Bayesian Analysis (2016) **11**, Number 2, pp. 421–445

Expert Information and Nonparametric Bayesian Inference of Rare Events

Hwan-sik Choi[*]

Abstract. Prior distributions are important in Bayesian inference of rare events because historical data information is scarce, and experts are an important source of information for elicitation of a prior distribution. I propose a method to incorporate expert information into nonparametric Bayesian inference on rare events when expert knowledge is elicited as moment conditions on a finite dimensional parameter θ only. I generalize the Dirichlet process mixture model to merge expert information into the Dirichlet process (DP) prior to satisfy expert's moment conditions. Among all the priors that comply with expert knowledge, we use the one that is closest to the original DP prior in the Kullback–Leibler information criterion. The resulting prior distribution is given by exponentially tilting the DP prior along θ. I provide a Metropolis–Hastings algorithm to implement this approach to sample from posterior distributions with exponentially tilted DP priors. The proposed method combines prior information from a statistician and an expert by finding the least-informative prior given expert information.

Keywords: Dirichlet process mixture, defaults, Kullback–Leibler information criterion, maximum entropy, Metropolis–Hastings algorithm.

1 Introduction

I develop a nonparametric Bayesian framework that incorporates expert information for inference of rare events. Inference of rare events such as defaults in a high grade portfolio, extreme losses, and other catastrophic events is critical in measuring credit risk and systemic risk, which are important in risk management for optimal hedging and economic capital calculations. But inference of rare events is difficult because of lack of historical data information. Therefore, it is important to use all sources of information including non-data information, and the Bayesian method provides a natural and coherent mechanism of combining all available information. For rare events, the use of non-informative or objective priors in a Bayesian model is often not satisfactory because of the scarcity of information in data. Kiefer (2009, 2010) proposes to use expert information for default estimation as an additional source of information, and argues that the Bayesian approach is a natural theoretical framework to include expert information for inference on rare events. He uses expert information to elicit the Beta distribution prior or a mixture of the Beta distributions for a default probability. Although Kiefer's approach is illustrated with the binomial sampling distribution for the number of defaults in a portfolio, it can be easily extended to any general parametric models.

[*] Department of Economics, Binghamton University, 4400 Vestal Parkway E., P.O. Box 6000, Binghamton, NY 13902, U.S.A., hwansik@binghamton.edu

DOI: 10.1214/15-BA956

In this paper, I argue that nonparametric models are useful for rare events, and show that expert information can be effectively incorporated in nonparametric models too. Consider the following situation for motivation of using nonparametric models for rare events. Rare events are often associated with the tails of sampling distributions. For this reason, rare events are also called tail-risk events. For a parametric model, a finite number of parameters would determine the entire distribution, and inference of tail probabilities can be done from inference of the finite dimensional parameter. In this case, the problem of scarce tail-risk events is remedied by the use of the parametric assumption, which is a strong form of non-data information. Consequently, a parametric model improves the efficiency of using data information. However, when there is a concern for misspecification, the impact of misspecification can be serious for inference of tail-risk events. Since frequent events would dominate data information, they would drive estimation of the parameters as well. But the frequent events may not be relevant to tail-risk events if the model is misspecified. Moreover, it is difficult to check the adequacy of a parametric model specifically for tail probabilities because of the scarcity of rare events. For example, the maximum likelihood (ML) estimator for the mean parameter of normal distributions with a known variance would be the sample mean, which is a sufficient statistic. All inference can be done with the sufficient statistic. But if the normal distributions are misspecified, it would be better to do inference on the tail probabilities with emphasizing the observations far away from the sample mean since they are more relevant. For this reason, using a nonparametric model can be appealing for rare events when we do not have strong confidence in a parametric model.

Of course, an important disadvantage of nonparametric models is reduction in efficiency of using data. Therefore, additional information becomes very desirable, and expert information is particularly useful as complementary information in this setting. However, elicitation of expert opinion becomes difficult when the dimension of a model parameter increases. Practically, it is easier for an expert to talk about a few aspects of rare events rather than the entire shape of sampling distributions. My approach is to elicit expert information on the distribution of a finite dimensional parameter derived from the nonparametric Bayesian model. In particular, we consider the Dirichlet process mixture (DPM) model, and incorporate the expert information in its prior.

The sampling distribution of the DPM model is an infinite mixture of a family of distributions, and the mixing distribution works as an infinite dimensional parameter of the model. The DPM model uses the Dirichlet process as a prior distribution over the mixing distributions. Since the development of the DP by Freedman (1963), its theoretical properties and usefulness in Bayesian analysis are shown by Ferguson (1973, 1974), Blackwell and MacQueen (1973), and Antoniak (1974). Especially, the DP is widely used as a prior distribution of nonparametric Bayesian models. See Lo (1984) for more information on the DP, and Ghosh and Ramamoorthi (2003) and Hjort et al. (2010) for an overview of nonparametric Bayesian methods.

The DPM model has become popular with the development of the Markov Chain Monte Carlo (MCMC) methods in Bayesian nonparametric modeling because of its flexibility compared to a parametric model or a finite mixture model with a fixed number of mixing components. To obtain the posterior distribution for the DPM model, various sampling schemes were developed. Blackwell and MacQueen (1973) develop Pólya urn schemes for DP priors. Escobar (1994) and Escobar and West (1995) provide generalized Pólya urn Gibbs samplers for the DPM model. MacEachern and Müller (1998) consider the DPM models without conjugacy restrictions. Neal (2000), Ishwaran and Zarepour (2000) and Ishwaran and James (2001) also provide generalized Pólya urn Gibbs samplers and blocked Gibbs samplers for the general stick-breaking priors that include DP priors as a special case. In economics, Hirano (2002) studies semiparametric Bayesian inference for a panel model with a countable mixture of normally distributed errors with mixing weights drawn from the DP. Griffin and Steel (2011) consider a serially correlated sequence of DPs with applications to financial time series and GDP data, and Jensen and Maheu (2010) use the DP for the distribution of errors that are multiplied by the volatility process in a stochastic volatility model. Taddy and Kottas (2010) apply the DPM model for inference in quantile regressions.

In the DPM model, a mixing distribution drawn from the DP determines the sampling distribution F that is a member of the space $\mathcal{F}$ of all sampling distribution functions of a nonparametric model. Therefore, the DP gives a distribution over the functional space $\mathcal{F}$. In the framework proposed in this paper, we consider a finite dimensional vector θ of which the elements are some functionals of $F \in \mathcal{F}$ related to the rare events that an expert has knowledge on. We elicit expert opinion about the distribution of θ rather than the infinite dimensional space $\mathcal{F}$. Elicitation of expert information requires a careful design. Generally, it is easier for an expert to think about probabilities or quantiles than moments. Kadane and Wolfson (1998) and Garthwaite et al. (2005) discuss the issues in elicitation of expert opinions. Kadane et al. (1996) study elicitation of a subjective prior for Bayesian unit root models. Chaloner and Duncan (1983) develop a computer scheme to elicit expert information from a predictive distribution.

In my approach, expert information is in the form of moment conditions on θ that the DP prior may not satisfy. We combine the expert information and the DP prior by modifying the DP prior to satisfy the moment conditions. Among all such modifications, we find the prior distribution that is closest to the original DP prior in the Kullback–Leibler information criterion. The resulting prior distribution is given by exponentially tilting the DP prior along θ. I also provide a Metropolis–Hastings algorithm to implement my approach to draw a sample from the exponentially tilted DP prior. The method proposed in this paper gives a simple way to combine the prior information from a statistician and an expert by finding the least-informative prior given expert information, based on a statistician's prior.

2 Expert information in nonparametric Bayesian inference

Consider the DPM model for n observations $\{y_i\}_{i=1}^n$ given by

$$\begin{aligned} y_i \mid \xi_i &\sim K_{\xi_i}, \\ \xi_i \mid G &\sim G, \\ G &\sim \mathcal{P}, \end{aligned}$$

where $X \sim F$ means the distribution of X is F, $K_{\xi_i} = K(\cdot|\xi_i)$ is a distribution function with a parameter $\xi_i \in \Xi \subset \mathbb{R}^d$, and ξ_i is a random draw from a mixing measure G defined on a probability space $(\Xi, \mathcal{B})$, where $\mathcal{B}$ is a σ-field of subsets of Ξ. For notational simplicity, I omit the subscript i. Let $\mathcal{G}$ be the space of all mixing measures on Ξ. In the DPM model, a measure $G \in \mathcal{G}$ is thought to be an infinite dimensional parameter, and the prior distribution of G is given by a Dirichlet process $\mathcal{P} = \mathrm{DP}(\alpha_0 G_0)$ with a concentration parameter $\alpha_0 > 0$ and a base distribution $G_0 \in \mathcal{G}$. The mean of $\mathrm{DP}(\alpha_0 G_0)$ is G_0, so we can write

$$\mathbf{E}G = G_0, \tag{1}$$

where

$$\mathbf{E}G = \int G \, \mathcal{P}(dG).$$

The concentration parameter α_0 is inversely related to the variance of $\mathrm{DP}(\alpha_0 G_0)$, and we have $G \overset{d}{\to} G_0$ as $\alpha_0 \to \infty$. Any finite-dimensional distribution of $\mathrm{DP}(\alpha_0 G_0)$ is a Dirichlet distribution, i.e., for an arbitrary finite measurable partition $(B_1, B_2, \ldots, B_J)$ of Ξ, we have

$$(G(B_1), G(B_2), \ldots, G(B_J)) \sim \mathrm{Dirichlet}(\alpha_0 G_0(B_1), \alpha_0 G_0(B_2), \ldots, \alpha_0 G_0(B_J)),$$

and $\mathbf{E}G(B) = G_0(B)$ for all $B \in \mathcal{B}$. Because of $\sum_{j=1}^{J} \alpha_0 G_0(B_j) = \alpha_0 G_0(\Xi) = \alpha_0$, we also call α_0 the total mass parameter.

