Predictive computable ⟺ posterior computable

Cameron E. Freer
Massachusetts Institute of Technology
`freer@math.mit.edu`

Daniel M. Roy
Massachusetts Institute of Technology
`droy@csail.mit.edu`

1   Introduction

As we devise more complicated prior distributions, will we ever find ourselves unable to compute the conditional distributions necessary for posterior analysis?

In many settings, computing posterior distributions is straightforward. For a discrete random variable Y and event A, the formula

$$\mathbf{P}(A \mid Y = a) = \frac{\mathbf{P}(A, \, Y = a)}{\mathbf{P}(Y = a)} \tag{1}$$

gives us the conditional probability of A given $Y = a$, provided that $\mathbf{P}(Y = a) > 0$. In the case of a continuous random variable, the situation is far more complicated. On sufficiently nice spaces, conditional distributions are always well-defined; but their implicit definition, given by Kolmogorov [Kol33], gives no recipe for their calculation.

A common setting is when the likelihood model is *dominated*. We say that a likelihood, given by, e.g., a conditional distribution $P[X \mid \Theta]$, is dominated when there is a measure μ with respect to which almost all distributions $P[X \mid \Theta = \theta]$ have a density $p(x \mid \theta)$. When the likelihood model is dominated, a Bayes rule exists and so a ratio of probability densities, just like the above elementary rule (1), gives an explicit form for the conditional distribution. This is the setting for most of finite-dimensional statistics. Even in this setting, there exist computable dominated models with noncomputable posteriors, as shown in [AFR].

However, in the nonparametric setting, the problem of computing posterior distributions is even more complicated, in part because the likelihood is often not dominated. In these cases, someone must derive the form of the conditional without the help of a density. For example, let $\alpha > 0$ and let H be a continuous distribution, and define μ to be the distribution of a Dirichlet process with parameter αH (see [Fer73]). Then two independent random measures, each with distribution μ, will have overlapping support with probability zero. Therefore, no Bayes rule exists.

In the case of the Dirichlet process, Ferguson derived a closed form expression for the conditional distribution. The negative result by [AFR] implies that a computable conditional distribution, let alone a closed form, does not always exist. Modern nonparametric models have grown considerably in complexity, and care is needed when building a model in order to ensure that one can determine the posterior distribution. Can we build models in such a way that we can always compute the posterior?

Orbanz [Orb10] describes a method for building conjugate nonparametric models with closed form update rules for posterior analysis. In particular, he shows how to construct a nonparametric model as the limit of a conditionally projective family of finite dimensional conditional distributions.

Freer and Roy [FR] investigate a natural minimal requirement: when there exists *some program*, not necessarily a closed form expression, that computes the posterior distribution, in the sense of a numerical algorithm which provides arbitrary precision. Although there are circumstances in which there is provably no such algorithm, they show that one can

exploit *exchangeability* to obtain a general procedure that transforms a numerical algorithm that can sample a stochastic process into a numerical algorithm that computes the posterior distribution of the latent variables governing the distribution of the stochastic process.

2 Posterior analysis of exchangeable sequences

Let $X = \{X_i\}_{i \geq 1}$ be an infinite sequence of real random variables. We say that X is *exchangeable* if, for every finite set $\{k_1, \dots, k_j\}$ of distinct indices, $(X_{k_1}, \dots, X_{k_j})$ is equal in distribution to $(X_1, \dots, X_j)$. Recall the classic theorem of de Finetti, which characterizes exchangeable sequences as mixtures of i.i.d. sequences.

Theorem 1 (de Finetti [Kal05, Chap. 1.1])**.** *Let $X = \{X_i\}_{i \geq 1}$ be an exchangeable sequence of random variables with distribution χ. There is an (almost surely unique) random probability measure ν on $\mathbb{R}$ such that X is conditionally i.i.d. with respect to ν:*

$$\mathbf{P}[X_1 \in B_1, \ X_2 \in B_2, \ \dots \mid \nu] = \prod_{i \geq 1} \nu(B_i) \quad a.s.,$$

for $B_i \in \mathcal{B}_{\mathbb{R}}$. □

The random measure ν is called the *directing random measure*. Its distribution μ (a measure on probability measures) is called the *mixing measure* or the *de Finetti measure*. Note that ν is, in general, an infinite dimensional object.

