Computational Statistics and Data Analysis 55 (2011) 1581–1595

Contents lists available at ScienceDirect

Computational Statistics and Data Analysis

journal homepage: www.elsevier.com/locate/csda

Usage of a pair of **S**-paths in Bayesian estimation of a unimodal density

Man-Wai Ho

Department of Statistics and Applied Probability, National University of Singapore, 6 Science Drive 2, S117546, Republic of Singapore

ARTICLE INFO

Article history:
Received 18 January 2010
Received in revised form 24 August 2010
Accepted 10 November 2010
Available online 21 November 2010

Keywords:
Species sampling mixture model
Partition
Rao–Blackwell improvement
Sequential importance sampling
Weighted Chinese restaurant process

ABSTRACT

This paper aims at illustrating the importance of using **S**-paths in Bayesian estimation of a unimodal density on the real line. A class of species sampling mixture models containing random densities that are unimodal and not necessarily symmetric is considered. A novel and explicit characterization of the posterior distribution expressible as a finite mixture over pairs of two dependent **S**-paths is derived, resulting in closed-form and tractable Bayes estimators for both the density and the mode as finite sums over the pairs. These results are statistically important as they are proved to be Rao–Blackwell improvements over existing results expressible in terms of partitions, and thus can be estimated with less variability. Extending an effective and newly-developed sequential importance sampling (SIS) scheme for sampling one **S**-path at a time, an SIS scheme is proposed to approximate the density estimates or any other posterior quantities of the model that are expressible in terms of two **S**-paths. Simulation results are reported to demonstrate practicality of our methodology and its effectiveness over an existing class of non-iterative algorithms that are based on sampling partitions. Indeed, the latter commonly-used algorithms, widely believed to be feasible, are shown to be ineffective and unreliable, implying that there exists hardly any practical non-iterative algorithm in this context. This prompts the essentiality of a practically useful algorithm for the problem.

1. Introduction

Statistical methods are usually developed by assuming that the underlying distribution of data is symmetric and unimodal at zero on the real line, such as a normal or a Student's t distribution. Such a restrictive assumption can easily fail when dealing with real-life data, for example, stock prices and rainfall, that are unimodally distributed but may not be symmetric about its mode possibly different from zero. There is a vast amount of literature on nonparametric estimation of unimodal densities and the mode including early works of Grenander (1956), Parzen (1962), Chernoff (1964), Venter (1967), Prakasa Rao (1969) and Wegman (1969, 1970a,b, 1971). For more recent works, see Dharmadhikari and Joag-Dev (1988), Bertin et al. (1997) and Genton (2004), and the references therein.

Tackling the problem from a nonparametric Bayes approach, we look at the model,

$$f(t|G,\theta)=\int\frac{1}{Y}\left[\,\mathbb{I}(0<t-\theta\le Y)-\mathbb{I}(Y\le t-\theta<0)\,\right]G(\mathrm{d}Y),\quad t\in\mathcal{R}=(-\infty,\infty),\tag{1}$$

where $\mathbb{I}(A)$ is the indicator of an event A, and G is an unknown probability measure over $\mathcal{R}$, in analogy to Ishwaran and James (2003), taken from the class of species sampling models (SSM) developed in Pitman (1995, 1996). The SSM includes a large number of random probability measures, for instance, the Dirichlet process (Ferguson, 1973), the Poisson–Dirichlet (PD) process (Pitman and Yor, 1997), the class of finite-dimensional Dirichlet priors discussed in detail in Ishwaran and Zarepour (2002a,b), and the homogeneous normalized random measures with independent increments discussed in

E-mail addresses: stahmw@nus.edu.sg, ismtwaiho@hotmail.com.