The distribution function $F(y|G)$ of y given G can be written in a mixture form as

$$F(y|G) = \int K(y|\xi) \, G(d\xi), \tag{2}$$

where $K(y|\xi)$ is used as the mixing kernel. For convenience of exposition of the main idea, we also assume that the probability density function of $K(y|\xi)$ exists for all ξ. Then we can write the sampling density of y given G as

$$f(y|G) = \int k(y|\xi) \, G(d\xi), \tag{3}$$

where the mixing kernel $k(y|\xi)$ is the probability density function of $K(y|\xi)$.

In our framework, experts have information on an r-dimensional parameter of interest

$$\theta = \varphi(F(y|G)), \tag{4}$$

where $\varphi(\cdot)$ is an $r \times 1$ vector of functionals of the sampling distribution $F(y|G)$ of y. Note that the expected value of θ can be written as

$$\mathbf{E}\theta = \int \varphi(F(y|G)) \, \mathcal{P}(dG).$$

The functional φ can be a linear functional such as moments of $F(y|G)$, or a value of the distribution function $F(c|G)$ at a fixed point $y = c$. But it can also be a nonlinear functional such as quantiles, inter-quantile ranges, hazard functions of $F(y|G)$, or the quantities regarding the distribution of the maximum or minimum of n observations. See Gelfand and Mukhopadhyay (1995) and Gelfand and Kottas (2002) for more examples of φ and their properties. When φ is linear, from (2), we can simplify θ to a mixture of $\varphi(K(y|\xi))$,

$$\theta = \int \varphi(K(y|\xi))\, G(d\xi), \tag{5}$$

and using (1), the expected value of θ becomes

$$\mathbf{E}\theta = \int \varphi(K(y|\xi))\, G_0(d\xi). \tag{6}$$

An important class of a linear functional φ is given by

$$\varphi(F(y|G)) = \int h(y) f(y|G)\, dy,$$

where $h : \mathbb{R} \to \mathbb{R}^r$. Note that if $h(y) = y^p$, then θ becomes the pth order moment of $F(y|G)$, and if $h(y) = I\{y \leq c\}$, where $I(\cdot)$ is an indicator function, then θ is the cumulative distribution function evaluated at c. From (3), we can easily prove (5) by

$$\begin{aligned}\theta &= \int h(y) \int k(y|\xi)\, G(d\xi)\, dy \\ &= \iint h(y) k(y|\xi)\, dy\, G(d\xi) \\ &= \int \varphi(K(y|\xi))\, G(d\xi).\end{aligned}$$

But for a general nonlinear functional φ, (5) or (6) would not hold. As suggested by Kadane and Wolfson (1998), elicitation of expert information should include questions about quantiles (nonlinear φ) or probabilities such as values of distribution functions at fixed points (linear φ) rather than moments because moments are sensitive to tail probabilities, thus, difficult to have good knowledge of.

We assume that expert's information is in the form of l moment restrictions

$$\mathbf{E}g(\theta) = 0, \tag{7}$$

where $g(\cdot)$ is a function $g : \mathbb{R}^r \to \mathbb{R}^l$. From (4), we can write (7) as

$$\mathbf{E}\tilde{\varphi}(F(y|G)) = 0,$$

where $\tilde{\varphi} = g \circ \varphi$ is an l-dimensional functional. If both g and φ are linear, $\tilde{\varphi}$ is a linear functional and $\mathbf{E}g(\theta)$ is simplified to

$$\mathbf{E}g(\theta) = \iint \tilde{\varphi}(K(y|\xi))\, G(d\xi)\, \mathcal{P}(dG)$$

$$= \int \tilde{\varphi}(K(y|\xi))\, G_0(d\xi). \tag{8}$$

This implies that $\mathbf{E}g(\theta)$ depends on G_0 only, when $\tilde{\varphi}$ is linear and G is from $\mathrm{DP}(\alpha_0 G_0)$.

Although the condition in (7) is in the form of moment restrictions, it accommodates expert elicitation for a very wide range of quantities related to the distribution of θ because of the generality of $g(\cdot)$. Possible questions to experts include the α-level quantile q_α for $\alpha \in [0,1]$ using $g(\theta) = I\{\theta \leq q_\alpha\} - \alpha$, the cumulative distribution function $F(a)$ at $a \in \mathbb{R}$ using $g(\theta) = I\{\theta \leq a\} - F(a)$, the probability α on any measurable interval A using $g(\theta) = I\{\theta \in A\} - \alpha$, and the pth moments m_p using $g(\theta) = \theta^p - m_p$. Moreover, it is convenient to use the moment conditions to develop our method in the information theoretic framework because there are well established results under moment restrictions such as Csiszár (1975) in the information geometry literature.

Let

$$D(\mathcal{Q} \parallel \mathcal{P}) = \int \log\left(\frac{d\mathcal{Q}}{d\mathcal{P}}\right) d\mathcal{Q}$$

be the Kullback–Leibler information criterion (KLIC, Kullback and Leibler (1951), White (1982)) from a measure $\mathcal{Q}$ to an equivalent measure $\mathcal{P}$. The goal of this paper is to merge the expert information (7) into the DP prior $\mathcal{P}$ by finding the prior that is closest to $\mathcal{P}$ in the KLIC among the measures that satisfy (7). We get the new prior by solving

$$\min_{\mathcal{Q} \in \mathbb{Q}} D(\mathcal{Q} \parallel \mathcal{P}), \tag{9}$$

where

$$\mathbb{Q} = \{\mathcal{Q} : \mathbf{E}^{\mathcal{Q}} g(\theta) = 0\},$$

and $\mathbf{E}^{\mathcal{Q}} g(\theta) = \int g(\theta)\, \mathcal{Q}(dG)$ is the expectation of $g(\theta)$ with respect to $\mathcal{Q}$. The solution $\mathcal{Q}^*$ to (9) is well known and given by the Gibbs canonical density

$$\pi^* = \frac{d\mathcal{Q}^*}{d\mathcal{P}} = \frac{\exp(\lambda_*' g(\theta))}{\mathbf{E}^{\mathcal{P}} \exp(\lambda_*' g(\theta))}. \tag{10}$$

The coefficient vector λ_* in (10) can be calculated by the unconstrained convex problem

$$\min_{\lambda} \mathbf{E}^{\mathcal{P}} \exp(\lambda' g(\theta)), \tag{11}$$

and the minimum KLIC with $\mathcal{Q}^*$ is given by

$$D(\mathcal{Q}^* \parallel \mathcal{P}) = -\log(\mathbf{E}^{\mathcal{P}} \exp(\lambda_*' g(\theta))).$$

See Kitamura and Stutzer (1997) Section 2.2 for a short discussion on this result and further reference. Note also that since $D(\mathcal{Q} \parallel \mathcal{P})$ is convex in $\mathcal{Q}$, the uniqueness of the solution $\mathcal{Q}^*$ is guaranteed by the convexity of $\mathbb{Q}$ (Csiszár (1975), p. 147). Using the first order condition

$$\int g(\theta)\ \exp(\lambda_*' g(\theta))\, \mathcal{P}(dG) = 0$$

of (11), we can easily see that $\mathcal{Q}^*$ satisfies (7) from

$$\begin{aligned}\mathbf{E}^{\mathcal{Q}^*} g(\theta) &= \int g(\theta)\, \mathcal{Q}^*(dG) \\ &= \int g(\theta) \left(\frac{\exp(\lambda_*' g(\theta))}{\mathbf{E}^{\mathcal{P}} \exp(\lambda_*' g(\theta))} \right) \mathcal{P}(dG) \\ &= 0.\end{aligned}$$

Intuitively, the solution $\mathcal{Q}^*$ gives the least informative prior in $\mathbb{Q}$. Geometrically speaking, $\mathcal{Q}^*$ is a projection of $\mathcal{P}$ on $\mathbb{Q}$ in the sense that $\mathcal{Q}^*$ is the closest point on $\mathbb{Q}$ from $\mathcal{P}$ in the divergence measure $D(\mathcal{Q} \parallel \mathcal{P})$. Of course, if we use a different divergence measure such as the KLIC $D(\mathcal{P} \parallel \mathcal{Q})$ from $\mathcal{P}$ to $\mathcal{Q}$ as the minimization criterion, we would get a different projection point. By considering more general divergence measures between $\mathcal{P}$ and $\mathcal{Q}$, we can understand our projection $\mathcal{Q}^*$ in a larger perspective. Specifically, Amari (1982) considers a class of divergence measures $\mathcal{D}^{(\alpha)}(\mathcal{P} \parallel \mathcal{Q})$ called the α-divergence ($|\alpha| \leq 1$) from $\mathcal{P}$ to $\mathcal{Q}$ defined as

$$\mathcal{D}^{(\alpha)}(\mathcal{P} \parallel \mathcal{Q}) = \int f_\alpha \left(\frac{d\mathcal{Q}}{d\mathcal{P}} \right) d\mathcal{P},$$

where

$$f_\alpha(u) = \begin{cases} \frac{4}{1-\alpha^2} \left\{ 1 - u^{(1+\alpha)/2} \right\} & \text{for } \alpha \neq \pm 1, \\ u \log u & \text{for } \alpha = 1, \\ -\log(u) & \text{for } \alpha = -1. \end{cases}$$

The α-divergence has the duality

$$\mathcal{D}^{(\alpha)}(\mathcal{P} \parallel \mathcal{Q}) = \mathcal{D}^{(-\alpha)}(\mathcal{Q} \parallel \mathcal{P}),$$

and the (-1)-divergence $\mathcal{D}^{(-1)}(\mathcal{P} \parallel \mathcal{Q})$ is the KLIC from $\mathcal{P}$ to $\mathcal{Q}$, i.e., $\mathcal{D}^{(-1)}(\mathcal{P} \parallel \mathcal{Q}) = D(\mathcal{P} \parallel \mathcal{Q})$. See also Amari (1985) Section 3.5 and Amari and Nagaoka (2000) Section 3.2 for further discussion. The α-divergence is a special class of the f-divergence of Csiszár (1967a,b). For discrete measures, the α-divergence is also equivalent to Rényi's information divergence (Rényi (1961)), of which another equivalent form is known as the Cressie–Read power divergence family (Cressie and Read (1984)) in the generalized empirical likelihood literature such as Newey and Smith (2004). From the duality of the α-divergence, we can easily see that our minimization criterion $D(\mathcal{Q} \parallel \mathcal{P})$ in (9) is the 1-divergence $\mathcal{D}^{(1)}(\mathcal{P} \parallel \mathcal{Q})$ from $\mathcal{P}$ to $\mathcal{Q}$, or equivalently, the (-1)-divergence from $\mathcal{Q}$ to $\mathcal{P}$. In Amari's terminology, $\mathcal{Q}^*$ is called the 1-projection of $\mathcal{P}$ on $\mathbb{Q}$ because it minimizes the 1-divergence from $\mathcal{P}$ to $\mathcal{Q}$. In Csiszár (1975), $\mathcal{Q}^*$ is also called the I-projection of $\mathcal{P}$. Because of the exponential form in (10), the 1-projection $\mathcal{Q}^*$ is said to be given by the exponential tilting of $\mathcal{P}$. We call $\mathcal{Q}^*$ the exponentially tilted DP prior or the ETDP prior.

