and is also applicable for multivariate θ [3].

Assume the samples $(X_i, \theta_i)_{i=1}^n$ are i.i.d. The variance of the estimator $\hat{\lambda}$ is given by

$$\mathrm{Var}(\hat{\lambda}) = \frac{1}{n} \left\{ \int_{\theta_1}^{\theta_2} \int_{\mathcal{X}} \frac{U^2(x, \theta)}{\nu(\theta)} \frac{p(x \,|\, \theta)}{Z(\theta)} \mu(\mathrm{d}x) \, \mathrm{d}\theta - \lambda^2 \right\}.$$

Equivalently, letting $\pi(x \,|\, \theta) = p(x \,|\, \theta)/Z(\theta)$ denote the normalized density, we have

$$\begin{aligned} n\mathrm{Var}(\hat{\lambda}) &= \int_{\theta_1}^{\theta_2} \int_{\mathcal{X}} \left(\frac{\partial \log \pi(x \,|\, \theta)}{\partial \theta} + \frac{\partial \log Z(\theta)}{\partial \theta} \right)^2 \frac{\pi(x \,|\, \theta)}{\nu(\theta)} \mu(\mathrm{d}x) \, \mathrm{d}\theta - \lambda^2 \\ &= \int_{\theta_1}^{\theta_2} \int_{\mathcal{X}} \left(\frac{\partial \log \pi(x \,|\, \theta)}{\partial \theta} \right)^2 \frac{\pi(x \,|\, \theta)}{\nu(\theta)} \mu(\mathrm{d}x) \, \mathrm{d}\theta + \int_{\theta_1}^{\theta_2} \left(\frac{\partial \log Z(\theta)}{\partial \theta} \right)^2 \frac{1}{\nu(\theta)} \mathrm{d}\theta - \lambda^2. \end{aligned}$$

The following result is from [1]:

Theorem 8.1. *Under the assumptions given at the beginning of this subsection,*

$$\mathrm{Var}(\hat{\lambda}) \geq \frac{4}{n} H^2(\pi_1, \pi_2),$$

where $H^2(\pi_1, \pi_2) = \int_{\mathcal{X}} \left(\sqrt{\pi_1(x)} - \sqrt{\pi_2(x)} \right)^2 \mu(\mathrm{d}x)$ *is the Hellinger distance between the two distributions.*

Exercise 8.1. Prove (3).

Exercise 8.2. Use Cauchy-Schwarz inequality to prove Theorem 8.1.

References

[1] Ming-Hui Chen and Qi-Man Shao. On Monte Carlo methods for estimating ratios of normalizing constants. *The Annals of Statistics*, 25(4):1563–1594, 1997.

[2] Alan E Gelfand and Dipak K Dey. Bayesian model choice: asymptotics and exact calculations. *Journal of the Royal Statistical Society: Series B (Methodological)*, 56(3):501–514, 1994.

[3] Andrew Gelman and Xiao-Li Meng. Simulating normalizing constants: From importance sampling to bridge sampling to path sampling. *Statistical science*, pages 163–185, 1998.

[4] Charles J Geyer. Estimating normalizing constants and reweighting mixtures in Markov Chain Monte Carlo. 1991.

[5] Holden Lee, Andrej Risteski, and Rong Ge. Beyond log-concavity: Provable guarantees for sampling multi-modal distributions using simulated tempering langevin Monte Carlo. *Advances in neural information processing systems*, 31, 2018.

[6] Xiao-Li Meng and Wing Hung Wong. Simulating ratios of normalizing constants via a simple identity: a theoretical exploration. *Statistica Sinica*, pages 831–860, 1996.