Journal of Statistical Planning and Inference 166 (2015) 87–101

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Bayesian adaptation

Catia Scricciolo

Department of Decision Sciences, Bocconi University, Via Röntgen 1, 20136 Milano, Italy

ARTICLE INFO

Article history:
Received 24 February 2014
Received in revised form 22 September 2014
Accepted 3 December 2014
Available online 12 December 2014

Keywords:
Adaptive estimation
Empirical Bayes
Gaussian process priors
Kernel mixture priors
Nonparametric credibility regions
Posterior distributions
Rates of convergence
Sieve priors

ABSTRACT

In the need for low assumption inferential methods in infinite-dimensional settings, Bayesian adaptive estimation via a prior distribution that does not depend on the regularity of the function to be estimated nor on the sample size is valuable. We elucidate relationships among the main approaches followed to design priors for minimax-optimal rate-adaptive estimation meanwhile shedding light on the underlying ideas.

1. Introduction

Nonparametric curve estimation is a fundamental problem that has been intensively studied in a Bayesian framework only in the last decade, with more than a ten-years delay over the ponderous progress made in the frequentist literature where rates for point estimators have been developed in many aspects: adaptation, sharp minimax adaptive constants etc., see, e.g., Goldenshluger and Lepski (2012) for recent progress in the area. Bayesian adaptive estimation is a main theme: it accounts for designing a prior probability measure on a function space so that the posterior distribution contracts at "the truth" at optimal rate, in the minimax sense, relative to the distance defining the risk. The rate then has the desirable property of automatically adapting to the unknown regularity level of the estimandum: the correct rate stems, whichever the true value of the regularity parameter, even if knowledge of it is not available to be exploited in the definition of the prior. As the amount of data grows, the posterior distribution learns from the data so that the derived estimation procedure, despite lack of knowledge of the smoothness, performs as well as if the regularity level were known and this information could be incorporated into the prior. In this sense, adaptation may be regarded as an oracle property of the prior distribution providing a frequentist large-sample validation of it and, above all, a success of Bayesian nonparametric methods for low assumption inference in infinite-dimensional settings.

Early influential contributions to Bayesian adaptation are due to Belitser and Ghosal (2003) and Huang (2004). The former article deals with the prototypical problem of adaptive estimation of the mean of an infinite-dimensional normal distribution which is assumed to live in a Sobolev space of unknown smoothness level; the latter provides general sufficient conditions for adaptive density and regression estimation which are then applied to illustrate full exact minimax-optimal rate adaptation in density and regression estimation over Sobolev spaces using log-spline models and full minimax-optimal rate adaptation in

E-mail address: catia.scricciolo@unibocconi.it.

http://dx.doi.org/10.1016/j.jspi.2014.12.002
0378-3758/

density estimation over Besov spaces with the Haar basis but at the price of an extra logarithmic term. A third breakthrough contribution is given in the article of van der Vaart and van Zanten (2009), where adaptation is considered in the statistical settings of density estimation, regression and classification by introducing as a prior for the functional parameter a re-scaling of the sample paths of a smooth Gaussian random field on $[0,\ 1]^d$, $d \geq 1$, by an independent gamma random variable. These three articles are paradigmatic of the main approaches followed for Bayesian adaptation:

(a) the approach that considers the regularity level as a hyper-parameter and puts a prior on it;
(b) the approach that puts a prior on a discrete random variable which may represent the model dimension, the dimension of the space where the function is projected or the number of basis functions used in the approximation;
(c) the approach based on the re-scaling of a smooth Gaussian random field.

Approach (a), which considers hierarchical models with regularity hyper-parameter, is proposed in Belitser and Ghosal (2003), where the unknown regularity level is endowed with a prior supported on at most countably many values. The overall prior is then a mixture of priors on different models indexed by the regularity parameter and leads to exact optimal posterior contraction rates simultaneously for all regularity levels. The same philosophy is followed in Scricciolo (2006), where full exact optimal rate adaptive estimation of log-densities in Sobolev ellipsoids is achieved by considering only a finite number of competing models. In both articles, the key ideas are the following:

(i) the posterior probability of selecting a coarser model than the best one asymptotically vanishes;
(ii) the posterior distribution resulting from the prior restricted to smaller models asymptotically accumulates on a fixed ellipsoid in the correct space;
(iii) the posterior distribution corresponding to the restricted prior concentrates on Hellinger/ℓ^2-balls around the truth at optimal rate.