If both expert's moment function $g(\cdot)$ and functional $\varphi(\cdot)$ are linear, then, from (8), $\mathbf{E}g(\theta) = 0$ becomes moment conditions that depend on the base distribution G_0 only.

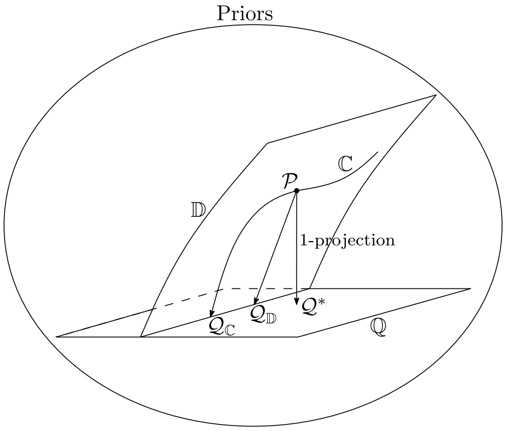

Figure 1: Geometry of exponential tilting of DP priors. The set $\mathbb{Q}$ represents the priors that satisfy expert's moment conditions, $\mathbb{D}$ is the space of all DP priors, and $\mathbb{C} \subset \mathbb{D}$ is the conjugate family of DP priors. In our approach, a DP prior $\mathcal{P} = \mathrm{DP}(\alpha_0 G_0)$ is projected onto $\mathbb{Q}$ by the 1-projection of Amari (1982) minimizing Amari's 1-divergence. $\mathcal{Q}^*$ represents the projected prior. Alternatively, $\mathcal{P}$ may be projected to the space $\mathbb{D} \cap \mathbb{Q}$ of the DP priors that satisfy the moment conditions. The prior $\mathcal{Q}_{\mathbb{D}} \in \mathbb{D} \cap \mathbb{Q}$ represents the projected prior with this method. If there are conjugate priors that satisfy the moment conditions, i.e., $\mathbb{C} \cap \mathbb{Q} \neq \varnothing$, $\mathcal{P}$ can be projected to the space $\mathbb{C} \cap \mathbb{Q}$. The prior $\mathcal{Q}_{\mathbb{C}} \in \mathbb{C} \cap \mathbb{Q}$ represents the projected prior in this case.

Consequently, $\mathcal{Q}^*$ would depend on G_0 only. If φ is nonlinear, the concentration parameter becomes important also. If the concentration parameter is small, $\mathcal{Q}^*$ would depend less on G_0.

In Figure 1, I present the geometric interpretation of our approach and discuss some alternative methods graphically. In the figure, the DP prior $\mathcal{P}$ is in the space $\mathbb{D}$ of all DP priors. The ETDP prior $\mathcal{Q}^*$ represents the 1-projection of $\mathcal{P}$ on $\mathbb{Q}$. Although $\mathcal{Q}^*$ is the closest prior to $\mathcal{P}$ among all the priors that comply the expert opinion, we may consider sub-optimal priors in consideration of convenience. For example, if we consider only the DP priors that comply the expert opinion ($\mathbb{D} \cap \mathbb{Q}$), the 1-projection $\mathcal{Q}_{\mathbb{D}}$ of $\mathcal{P}$ on $\mathbb{D} \cap \mathbb{Q}$ would be the prior that is closest to $\mathcal{P}$. Another possibility is to consider only conjugate DP priors shown as $\mathbb{C} \subset \mathbb{D}$ in the figure. If $\mathbb{C} \cap \mathbb{Q} \neq \varnothing$, then the 1-projection $\mathcal{Q}_{\mathbb{C}}$ of $\mathcal{P}$ on $\mathbb{C} \cap \mathbb{Q}$ is the closest prior to $\mathcal{P}$ among the conjugate DP priors that satisfy the expert moment conditions. Although $\mathcal{Q}_{\mathbb{D}}$ and $\mathcal{Q}_{\mathbb{C}}$ are more familiar priors to use, it may not be easy to find them. Also, $\mathcal{Q}_{\mathbb{C}}$ may not exist when $\mathbb{C}$ is low-dimensional and there are many expert moment conditions to satisfy. Although we do not investigate these alternative methods further, these are important problems that deserve more attention.

An alternative approach to the method proposed in this paper is to impose expert information on the base distribution G_0 by asking experts about G_0 directly. The main problem of this approach is that it is difficult to elicit moment conditions from experts because it is hard for an expert to imagine the relationship between G_0 and the distribution of the quantity of interest θ such as default probabilities. Experts would have a clearer idea on the distribution of θ rather than the distribution implied by G_0. Moreover, the distribution of θ also depends on the DP prior's concentration parameter on which expert information is not solicited. It is also difficult to interpret the resulting prior since it would not be the closest prior to the original DP prior in KLIC. On the other hand, this approach has the advantage that the resulting prior after the exponential tilting of G_0 is also a DP, which allows us to use existing posterior simulation techniques for DPM models.

The present paper is closely related to Kessler et al. (2015). In their work, expert information is provided through specifying a marginal prior on θ. The expert's knowledge in my paper is limited to some aspects (moment conditions on θ) of the marginal distribution of θ. In practice, it may be difficult to elicit from an expert the entire shape of the marginal distribution of θ, but it is easy to ask some questions on various features of the marginal distribution. In that case, the method proposed in the present paper would be more useful. Also, the ETDP prior from this paper would be naturally closer to the original non-expert DP prior than the one given by Kessler et al. (2015), because the resulting prior is less informative than the one from Kessler et al. (2015). However, the approach in Kessler et al. (2015) would be more reasonable when an expert strongly believes that the marginal distribution of θ belongs to a certain parametric family. This belief represents a much larger amount of information than what is considered in the present paper.

I briefly discuss the issue of pooling information from multiple experts. When there are more than one expert opinion, pooling of potentially conflicting information becomes an important issue. A simple pooling mechanisms would be the mixture pooling. For example, suppose that we elicit opinions on the moment of a vector $h(\theta)$ of some functions of θ from two experts, and they give us two conflicting moment conditions $\mathbf{E}h(\theta) = c_1$ and $\mathbf{E}h(\theta) = c_2$. Let $\mathcal{Q}_1^*$ and $\mathcal{Q}_2^*$ be exponentially tilted priors based on the two moment conditions. The mixture pooling considers the mixture prior $\mathcal{Q}^*$ given by

$$d\mathcal{Q}^* = (1-w)\, d\mathcal{Q}_1^* + w\, d\mathcal{Q}_2^*$$

with some weight $w \in [0,\, 1]$. The weights may be determined by some confidence measures on the quality of the expert information. The mixture pooling leads to the mixture of the two moments because $\mathbf{E}^{\mathcal{Q}^*} h(\theta) = (1-w)\, \mathbf{E}^{\mathcal{Q}_1^*} h(\theta) + w\, \mathbf{E}^{\mathcal{Q}_2^*} h(\theta) = (1-w)\, c_1 + w\, c_2$. Another possibility is the exponential pooling. The exponential pooling is to combine the two priors $\mathcal{Q}_1^*$ and $\mathcal{Q}_2^*$ by

$$\log d\mathcal{Q}^* = c + (1-w) \log d\mathcal{Q}_1^* + w \log d\mathcal{Q}_2^*$$

with a normalizing constant c. Because $d\mathcal{Q}_i^*/d\mathcal{P} \propto \exp\{\lambda_i' g(\theta)\}$, where λ_i is the solution from (11) for expert $i = 1, 2$, the exponential pooling would lead to the mixture of the exponential tilting parameters λ_1 and λ_2. The resulting prior $\mathcal{Q}^*$ after pooling information is given by

$$\frac{d\mathcal{Q}^*}{d\mathcal{P}} \propto \exp\{((1-w)\lambda_1' + w\lambda_2')g(\theta)\}.$$

Of course, there are other alternative averaging schemes, and it would be interesting to study relative advantages of different pooling mechanisms. In fact, pooling potentially conflicting opinions is an important issue for far more general problems than the inference of rare events. A thorough discussion of this matter would require a separate paper.

In the following section, I develop a practical method to find $\mathcal{Q}^*$ and implement nonparametric Bayesian inference with $\mathcal{Q}^*$ as a prior distribution.

3 Exponential tilting of Dirichlet process prior

3.1 Estimation of Gibbs measure

We find $\mathcal{Q}^*$ by solving λ_* from the sample version of (11) using simulated θ. To simulate θ, we first generate G from $\mathrm{DP}(\alpha_0 G_0)$ using the stick-breaking process of Sethuraman (1994). Compared to the methods based on the standard Pólya urn Gibbs samplers, the stick-breaking process is particularly useful for our approach because it does not marginalize out θ for posterior simulation. A random measure G from the DP prior is discrete almost surely, and can be written in a countable sum,

$$G = \sum_{j=1}^{\infty} w_j \delta_{\xi_j}, \tag{12}$$

where w_j are random weights independent of ξ_j, and δ_ξ is the distribution concentrated at a random point ξ. An important consequence of the almost sure discreteness of G is that the sample from G shows clustering. If α_0 is small, the clustering becomes more severe, and there will be a fewer number of distinct clusters. The stick-breaking process defines the DP through (12) by

$$\xi_j \overset{\text{iid}}{\sim} G_0, \tag{13}$$

$w_1 = V_1$, and

$$w_j = V_j \prod_{k=1}^{j-1} (1 - V_k), \text{ for } j = 2, 3, \ldots \tag{14}$$

where $V_k \overset{\text{iid}}{\sim} \mathrm{Beta}(1, \alpha_0)$. Intuitively, the stick-breaking scheme starts with a stick with length 1, and keeps cutting the fraction V_j of the remaining stick of length $\prod_{k=1}^{j-1}(1-V_k)$. Then w_j are the lengths of the pieces cut.