In most cases, our knowledge of an exchangeable sequence is given by an algorithm for sampling from $\mathbf{P}[X_{k+1}|X_{1:k}]$, i.e., the posterior predictive — a rule that, given samples for a prefix $X_{1:k}$ of the sequence, describes the distribution of the next element, X_{k+1}. The following theorem of [FR] characterizes our ability to perform posterior inference in these settings.

Theorem 2 (Freer and Roy [FR])**.** *Let $X = \{X_i\}_{i \geq 1}$ be an exchangeable sequence of random variables with directing random measure ν. Then the posterior distribution $\mathbf{P}[\nu \mid X_{1:k}]$ is computable if and only if the posterior predictive $\mathbf{P}[X_{k+1} \mid X_{1:k}]$ is computable.* □

By induction, a posterior predictive rule can be used to subsequently sample X_{k+2} given $X_{1:k+1} = \{X_1, \dots, X_{k+1}\}$, and so on, "hallucinating" an entire infinite exchangeable sequence given the original prefix. This result states that the ability to consistently extend a prefix to an infinite sequence is equivalent to being able to compute the posterior distribution of the latent distribution that is generating the sequence.

In summary:

Given a computable posterior predictive rule for an exchangeable stochastic process, one can compute the posterior distribution of the mixing measure.

The proof is uniform: an algorithm for the predictive can be automatically transformed into an algorithm for the posterior (and vice versa). The key ingredient is a computable version of de Finetti's theorem [FR09].

2.1 Example: Dirichlet process

This result captures a common setting of nonparametric modeling, where a model is given by a rule for generating a sequence of exchangeable observations in terms of previous samples. This representation exists even without a Bayes rule.

Consider the following exchangeable sequence whose combinatorial structure is known as the Chinese restaurant process (CRP) [Ald85]. Let $\alpha > 0$ and let H be a distribution on $\mathbb{R}$ and sample $X_{k+1} \sim \frac{k}{k+\alpha} \sum_{i=1}^{k} \delta_{X_i} + \frac{\alpha}{k+\alpha} H$. The sequence $\{X_i\}_{i \geq 1}$ is exchangeable and the directing random measure is a Dirichlet process with parameter αH. Note that this sampling rule satisfies the hypotheses of Theorem 2. The proof (if implemented as code) automatically transforms the CRP rule into the computable map

$$\{X_i\}_{i \leq k} \mapsto \mathrm{DP}(\alpha H + \textstyle\sum_{i=1}^{k} \delta_{X_i}). \tag{2}$$

References

[AFR] N. L. Ackerman, C. E. Freer, and D. M. Roy, *The computability of conditional probability*, Preprint.

[Ald85] D. J. Aldous, *Exchangeability and related topics*, École d'été de probabilités de Saint-Flour, XIII—1983, Lecture Notes in Math., vol. 1117, Springer, Berlin, 1985, pp. 1–198.

[Fer73] T. S. Ferguson, *A Bayesian analysis of some nonparametric problems*, Ann. Statist. **1** (1973), 209–230.

[FR] C. E. Freer and D. M. Roy, *For exchangeable sequences, the predictive is computable if and only if the posterior is computable*, Preprint.

[FR09] _____, *Computable exchangeable sequences have computable de Finetti measures*, Mathematical Theory and Computational Practice (CiE 2009), Proc. of the 5th Conf. on Computability in Europe (K. Ambos-Spies, B. Löwe, and W. Merkle, eds.), Lecture Notes in Comput. Sci., vol. 5635, Springer, 2009, pp. 218–231.

[Kal05] O. Kallenberg, *Probabilistic symmetries and invariance principles*, Springer, New York, 2005.

[Kol33] A. N. Kolmogorov, *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Springer, 1933.

[Orb10] P. Orbanz, *Construction of nonparametric Bayesian models from parametric Bayes equations*, Adv. in Neural Inform. Processing Syst. 22, 2010.

[Set94] J. Sethuraman, *A constructive definition of Dirichlet priors*, Statistica Sinica **4** (1994), 639–650.