In both articles, full minimax-optimal rate adaptation is achieved when the prior on the regularity level can only take countably many values, while continuous spectrum adaptation is obtained at the price of a genuine power of n in Belitser and Ghosal (2003) and of an extra logarithmic factor in Lian (2014). In the latter article, adaptation to the regularity level of the Besov space where the true signal of a Gaussian white noise model is assumed to live is achieved, up to a log-factor, over the full scale of possible regularity values by considering a spike-and-slab type prior, with a point mass at zero mixed with a Gaussian distribution, on the single wavelet coefficients of the signal and a prior on a parameter related to the regularity of the space, but the overall prior is restricted to a fixed Besov ellipsoid. Another extension of Belitser and Ghosal (2003) to continuous spectrum is Knapik et al. (2012). Also the Bayesian adaptation scheme proposed by Ghosal et al. (2003) and Lember and van der Vaart (2007) can be ascribed to approach (a). It puts a prior on every model of a collection, each one expressing a qualitative prior guess on the true density, possibly a regularity parameter, and next combines these priors into an overall prior by equipping the abstract model indices with special sample-size-dependent prior weights giving more relevance to "smaller" models, that is, those with faster convergence rates. Illustrations include finite discrete priors based on nets and priors on finite-dimensional models for adaptive estimation over scales of Banach spaces like Hölder spaces. A closely related problem is that of model selection which is dealt with using similar ideas in Ghosal et al. (2008), where it is shown that the posterior distribution gives negligible weights to models that are bigger than the one that best approximates the true density from a given list, thus automatically selecting the optimal one.

Approach (b) that considers hierarchical models with dimension reduction hyper-parameter is followed in Huang (2004) and relies on the construction of a fairly simple compound prior called "sieve prior" by Shen and Wasserman (2001). A sieve prior is a mixture of priors,

$$\Pi = \sum_{k=1}^{\infty} \rho(k)\Pi_k,$$

with $\rho(k) \geq 0$, $\sum_{k=1}^{\infty} \rho(k) = 1$ and, where every single prior Π_k is supported on a space of densities $\mathscr{F}_k$ which is typically finite-dimensional and can be represented as $\{f_\theta : \theta \in \Theta_k\}$. As previously mentioned, the index k may represent the dimension of the space where the function is projected, the number of basis functions for the approximation or the model dimension. A sieve prior can be thought of as generated in two steps: first the index k of a model is selected with probability $\rho(k)$, next a probability measure is generated from the chosen model $\mathscr{F}_k$ according to a prior Π_k on it. Such finite-dimensional models may arise from the approximation of a collection of target densities through a set of basis functions (e.g., trigonometric functions, splines or wavelets), where a model of dimension k is generated by a selection of k basis functions. This adaptive scheme is based on a set of assumptions such that they give control in terms of covering numbers of the local structure of each Θ_k, they guarantee the existence of a model $\mathscr{F}_{k_n}$ receiving enough prior weight $\rho(k_n)$, the existence of a density $f_{\beta_{k_n}} \in \mathscr{F}_{k_n}$ close to f_0 and of neighborhoods of this approximating density being charged enough prior mass by Π_{k_n}. Several examples treated in Huang (2004) using scales of finite-dimensional models are covered with different priors in Lember and van der Vaart (2007). Further references on adaptive curve estimation via sieve priors are Scricciolo (2008) and Arbel et al. (2013). Bayesian adaptive procedures via sieve priors on the unit interval include piecewise constant and polygonally smoothed priors based on the Dirichlet process as in Scricciolo (2007), Bernstein–Dirichlet polynomials as in Kruijer and van der Vaart (2008), mixtures of beta densities as in Rousseau (2010). Other contributions clearly belonging to this category, while not dealing with Dirichlet mixtures, are de Jonge and van Zanten (2010, 2012), Ray (2013) and Belitser and