In fact, the discrete measure in (12) gives more general classes of random measures than DP priors. By generalizing the probability distribution of V_k, the stick-breaking method can generate other classes of priors such as the two parameter Beta process of Ishwaran and Zarepour (2000) and the two-parameter Poisson–Dirichlet process (Pitman–Yor process) of Pitman and Yor (1997). The priors from the generalized stick-breaking schemes applied to (12) are called the stick-breaking priors which include DP priors as a simple special case. We consider DP priors only in this paper, but the main idea can be easily extended to other stick-breaking priors.

Once we have G, we can calculate θ from (4). Since G defined in (12) is an infinite sum, we would have to approximate G by the truncated stick-breaking method. The method is given by the finite sum

$$G_{\bar{N}} = \sum_{j=1}^{\bar{N}} w_j \delta_{\xi_j}, \tag{15}$$

where w_j and ξ_j are from (13) and (14) except that $V_{\bar{N}} = 1$. Let $\mathcal{P}_{\bar{N}}$ be the distribution of (15) generated by the truncated stick-breaking method. Ishwaran and Zarepour (2002) show that, as $\bar{N} \to \infty$,

$$\iint f(x) G(dx) \mathcal{P}_{\bar{N}}(dG) \to \iint f(x) G(dx) \mathcal{P}(dG),$$

for any bounded continuous real function f. Ishwaran and Zarepour (2000) show that the approximation error of using $\mathcal{P}_{\bar{N}}$ can be substantial when α_0 is large, and $\bar{N}$ should increase as α_0 increases. When the mixing kernel $K(\cdot|\xi)$ is a normal distribution, Theorem 1 of Ishwaran and James (2002) show that the distance $\int |f_{\mathcal{P}_{\bar{N}}}(y) - f_{\mathcal{P}}(y)| dy$ between the marginal densities $f_{\mathcal{P}_{\bar{N}}}(y)$ of y from $\mathcal{P}_{\bar{N}}$ and $f_{\mathcal{P}}(y)$ from $\mathcal{P}$ is proportional to $\exp(-(\bar{N}-1)/\alpha_0)$ for large $\bar{N}$. This implies that $\bar{N}$ should increase proportionally as α_0 increases to get the same level of asymptotic approximation.

Once we have $G_{\bar{N}}$, we can estimate λ_* with the following method. We solve the minimization problem in (11) by substituting the expectation with the Monte Carlo integration calculated with simulated θ using $G_{\bar{N}}$. Let $\{\theta_m^{\bar{N}}\}_{m=1}^M$, where $\theta_m^{\bar{N}} = \sum_{j=1}^{\bar{N}} w_{mj} \varphi(K(y|\xi_{mj}))$, be M values of simulated θ. Although $\{\theta_m^{\bar{N}}\}$ depends on $\bar{N}$, we drop the superscript $\bar{N}$ for notational simplicity. The estimator $\hat{\lambda}_M$ of λ_* is given by

$$\hat{\lambda}_M = \operatorname*{argmin}_{\lambda \in \Lambda} M^{-1} \sum_{m=1}^{M} \exp(\lambda' g(\theta_m)), \tag{16}$$

where Λ is a compact subset of $\mathbb{R}^l$.

The estimator $\hat{\lambda}_M$ above has the following maximum entropy interpretation. Let $\{\pi_m\}_{m=1}^M$ be a discrete probability measure on the points $\{\theta_m\}$ such that $0 \leq \pi_m \leq 1$ and $\sum_{m=1}^M \pi_m = 1$. Without any other constraints, the discrete measure that maximizes the entropy

$$\sum_{m=1}^{M} \pi_m \log(1/\pi_m)$$

is given by the uniform discrete measure with $\pi_m = \frac{1}{M}$ for $m = 1, \ldots, M$. Consider the expert information as the moment condition $\sum_{m=1}^M \pi_m g(\theta_m) = 0$ with respect to the discrete measure. Of course, a discrete measure that incorporates the moment information should have lower entropy than the uniform measure. We can find the discrete measure that maximizes entropy under the expert's moment constraints by solving

$$\max_{(\pi_1,\ldots,\pi_M)} \sum_{m=1}^{M} \pi_m \log(1/\pi_m) \tag{17}$$

subject to

$$\sum_{m=1}^{M} \pi_m g(\theta_m) = 0.$$

The solution $\{\hat{\pi}_m\}$ to (17) is given by

$$\hat{\pi}_m = \frac{\exp\left(\hat{\lambda}'_M g(\theta_m)\right)}{\sum_{m=1}^{M} \exp\left(\hat{\lambda}'_M g(\theta_m)\right)},$$

where the Lagrange multiplier $\hat{\lambda}_M$ turns out to be identical to the vector obtained from (16). The entropy maximization property of the discrete measure $\{\hat{\pi}_m\}$ is well known in the empirical likelihood literature. See Kitamura and Stutzer (1997) and Newey and Smith (2004) for further discussions.

It is easy to show that $\hat{\lambda}_M$ is an asymptotically consistent estimator of λ_* as M, $\bar{N} \to \infty$. Let

$$L_M^{\bar{N}}(\lambda) = M^{-1} \sum_{m=1}^{M} \exp(\lambda' g(\theta_m))$$

be the function to be minimized in (16), and $L(\lambda) = \mathbf{E}^{\mathcal{P}} \exp(\lambda' g(\theta))$ be the objective function from (11). Noting that $\hat{\lambda}_M = \operatorname{argmin}_{\lambda \in \Lambda} L_M^{\bar{N}}(\lambda)$ and $\lambda_* = \operatorname{argmin}_{\lambda} L(\lambda)$, we prove $\hat{\lambda}_M \overset{a.s.}{\to} \lambda_*$ as M, $\bar{N} \to \infty$.

Assumption 1. *The solution $\lambda_* = \operatorname{argmin}_{\lambda} \mathbf{E}^{\mathcal{P}} \exp(\lambda' g(\theta))$ of (11) is unique in Λ.*

The above assumption ensures that the expert moment conditions are not linearly dependent. Essentially, it implies that there is no redundant expert information.

Assumption 2. *$L_M^{\bar{N}}(\lambda) \overset{a.s.}{\to} L(\lambda)$ uniformly as M, $\bar{N} \to \infty$.*

Assumption 3. $L(\lambda)$ is continuous on Λ.

Theorem 4. *Let $\hat{\pi}^M$ be the empirical distribution of $\{\hat{\pi}_m\}$, and π^* be the change of measure in (10). If Assumptions 1–3 hold, as M, $\bar{N} \to \infty$, we have $\hat{\lambda}_M \overset{a.s.}{\to} \lambda_*$.*

Proof. We use the argument in Theorem 4.2.1 of Bierens (1994). We first show

$$L(\hat{\lambda}_M) \overset{a.s.}{\to} L(\lambda_*).$$

We have

$$\begin{aligned} 0 \le L(\hat{\lambda}_M) - L(\lambda_*) &= L(\hat{\lambda}_M) - L_M^{\bar{N}}(\hat{\lambda}_M) + L_M^{\bar{N}}(\hat{\lambda}_M) - L(\lambda_*) \\ &\le L(\hat{\lambda}_M) - L_M^{\bar{N}}(\hat{\lambda}_M) + L_M^{\bar{N}}(\lambda_*) - L(\lambda_*) \\ &\le 2 \sup_{\lambda \in \Lambda} |L_M^{\bar{N}}(\lambda) - L(\lambda)| \overset{a.s.}{\to} 0, \end{aligned} \tag{18}$$

as M, $\bar{N} \to \infty$, from $L_M^{\bar{N}}(\hat{\lambda}_M) \le L_M^{\bar{N}}(\lambda_*)$ and almost sure uniform convergence of $L_M^{\bar{N}}(\lambda)$ to $L_M^{\bar{N}}(\lambda)$. Since $L(\lambda)$ is continuous and λ_* is unique, we have $\hat{\lambda}_M \overset{a.s.}{\to} \lambda_*$ from (18). □

3.2 Implementation with a Metropolis–Hastings algorithm

Once we obtain $\hat{\lambda}_M$ from (16), we can proceed with posterior simulations. I present a Metropolis–Hastings (M–H) method to draw a sample from the posterior distribution with the exponentially tilted DP prior. Let $\mathbf{y} = (y_1, \ldots, y_n)$ be data and

$$\Psi(dG|\mathbf{y}) \propto \prod_{i=1}^{n} f(y_i|G)\mathcal{P}(dG),$$

be the density Ψ of the posterior distribution of G in the DPM model. With expert information, the posterior becomes

$$\Psi^*(dG|\mathbf{y}) \propto \prod_{i=1}^{n} f(y_i|G)\pi^*(\theta)\mathcal{P}(dG) = \prod_{i=1}^{n} f(y_i|G)\mathcal{Q}^*(dG).$$

We use an MCMC method to get simulated observations from the posterior distribution Ψ^*. Because of the simple relationship between $\mathcal{P}$ and $\mathcal{Q}^*$, the independence chain M–H algorithm is particularly convenient for our problem. See Chib and Greenberg (1994) and Chib and Greenberg (1995) for a review on various M–H algorithms. Considering the relationship

$$\frac{\Psi^*(dG|\mathbf{y})}{\Psi(dG|\mathbf{y})} = \pi^*,$$

we use the posterior distribution Ψ with the original DP prior $\mathcal{P}$ as the proposal distribution of the independence chain M–H algorithm. Then the acceptance of tth sample $G^{(t)}$ in the MCMC chain depends on the quantity

$$\frac{\pi^*(\theta^{(t)})}{\pi^*(\theta^{(t-1)})},$$

which is approximated by

$$\frac{\exp\left(\hat{\lambda}_M' g(\theta^{(t)})\right)}{\exp\left(\hat{\lambda}_M' g(\theta^{(t-1)})\right)},$$

with the estimated $\hat{\lambda}_M$ following the method described in the previous section. The tth draw of $G^{(t)}$ is accepted with probability

$$\min\left\{1, \frac{\exp\left(\hat{\lambda}_M' g(\theta^{(t)})\right)}{\exp\left(\hat{\lambda}_M' g(\theta^{(t-1)})\right)}\right\},$$

or $G^{(t)} = G^{(t-1)}$ if rejected.