0167-9473/$ – see front matter
doi:10.1016/j.csda.2010.11.007

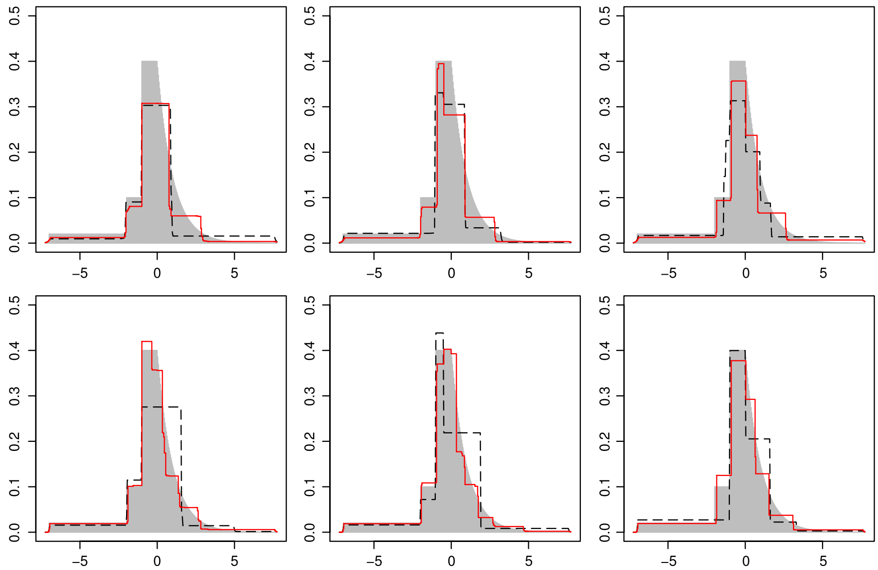

Fig. 1. Density estimates produced by the SIP(θ_0) sampler (solid line) and by GWCR (dashed line) with fixed modes $\theta = -1, -0.5, 0$ (from left to right) based on samples of size $N = 500$ (top) and 2000 (bottom) from true unimodal density $\lambda_1(t)$ (represented by the grey area) when G in (1) is $\mathcal{PD}(H; 0.9, 100)$ with H defined in (23).

Regazzini et al. (2003); see Pitman (1996) and Ishwaran and James (2003, Section 2) for more discussion. Validity of the above mixture representation can be justified by noting equality between its integral when $\theta = 0$ and the distribution function of any unimodal density with mode at zero given in Feller (1971, page 158); see also discussion in Brunner (1992). One can alternatively express models in (1) as a hierarchical model as discussed in Section 5.

Posterior analysis in model (1) through usage of random partitions follows from the general results in Ishwaran and James (2003). Posterior quantities such as Bayes estimate of the unimodal density expressible as a finite sum over partitions can be evaluated by the generalized weighted Chinese restaurant (GWCR) algorithm. The GWCR algorithm is in general an extension of many other sequential importance sampling (SIS; Kong et al., 1994) algorithms (see, for example Lo et al., 1996; Quintana, 1998; MacEachern et al., 1999) for fitting Dirichlet process mixture models (Lo, 1984) based on sampling random partitions. They are regarded as one of the most versatile non-iterative methods in fitting Bayesian nonparametric or semiparametric models; see, for example, Ishwaran and Takahara (2002) and Naskar and Das (2006). Generalizing the work of Basu and Chib (2003), all these algorithms can serve as essential numerical methods for computing marginal likelihoods in model selection problems involving semiparametric models. Nevertheless, unreliable estimates of a unimodal density produced by the GWCR algorithm cast doubt on appropriateness of usage partitions in this model; compare dashed lines with grey areas in Figs. 1–3.

An **S**-path or its equivalent, an integer-valued vector (defined in Section 2.1) that can be viewed as a combinatorial reduction of a partition, appears in many Bayesian mixture models under monotonicity constraints, such as monotone hazard rates by Dykstra and Laud (1981), Lo and Weng (1989) and Ho (2006a), monotone or unimodal symmetric densities by Brunner and Lo (1989) and Ho (2006b), and unimodal and rotationally symmetric density on a sphere by Brunner and Lo (1994). Ho (2006b) illustrates the importance and advantages of using **S**-paths in estimation of a class of symmetric unimodal densities, which is another special case of (1) resulting when G is symmetric about $\theta = 0$. For instance, posterior analysis in such a special model can be carried out via one random **S**-path. More importantly, the path-sum estimator of the density is shown to be a Rao–Blackwell (RB) improvement of the partition-sum counterpart available from the general results in Ishwaran and James (2003), and, in turn, to have a smaller variability. This is supported by empirical evidence through implementations of an iterative algorithm called the accelerated path (AP) sampler and the Gibbs version of the GWCR algorithm. Indeed, regarding all aforementioned models wherein **S**-path structures exist, analogous RB relationships between estimators in terms of respectively one **S**-path and a partition can be established, and thus it is always advantageous to use **S**-paths rather than partitions for inference (see Ho, 2006a, for discussion on a class of monotone hazard rates).

Contributions of this work start with a characterization of the posterior distribution of the models in (1) in terms of a pair of two statistically dependent **S**-paths (in Section 2). It yields closed-form Bayes estimators of both f and θ as finite sums over two **S**-paths. In principle, the pair of **S**-paths is also a combinatorial reduction of a partition. That is, the space of such pairs of **S**-paths is considerably coarser than that of partitions (see Appendix A.3). Hence, all these expressions via **S**-paths are much simpler for computational purposes though they look more mathematically involved than counterpart results in terms of partitions. An analogous RB improvement result between estimators of f of the two kinds is also established, that is, the path-sum estimator of f has a smaller variability and is always preferred to the partition counterpart. Section 3 proposes an SIS scheme for sampling one single **S**-path at a time, which is the first available effective non-iterative algorithm