For sampling from the proposal distribution Ψ, we use the blocked Gibbs algorithm of Ishwaran and James (2001). The implementation of our approach combines the blocked Gibbs algorithm and the independence chain M–H algorithm above. The posterior simulation cycles through the following steps. Let $\xi^{\bar{N}} = \{\xi_j\}_{j=1}^{\bar{N}}$ and $w^{\bar{N}} = \{w_j\}_{j=1}^{\bar{N}}$ from

$G_{\bar{N}}$ in (15). Let $\xi^* = \{\xi_i^*\}_{i=1}^n$ be the vector of n i.i.d. random variables $\xi_i^* \sim G_{\bar{N}}$ which are used to draw $y_i|\xi_i^* \sim K_{\xi_i^*}$. Since $G_{\bar{N}}$ is discrete, ξ_i^* are drawn from the elements of $\xi^{\bar{N}}$. Therefore, some ξ_i^* may be same and some elements of $\xi^{\bar{N}}$ may not be drawn at all. For each $j = 1, \ldots, \bar{N}$, define the set $I_j = \{i : \xi_i^* = \xi_j\}$ of the indices of ξ_i^* that are equal to ξ_j. When ξ_j was not drawn at all, then $I_j = \varnothing$.

1. Updating $\xi^{\bar{N}}$ given $w^{\bar{N}}$, ξ^*, and $\mathbf{y}$. For $j = 1, 2, \ldots, \bar{N}$, simulate $\xi_j \sim G_0$ for all j with $I_j = \varnothing$, and draw ξ_j for $I_j \neq \varnothing$ from the density
$$p(\xi_j|\xi^*, w^{\bar{N}}, \mathbf{y}) \propto G_0(d\xi_j) \prod_{i \in I_j} k(y_i|\xi_j),$$
where $k(\cdot|\cdot)$ is the density of the kernel function $K(\cdot|\cdot)$ given in (2) and (3).

2. Updating ξ^* given $\xi^{\bar{N}}$, $w^{\bar{N}}$, and $\mathbf{y}$. Draw $(\xi_i^*|\xi^{\bar{N}}, w^{\bar{N}}, \mathbf{y})$ independently from
$$\xi_i^*|(\xi^{\bar{N}}, w^{\bar{N}}, \mathbf{y}) \sim \sum_{j=1}^{\bar{N}} w_j \delta_{\xi_j}.$$

3. Updating $w^{\bar{N}}$ given $\xi^*, \xi^{\bar{N}}$, and $\mathbf{y}$. Draw $w_1 = V_1$, and
$$w_j = V_j \prod_{k=1}^{j-1} (1 - V_k), \text{ for } j = 2,\, 3, \ldots, \bar{N} - 1,$$
where $V_{\bar{N}} = 1$, $V_k \overset{\text{iid}}{\sim} \text{Beta}(1 + Card(I_k),\ \alpha_0 + \sum_{k+1}^{\bar{N}} Card(I_k))$ for $k = 1, \ldots, \bar{N}-1$, and $Card(I_k)$ represents the cardinality of the set I_k.

4. Calculate the tth MCMC sample $g(\theta^{(t)})$ from (4) replacing G with $G_{\bar{N}}$ defined from $(\xi^{\bar{N}}, w^{\bar{N}})$. The tth sample$(\xi^*, \xi^{\bar{N}}, w^{\bar{N}})$ is accepted with probability
$$\min \left\{ 1, \frac{\exp\left(\hat{\lambda}_M' g(\theta^{(t)})\right)}{\exp\left(\hat{\lambda}_M' g(\theta^{(t-1)})\right)} \right\},$$
or replaced by the previous sample.

In Step 1, the updating of $\xi^{\bar{N}}$ is simple if the base distribution G_0 and the kernel function K are conjugate. For this reason, we use normal–gamma conjugate distributions for the examples and empirical applications in this paper. Steps 1–3 are the original blocked Gibbs sampler, and our approach simply adds Step 4 for exponential tilting.

4 Dirichlet process mixture of normal distributions

To demonstrate the theory in this paper, we consider the DPM model with the family of normal distributions as a kernel function. A normal mixture model is convenient and popularly used because the conjugacy with the normal–gamma distribution, often used for G_0, makes the MCMC procedure simple. Also, the normal kernel function can generate fat-tailed distributions such as t-distributions with small degrees of freedom by variance mixture with inverse Gamma distributions, which is a useful property for inference of tail-risk events. We define the kernel function

$$K(y|\xi) = \Phi(y|\mu,\, 1/\tau),$$

where $\xi = (\mu,\, \tau)$ is defined on $\Xi = \mathbb{R} \times \mathbb{R}^+$, and $\Phi(\cdot|\mu,\, 1/\tau)$ is a normal distribution function with mean μ and variance $1/\tau$, where τ is a precision parameter. The prior $\mathrm{DP}(\alpha_0 G_0)$ on $\mathcal{G}$ is given by a concentration parameter α_0 and a base distribution G_0 on $(\mu, \tau) \in \mathbb{R} \times \mathbb{R}^+$ distributed as a normal–gamma distribution $G_0 = \mathbf{NG}(\mu_0,\, n_0,\, \nu_0,\, \sigma_0^2)$ for which

$$\mu|\tau \sim \mathbf{N}(\mu_0, (n_0\tau)^{-1}),$$

and

$$\tau \sim \Gamma(\nu_0/2,\, \nu_0\sigma_0^2/2),$$

where $\Gamma(a,\, b)$ is a gamma distribution with the shape parameter a and the rate parameter b. Therefore, $\mathrm{DP}(\alpha_0 G_0)$ is defined with 5 parameters $(\alpha_0,\, \mu_0,\, n_0,\, \nu_0,\, \sigma_0^2)$. The normal–gamma distribution is a convenient choice for the DPM of normal distribution because of its conjugacy with the normal kernel function. Note that α_0 represents the concentration of G around G_0, n_0 is related to the concentration of μ around μ_0 of G_0, and ν_0 is for the concentration of τ in G_0. Therefore, $(\alpha_0,\, n_0,\, \nu_0)$ jointly determines the effective sample size of the prior distribution, and smaller values of these parameters imply more diffuse priors.

In our numerical example, we assume a mixture of two normal distributions as the true distribution. We set $n = 20$ and generate i.i.d. data $\{y_i\}_{i=1}^{20}$ from the mixture of $\mathbf{N}(1, 1)$ and $\mathbf{N}(15, 1)$ with weights 5% and 95%, respectively. Specifically, $y_1, \ldots,\, y_{20}$ are calculated from

$$y_i = W_i Y_{1i} + (1 - W_i)Y_{2i}, \tag{19}$$

where $W_i = 1$ with probability 5% or 0 with 95%, and $Y_{1i} \sim \mathbf{N}(1, 1)$ and $Y_{2i} \sim \mathbf{N}(15, 1)$. This distribution has a small bump on the left tail which is hard to discover in data. We set the DP prior $\mathrm{DP}(\alpha_0 G_0)$ with $\mu_0 = 15$, $n_0 = 1$, $\nu_0 = 6$, $\sigma_0^2 = 15$, and $\alpha_0 = 4$. We use $\bar{N} = 80$ for $G_{\bar{N}}$ to approximate G. The parameter of interest θ is the left-tail probability $\theta = \mathbf{P}\{y_i \le 0\}$ given G, which would represent the probability of default if y_i is an equity value. The true value of θ is 0.793% calculated from (19). We assume that the expert information is

$$\mathbf{P}\{\theta \le 0.01\} = 50\% \quad \text{and} \quad \mathbf{P}\{\theta \le 0.005\} = 25\%.$$

This means $g(\theta) = (I\{\theta \le 0.01\} - 0.5,\ I\{\theta \le 0.005\} - 0.25)'$ for the moment condition $\mathbf{E}g(\theta) = 0$.

For estimation of λ_*, we first set $M = 5{,}000{,}000$ and simulate $\{\theta_m\}_{m=1}^{M}$ by using the truncated stick-breaking process in (15) with $\bar{N} = 80$ from $\mathrm{DP}(\alpha_0 G_0)$. Since a larger value of M would increase the precision of the estimator $\hat{\lambda}_M$, in practice, one can experiment with different values of M to check if $\hat{\lambda}_M$ is relatively stable for $M > M_0$ for a large M_0. For the examples in this paper, I find that, roughly speaking, $M > 100{,}000$ is sufficiently large. To implement the exponential tilting, we solve

$$\hat{\lambda}_M = \underset{\lambda \in \Lambda}{\operatorname{argmin}}\, M^{-1} \sum_{m=1}^{M} \exp(\lambda' g(\theta_m)),$$

to get $\hat{\lambda}'_M = (0.038, -0.66)$ from the simulated $\{\theta_m\}_{m=1}^{M}$. Since $\pi^* \propto \exp(\lambda'_* g(\theta))$, we can get some idea of the direction of the exponential tilting with the estimated sign of $\lambda'_* g(\theta)$. When $\theta > 0.01$, we have $\hat{\lambda}'_M g(\theta) = 0.146 > 0$, but $\hat{\lambda}'_M g(\theta) = -0.476 < 0$ if $\theta \leq 0.005$. This implies that the exponential tilting of $\mathrm{DP}(\alpha_0 G_0)$ would put higher weights on large θ than small θ by tilting the prior distribution of θ toward $\theta = 1$. In other words, the expert emphasizes that θ should be larger than what $\mathrm{DP}(\alpha_0 G_0)$ would generate. Consequently, the prior that complies with the expert information favors distributions with larger left-tail probabilities than $\mathrm{DP}(\alpha_0 G_0)$.

Since we have $\hat{\lambda}_M$, we can draw a sample from the posterior distribution using the M–H algorithm described in the previous section. For the MCMC procedure, we perform 500,100 MCMC iterations, discard the first 100 iterations for burn-in, and use every 10th state for thinning after the burn-in period. After thinning, we get 50,000 draws of θ.

In the first row of Figure 2, the prior (left panel) and posterior (right panel) densities of θ are shown with and without expert information, labeled as “Expert” and “DP”, respectively. The vertical lines represent the true value of θ. As expected from the estimated exponential tilting parameter $\hat{\lambda}_M$, the posterior distribution of θ is tilted toward large θ. The second row shows the first 500 MCMC draws of θ from the posterior simulations without (left panel) and with (right panel) the expert information to check good mixing of the MCMC sample. The effective sample sizes are calculated from the full 50,000 MCMC draws.

An interesting quantity in the analysis of rare events is the number of occurrences of rare events given a sample size. For the probability of default θ, the probability that we observe k defaults out of n i.i.d. observations is given by the binomial distribution

$$H_k^n = \binom{n}{k} \theta^k (1-\theta)^k. \tag{20}$$

We can calculate the posterior distribution of the probability H_1^{20} of having one default out of 20 observations. The third row of Figure 2 shows the prior and posterior densities of H_1^{20} with and without the expert information labeled as “Expert” and “DP”, respectively. Since the expert information favors the distributions with large left-tail probabilities, the distribution of the probability H_1^{20} with the expert information (“Expert”) is tilted toward $\theta = 1$ compared to the distribution without the expert information (“DP”).

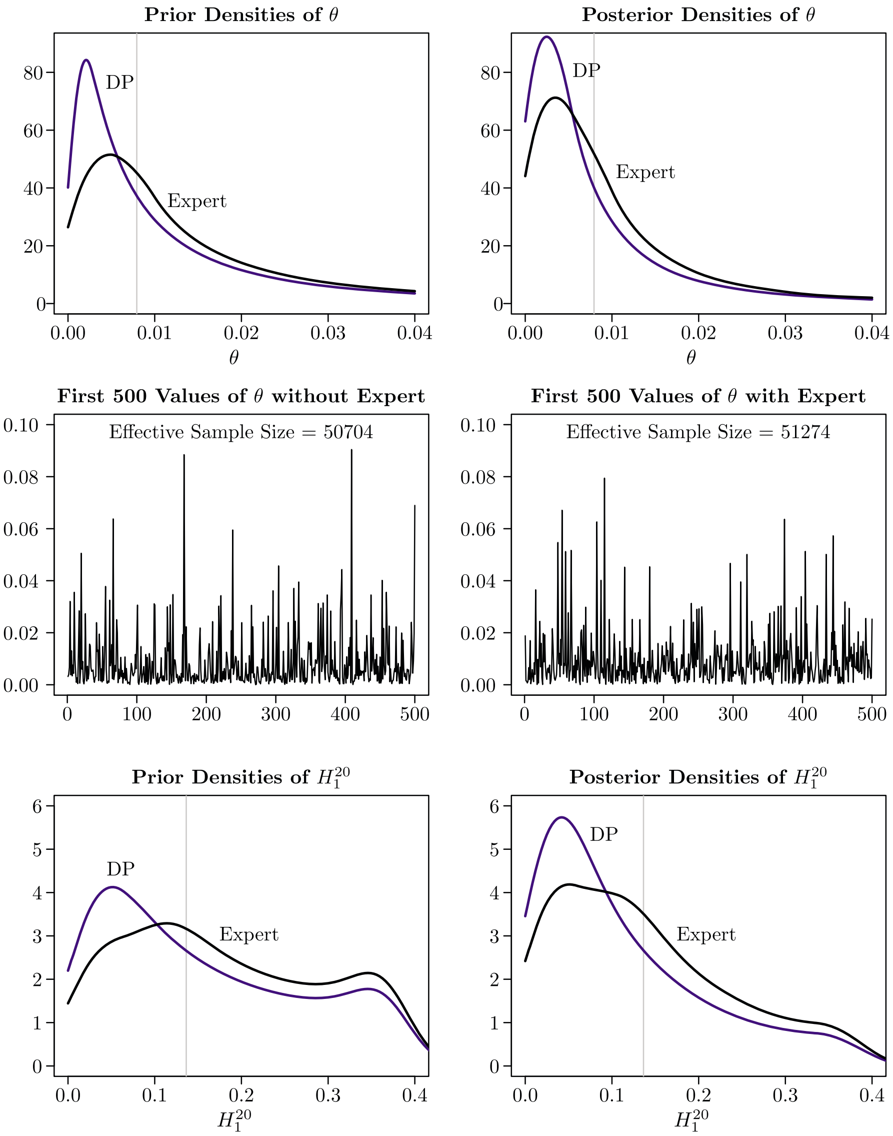

Figure 2: In the first row, we have the prior (left panel) and the posterior (right panel) distributions of θ with and without expert information (labeled as "Expert" and "DP", respectively). The vertical lines represent the true value of θ. The second row shows some values of θ from the posterior simulation without (left panel) and with (right panel) expert information. The third row shows the prior (left panel) and the posterior (right panel) distributions of the probability of one default out of 20 observations with ("Expert") and without ("DP") expert information.

Date	Return	Date	Return	Date	Return
10/21/1957	−2.93	5/24/1976	−1.80	12/18/1995	−1.55
11/24/1958	−2.60	7/27/1977	−1.63	3/8/1996	−3.08
8/10/1959	−2.09	10/31/1978	−2.01	10/27/1997	−6.87
9/19/1960	−2.27	10/9/1979	−2.96	8/31/1998	−6.80
4/24/1961	−2.08	3/17/1980	−3.01	10/15/1999	−2.81
5/28/1962	−6.68	8/24/1981	−2.89	4/14/2000	−5.83
11/22/1963	−2.81	10/25/1982	−3.97	9/17/2001	−4.92
2/5/1964	−2.70	1/24/1983	−2.70	9/3/2002	−4.15
6/28/1965	−1.76	2/8/1984	−1.82	3/24/2003	−3.52
8/29/1966	−2.46	7/29/1985	−1.46	8/5/2004	−1.63
5/31/1967	−1.56	9/11/1986	−4.81	4/15/2005	−1.67
3/14/1968	−1.90	10/19/1987	−20.47	1/20/2006	−1.83
7/28/1969	−1.90	1/8/1988	−6.77	2/27/2007	−3.47
5/4/1970	−3.00	10/13/1989	−6.12	10/15/2008	−9.03
6/18/1971	−1.52	8/6/1990	−3.02	1/20/2009	−5.28
5/9/1972	−1.32	11/15/1991	−3.66	5/20/2010	−3.90
11/19/1973	−3.05	4/7/1992	−1.86	8/8/2011	−6.66
11/18/1974	−3.67	2/16/1993	−2.40	6/1/2012	−2.46
3/24/1975	−2.36	2/4/1994	−2.27	6/20/2013	−2.50

Table 1: Worst daily returns (%) of the year from 1957 to 2013. Max= −1.32%, min= −20.47%, mean= −3.55%, median= −2.81%, and standard deviation= 2.87%.

5 Application to extreme losses from stock index investment

We consider inference of tail probabilities for stock index returns. We use the daily S&P500 index data (series ID: SP500) from the Federal Reserve Economic Data (FRED) maintained by the Federal Reserve Bank of St. Louis. We define the daily S&P500 index return on day d as $r_d = (p_d - p_{d-1})/p_{d-1}$ with daily S&P500 index values p_d ignoring missing values and holidays. Our main variable of interest is

$$X_t = \text{Worst daily return } r_d \text{ of the year } t,$$

for $t = 1957 \sim 2013$. We get 57 observations of X_t.

Table 1 shows the sample. The sample mean and median are −3.55% and −2.81%, respectively, the lower and upper empirical quartiles are −3.90% and −1.90%, and the sample standard deviation is 2.87%. The sample shows high skewness. The worst daily return −20.47% occurred on the black Monday 1987/10/19. The second worst daily return −9.03% was on 2008/10/15 during the recent recession. We are interested in the probability that the worst daily return of the next year would be lower than −10%, i.e.,

$$\theta = \mathbf{P}\{X_{2014} < -10\%\}.$$

For normalization, we use

$$Y_t = -\log(-X_t)$$

instead of X_t after assuming $\mathbf{P}(X_t < 0) = 1$. All observations are assumed to be i.i.d. given a mixing measure G. Suppose that our expert thinks that the chance that the worst daily return in 2014 falls below -10% is equally likely to be above or below 4%. If we condition that θ is below 4%, then the expert thinks that the chances of θ being above and below 2% are equally likely. This expert information is summarized by $\mathbf{P}\{\theta < 0.04\} = 50\%$ and $\mathbf{P}\{\theta < 0.02\} = 25\%$. As before, we use the DPM of normal distributions for Y_t. The DP prior $\mathrm{DP}(\alpha_0 G_0)$ used for this empirical application is from $\alpha_0 = 4$, and the normal–gamma base distribution $G_0 = \mathbf{NG}(\mu_0, n_0, \nu_0, \sigma_0^2)$ with $\mu_0 = 3.5$, $n_0 = 1$, $\nu_0 = 8$, $\sigma_0^2 = .25$.

Then we estimate the exponential tilting parameter λ_* from the minimization problem in (16) with simulated θ from $\mathrm{DP}(\alpha_0 G_0)$. As in the example in the previous section, we simulate five million values of θ with the truncated stick-breaking process in (15) with $\bar{N} = 80$, and get $\hat{\lambda}_M = (0.117, 0.505)$. From the estimated tilting parameter, we have $\hat{\lambda}'_M g(\theta) < 0$ for large θ but $\hat{\lambda}'_M g(\theta) > 0$ for small θ. Therefore, we know from $\pi^* \propto \exp(\lambda'_* g(\theta))$ that the exponentially tilted prior puts lower weights for large θ than small θ relative to the DP prior. This implies that the expert thinks that the probability of θ is generally smaller than what the DP prior suggests. Using $\hat{\lambda}_M$, we proceed to the posterior simulation with the M–H algorithm in Section 3. All the other settings such as burn-in and thinning parameters are identical to the example given in the previous section.

In the first row in Figure 3, the prior densities of θ with and without expert information are shown as “Expert” and “DP”, respectively, in the left panel. As discussed, the expert prior shows tilting toward small θ compared to the DP prior. The posterior densities with and without expert information in the right panel also show the effect of exponential tilting. The second row shows the first 500 MCMC draws of θ from the posterior simulations without (left panel) and with (right panel) expert information, which suggest good mixing of the MCMC sample. The effective sample sizes are calculated from the full 50,000 MCMC draws. The left panel of the third row in the figure shows the DP prior (“DP prior”) and posterior densities of the probability H_1^{57}, as defined in (20), that we observe exactly one year with the event $\{X_t < -10\%\}$ in the next 57 years. The posterior densities with and without the expert information are labeled as “Expert” and “DP”, respectively. With expert’s information, there is a slightly more chance to have θ around 3.5%. The right panel of the third row is the DP prior and the posterior densities with and without the expert information of the probability that the worst daily return of the year is lower than -10% during the next 57 years, which is equal to $1 - H_0^{57}$. With the expert opinion incorporated, we can see that the chance of having a large value of $1 - H_0^{57}$ is slightly smaller than what is from the posterior without expert information.

6 Conclusion and extensions

The method proposed in this paper deals with the problem of potential misspecification in parametric models and the scarcity of data information for inference of rare events

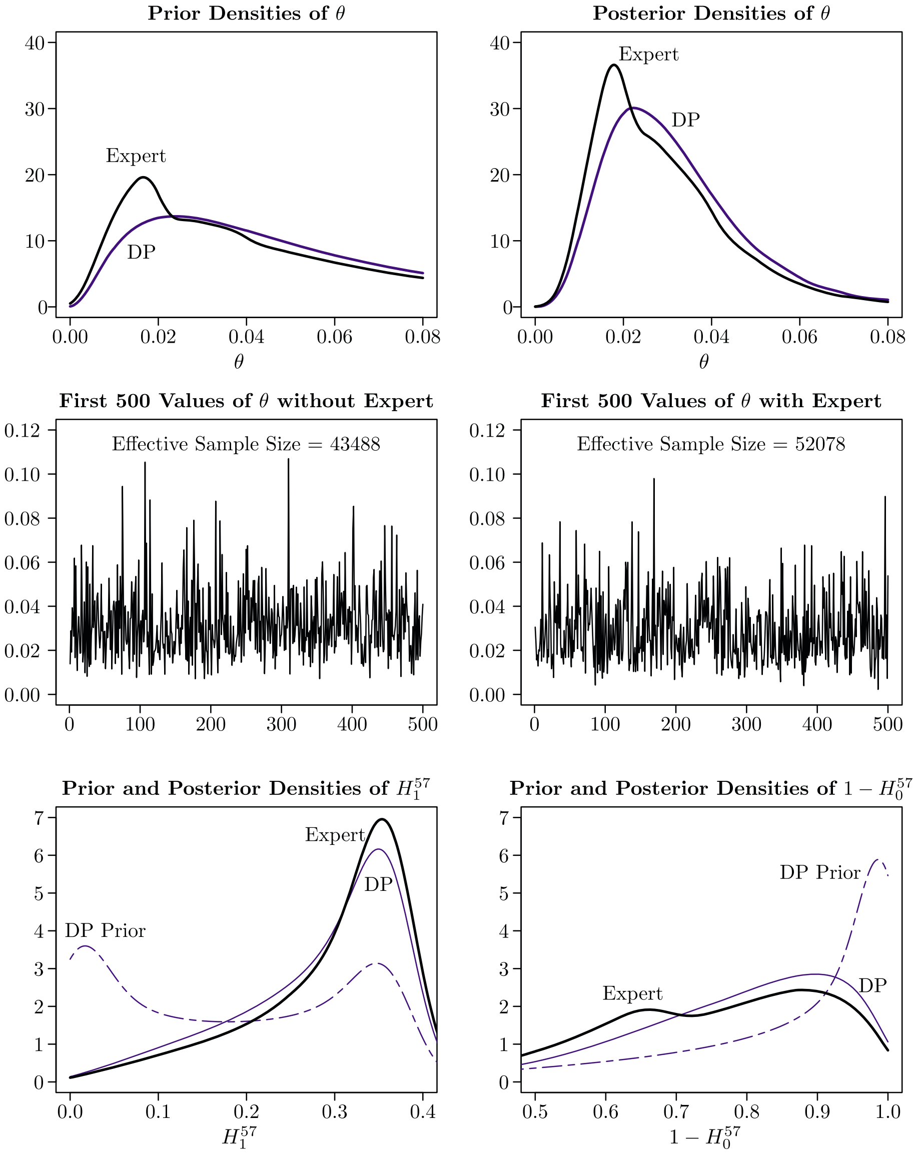

Figure 3: In the first row, we have the prior (left panel) and the posterior (right panel) distributions of θ with and without expert information (labeled as "Expert" and "DP", respectively). The second row shows some values of θ from the posterior simulation without (left panel) and with (right panel) expert information. The third row shows the DP prior (labeled as "DP prior") and posterior distributions with and without expert information (labeled as "Expert" and "DP", respectively) of the probability that $\{X_t < -10\%\}$ occurs exactly once in the next 57 years (left panel) and the probability that $\{X_t < -10\%\}$ occurs at least once in the next 57 years (right panel).

by using nonparametric models and incorporating expert information. Elicitation of expert information can be easily done with carefully designed interviews, and the prior that comply with the expert information can be obtained with the straightforward M–H method proposed in this paper. Therefore, when historical data are scarce and it is hard to find a reasonable parametric model, this paper's approach could be an attractive choice.

The main idea of this paper can be extended in several directions. First, an interesting topic would be an extension to a semiparametric model defined with a set of moment conditions on data such as $\int m(y;\beta)\, f(y|G)\, dy = 0$, where the function $m(y;\beta)$ is an γ_m-dimensional vector of moments of y for a γ_β-dimensional vector β of parameter of interest. Note that β, which is of interest to a modeler, may not be same as θ, which is of interest to an expert. When $\gamma_m = \gamma_\beta$, the parameter β is exactly identified by solving $\int m(y;\beta)\, f(y|G)\, dy = 0$ for each given G. This implies that the condition is not binding for G at all, and does not restrict the prior on G. But if $\gamma_m > \gamma_\beta$, we have over-identifying moment conditions that some G may not satisfy. Then the prior must be modified, so that G that violates the moment conditions are excluded. Therefore, imposing over-identifying moment conditions on data is equivalent to considering only the priors with a lower dimensional support, which entails a degenerate prior. For example, the zero-mean condition $\int y f(y|G)\, dy = 0$ reduces the dimension of the support of a prior $\mathcal{P}$ because some $G \sim \mathcal{P}$ that violate this moment condition must be excluded. Compare this with the expert's moment condition $\mathbf{E}(\theta) = 0$, where $\theta \equiv \int y f(y|G)\, dy$ is the mean. While the expert information reduces the dimension of the space of priors only, the zero-mean condition on data would reduce the dimension of the support of priors by restricting the prior space to the subspace with $\theta = 0$. In that sense, over-identifying moment conditions on data are an extremely strong form of expert information with which the expert is sure on certain aspects of data. Conceptually, this idea can be implemented in the following way. We first begin with the DP prior $\mathrm{DP}(\alpha_0 G_0)$. A random draw $G \in \mathcal{G}$ from $\mathrm{DP}(\alpha_0 G_0)$ would not satisfy $\int m(y;\beta)\, f(y|G)\, dy = 0$ for any β if $\gamma_m > \gamma_\beta$. Therefore, G may be projected onto the sub-space $\bar{\mathcal{G}}$ that satisfies the moment condition on data. For this projection, we can use Amari's 1-projection. We would first find the projection $G^*(\beta)$ that is closest from G with respect to the 1-divergence for each β. Among all $G^*(\beta)$, we may pick $G^*(\beta^*)$ that is closest to G with respect to the 1-divergence as the final projection. Applying this projection for all $G \in \mathcal{G}$ defines a degenerate prior $\mathcal{P}^*$ from which a random draw is in $\bar{\mathcal{G}}$ with probability one. Once we have the prior $\mathcal{P}^*$, then we can incorporate expert information by applying the 1-projection of $\mathcal{P}^*$ onto $\mathbb{Q}$ again. Essentially, this approach uses the information in moment conditions on data followed by expert information. See Kitamura and Otsu (2011) for an alternative framework for semiparametric models.

Another interesting topic for future research is to consider covariates in our approach. For example, we can consider implementing our method to Bayesian density regressions with DP priors that depend on covariates. MacEachern (1999, 2000) consider dependent Dirichlet processes and more general dependent nonparametric processes that include the hierarchical Dirichlet processes, nonparametric density regressions, and other models as special cases. Bush and MacEachern (1996) consider using a DP as the distribution of random effects that vary over sub-blocks of a sample. Li et al. (2011) propose a technique for identifiability of inference in a nonparametric Bayesian random effect model. It would be interesting to investigate how my approach extends to these settings.

References

Amari, S. I. (1982). "Differential geometry of curved exponential families – curvatures and information loss." *The Annals of Statistics*, 10(2): 357–385. MR0653513. 427, 428

— (1985). *Differential-Geometrical Methods in Statistics*. Lecture Notes in Statistics. Berlin: Springer-Verlag. MR0788689. doi: http://dx.doi.org/10.1007/978-1-4612-5056-2. 427

Amari, S. I. and Nagaoka, H. (2000). *Methods of Information Geometry*, volume 191 of *Translations of Mathematical Monographs*. Providence, RI: American Mathematical Society. Originally published in Japanese by Iwanami Shoten, Publishers, Tokyo, 1993. MR1800071. 427

Antoniak, C. (1974). "Mixtures of Dirichlet processes with applications to Bayesian nonparametric problems." *The Annals of Statistics*, 2(6): 1152–1174. MR0365969. 422

Bierens, H. (1994). *Topics in Advanced Econometrics: Estimation, Testing, and Specification of Cross-Section and Time Series Models*. Cambridge University Press. MR1291390. doi: http://dx.doi.org/10.1017/CBO9780511599279. 432

Blackwell, D. and MacQueen, J. (1973). "Ferguson distributions via Pólya urn schemes." *The Annals of Statistics*, 1(2): 353–355. MR0362614. 422, 423

Bush, C. A. and MacEachern, S. N. (1996). "A semiparametric Bayesian model for randomised block designs." *Biometrika*, 83(2): 275–285. 441

Chaloner, K. and Duncan, G. (1983). "Assessment of a beta prior distribution: PM elicitation." *The Statistician*, 32(1/2): 174–180. 423

Chib, S. and Greenberg, E. (1994). "Bayes inference in regression models with ARMA(p, q) errors." *Journal of Econometrics*, 64(1–2): 183–206. MR1310523. doi: http://dx.doi.org/10.1016/0304-4076(94)90063-9. 433

— (1995). "Understanding the Metropolis–Hastings algorithm." *The American Statistician*, 49(4): 327–335. 433

Cressie, N. and Read, T. R. C. (1984). "Multinomial goodness-of-fit tests." *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 46(3): 440–464. MR0790631. 427

Csiszár, I. (1967a). "Information type measures of difference of probability distributions and indirect observations." *Studia Scientiarum Mathematicarum Hungarica*, 2: 299–318. MR0219345. 427

— (1967b). "On topological properties of f-divergence." *Studia Scientiarum Mathematicarum Hungarica*, 2: 329–339. 427

— (1975). "*I*-divergence geometry of probability distributions and minimization problems." *The Annals of Probability*, 3: 146–158. MR0365798. 426, 427

Escobar, M. (1994). "Estimating normal means with a Dirichlet process prior." *Journal of the American Statistical Association*, 89(425): 268–277. MR1266299. 423

Escobar, M. and West, M. (1995). "Bayesian density estimation and inference using mixtures." *Journal of the American Statistical Association*, 90(430): 577–588. MR1340510. 423

Ferguson, T. (1973). "A Bayesian analysis of some nonparametric problems." *The Annals of Statistics*, 1(2): 209–230. MR0350949. 422

— (1974). "Prior distributions on spaces of probability measures." *The Annals of Statistics*, 2(4): 615–629. MR0438568. 422

Freedman, D. (1963). "On the asymptotic behavior of Bayes' estimates in the discrete case." *The Annals of Mathematical Statistics*, 34(4): 1386–1403. MR0158483. 422

Garthwaite, P., Kadane, J., and O'Hagan, A. (2005). "Statistical methods for eliciting probability distributions." *Journal of the American Statistical Association*, 100(470): 680–701. MR2170464. doi: http://dx.doi.org/10.1198/016214505000000105. 423

Gelfand, A. and Kottas, A. (2002). "A computational approach for full nonparametric Bayesian inference under Dirichlet process mixture models." *Journal of Computational and Graphical Statistics*, 11(2): 289–305. MR1938136. doi: http://dx.doi.org/10.1198/106186002760180518. 425

Gelfand, A. and Mukhopadhyay, S. (1995). "On nonparametric Bayesian inference for the distribution of a random sample." *Canadian Journal of Statistics*, 23(4): 411–420. MR1378937. doi: http://dx.doi.org/10.2307/3315384. 425

Ghosh, J. and Ramamoorthi, R. (2003). *Bayesian Nonparametrics*. Springer Verlag. MR1992245. 422

Griffin, J. and Steel, M. (2011). "Stick-breaking autoregressive processes." *Journal of Econometrics*, 162(2): 383–396. MR2795625. doi: http://dx.doi.org/10.1016/j.jeconom.2011.03.001. 423

Hirano, K. (2002). "Semiparametric Bayesian inference in autoregressive panel data models." *Econometrica*, 70(2): 781–799. MR1913831. doi: http://dx.doi.org/10.1111/1468-0262.00305. 423

Hjort, N. L., Holmes, C., Müller, P., and Walker, S. G. (eds.) (2010). *Bayesian Nonparametrics*. Cambridge University Press. MR2722987. doi: http://dx.doi.org/10.1017/CBO9780511802478. 422

Ishwaran, H. and James, L. (2001). "Gibbs sampling methods for stick-breaking priors." *Journal of the American Statistical Association*, 96(453): 161–173. MR1952729. doi: http://dx.doi.org/10.1198/016214501750332758. 423, 433

— (2002). “Approximate Dirichlet process computing in finite normal mixtures.” *Journal of Computational and Graphical Statistics*, 11(3): 508–532. MR1938445. doi: http://dx.doi.org/10.1198/106186002411. 431

Ishwaran, H. and Zarepour, M. (2000). “Markov chain Monte Carlo in approximate Dirichlet and beta two-parameter process hierarchical models.” *Biometrika*, 87(2): 371–390. MR1782485. doi: http://dx.doi.org/10.1093/biomet/87.2.371. 423, 430, 431

— (2002). “Exact and approximate sum representations for the Dirichlet process.” *Canadian Journal of Statistics*, 30(2): 269–283. MR1926065. doi: http://dx.doi.org/10.2307/3315951. 431

Jensen, M. and Maheu, J. (2010). “Bayesian semiparametric stochastic volatility modeling.” *Journal of Econometrics*, 157(2): 306–316. MR2661603. doi: http://dx.doi.org/10.1016/j.jeconom.2010.01.014. 423

Kadane, J., Chan, N., and Wolfson, L. (1996). “Priors for unit root models.” *Journal of Econometrics*, 75(1): 99–111. MR1414505. doi: http://dx.doi.org/10.1016/0304-4076(95)01771-2. 423

Kadane, J. and Wolfson, L. (1998). “Experiences in elicitation.” *The Statistician*, 47(1): 3–19. 423, 425

Kessler, D. C., Hoff, P. D., and Dunson, D. B. (2015). “Marginally specified priors for non-parametric Bayesian estimation.” *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 77(1): 35–58. MR3299398. doi: http://dx.doi.org/10.1111/rssb.12059. 429

Kiefer, N. (2009). “Default estimation for low-default portfolios.” *Journal of Empirical Finance*, 16(1): 164–173. 421

— (2010). “Default estimation and expert information.” *Journal of Business and Economic Statistics*, 28(2): 320–328. MR2681204. doi: http://dx.doi.org/10.1198/jbes.2009.07236. 421

Kitamura, Y. and Otsu, T. (2011). “Bayesian analysis of moment condition models using nonparametric priors.” Working paper, Yale University. 441

Kitamura, Y. and Stutzer, M. (1997). “An information-theoretic alternative to generalized method of moments estimation.” *Econometrica*, 65(4): 861–874. MR1458431. doi: http://dx.doi.org/10.2307/2171942. 426, 432

Kullback, S. and Leibler, R. (1951). “On information and sufficiency.” *The Annals of Mathematical Statistics*, 22(1): 79–86. MR0039968. 426

Li, Y., Müller, P., and Lin, X. (2011). “Center-adjusted inference for a nonparametric Bayesian random effect distribution.” *Statistica Sinica*, 21: 1201–1223. 441

Lo, A. (1984). “On a class of Bayesian nonparametric estimates: I. Density estimates.” *The Annals of Statistics*, 12(1): 351–357. MR0733519. doi: http://dx.doi.org/10.1214/aos/1176346412. 422

MacEachern, S. and Müller, P. (1998). "Estimating mixture of Dirichlet process models." *Journal of Computational and Graphical Statistics*, 7(2): 223–238. 423

MacEachern, S. N. (1999). "Dependent nonparametric processes." In: *ASA Proceedings of the Section on Bayesian Statistical Science*, 50–55. 441

— (2000). "Dependent Dirichlet processes." Unpublished manuscript, Department of Statistics, The Ohio State University. 441

Neal, R. (2000). "Markov chain sampling methods for Dirichlet process mixture models." *Journal of Computational and Graphical Statistics*, 9(2): 249–265. MR1823804. doi: `http://dx.doi.org/10.2307/1390653`. 423

Newey, W. K. and Smith, R. J. (2004). "Higher order properties of GMM and generalized empirical likelihood estimators." *Econometrica*, 72(1): 219–255. MR2031017. doi: `http://dx.doi.org/10.1111/j.1468-0262.2004.00482.x`. 427, 432

Pitman, J. and Yor, M. (1997). "The two-parameter Poisson–Dirichlet distribution derived from a stable subordinator." *The Annals of Probability*, 25(2): 855–900. MR1434129. doi: `http://dx.doi.org/10.1214/aop/1024404422`. 430

Rényi, A. (1961). "On measures of entropy and information." In: *Proceedings of Fourth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, 547–561. MR0132570. 427

Sethuraman, J. (1994). "A constructive definition of Dirichlet priors." *Statistica Sinica*, 4: 639–650. MR1309433. 430

Taddy, M. and Kottas, A. (2010). "A Bayesian nonparametric approach to inference for quantile regression." *Journal of Business and Economic Statistics*, 28(3): 357–369. MR2723605. doi: `http://dx.doi.org/10.1198/jbes.2009.07331`. 423

White, H. (1982). "Maximum likelihood estimation of misspecified models." *Econometrica*, 50(1): 1–26. MR0640163. doi: `http://dx.doi.org/10.2307/1912526`. 426

Acknowledgments

I am grateful for useful comments to Sid Chib, Nick Kiefer, Joon Park, Yuichi Kitamura, Michael Stutzer, and the seminar participants at the office of the comptroller of the currency, the info-metrics and nonparametric inference workshop (U.C. Riverside), the NBER-NSF seminar on Bayesian inference in econometrics and statistics (Washington University at St. Louis), and Midwest Econometrics Group conference (Indiana University). I also thank the Associate Editor and referees for their helpful suggestions on a previous version of this paper.