Journal of Statistical Planning and Inference 145 (2014) 92–101

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

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

Adaptive rates of contraction of posterior distributions in Bayesian wavelet regression

Heng Lian *

Division of Mathematical Sciences, School of Physical and Mathematical Sciences, Nanyang Technological University, Singapore 637371, Singapore

A R T I C L E I N F O

Article history:
Received 10 May 2013
Received in revised form
5 September 2013
Accepted 5 September 2013
Available online 17 September 2013

Keywords:
Besov spaces
Infinite-dimensional Bayesian procedure
Wavelet thresholding

A B S T R A C T

In the last decade, many authors studied asymptotic optimality of Bayesian wavelet estimators such as the posterior median and the posterior mean. In this paper, we consider contraction rates of the posterior distribution in Bayesian wavelet regression in L_2/l_2 neighborhood of the true parameter, which lies in some Besov space. Using the common spike-and-slab-type of prior with a point mass at zero mixed with a Gaussian distribution, we show that near-optimal rates (that is optimal up to extra logarithmic terms) can be obtained. However, to achieve this, we require that the ratio between the log-variance of the Gaussian prior component and the resolution level is not constant over different resolution levels. Furthermore, we show that by putting a hyperprior on this ratio, the approach is adaptive in that knowledge of the value of the smoothness parameter is no longer necessary. We also discuss possible extensions to other priors such as the sieve prior.

1. Introduction

In this paper, we study Bayesian estimation of a function f with observations from the white noise model

$$dX(t) = f(t)dt + dW(t)/\sqrt{n}. \tag{1}$$

We assume that $f \in B^s_{p,q}$ is in the Besov space and W is the standard Brownian motion. Due to the acclaimed approximation property of wavelets for functions in Besov spaces, wavelets expansion is typically used for estimation. This white noise model (1) is closely related to the nonparametric regression model (Brown and Low, 1996; Donoho et al., 1995):

$$Y_i = f\left(\frac{i}{n}\right) + z_i \tag{2}$$

with standard normal noise. We choose to work with formulation (1) for its simplicity. Johnstone and Silverman (2005) and Pensky (2006) have established explicit connections between models (1) and (2).

The Bayesian approach for estimating f consists of putting a prior on f and computing its posterior distribution, which is usually performed via Markov chain Monte Carlo (MCMC) methods except in the special case where the prior is conjugate. Motivated by the superior performance of the Bayesian approach in practice, several theoretical studies (Abramovich et al., 2004; Johnstone and Silverman, 2005; Pensky, 2006) have investigated the asymptotic frequentist optimality properties of

* Tel.: +65 65137175.
E-mail address: henglian@ntu.edu.sg

0378-3758/$ - see front matter
http://dx.doi.org/10.1016/j.jspi.2013.09.002

different Bayesian estimators, with posterior mean and posterior median being the most commonly studied ones. These works give theoretical justifications for the use of Bayesian estimators in practice. Abramovich et al. (2007) and Abramovich and Grinshtein (2010) considered convergence rates of an estimator called "Bayesian testimation" that does not seem fully Bayesian. Castillo and van der Vaart (2012) considered estimation of a sparse sequence in the l_s ball. Their objective is discovery of sparsity patterns which we do not consider in this paper. For this reason, they consider the class of hierarchical priors that more explicitly exploits sparsity properties and, hence, is very different from ours.

Our interest lies in studying the frequentist properties of the posterior distribution, for which a general theory has been presented in Ghosal et al. (2000), Ghosal and Van Der Vaart (2007). In some sense the characterization of the posterior distribution is more general than Bayesian estimators, since the latter only focuses on limited aspects of the posterior distribution. Like frequentists, we assume that the data X has been generated according to model (1) with a given true parameter f_0. We investigate at what rate the posterior distribution contracts to f_0 as $n\to\infty$. That is, we investigate how small a "neighborhood" A of f_0 should be so that it still allows $\Pi^n(A|X)\to 1$ in probability. In this paper we will use Π^n for the posterior distribution and Π_n for the prior distribution.

In this type of analysis, prior concentration rate characterized by the Kullback–Leibler distance is the most important deciding factor of the convergence rates when considering choice of priors. This means that the prior distribution should have support sufficiently large in the neighborhood of the true parameter to obtain the correct contraction rates. Demonstration of this large support property composes the main part of our proofs.

Another important theoretical contribution is that by imposing a hyperprior on a parameter related to the smoothness parameter of the Besov space, optimal rate of convergence can be obtained without knowing the value of the smoothness parameter. Without such a hyperprior, even when the smoothness parameter is known, our strategy of proof does not produce an optimal rate of convergence.

In conclusion, we propose that one should put a hyperprior on the ratio of the log-variance of the Gaussian prior component and the resolution level.

2. Main results

2.1. Background

For simplicity we assume that f is periodic and use periodic orthonormal wavelet basis on [0, 1]. Using wavelet basis with sufficient regularity, the function f can be expanded as

$$f=\sum_{k=0}^{2^{j_0}-1}\alpha_{j_0k}\phi_{j_0k}+\sum_{j\ge j_0}\sum_{k=0}^{2^j-1}\beta_{jk}\psi_{jk},$$

where ϕ_{j_0k} are the scaling functions and ψ_{jk} are the mother wavelets at resolution j, and j_0 is the coarsest resolution in the expansion. We assume $j_0=0$ for simplicity of notations below.

The Besov spaces include the well-known Sobolev and Hölder classes of function and also nearly contain the space of functions of bounded variation. The norm for the Besov space with parameter $s>\max(0,1/p-1/2), 1\le p\le\infty$, and $1\le q\le\infty$ is defined as

$$\|f\|_{B^s_{p,q}}=\|P_0(f)\|_{L^p}+\left(\sum_{j\ge 0}(2^{js}\|Q_j(f)\|_{L^p})^q\right)^{1/q},$$

where $P_0(f)=\alpha_{00}\phi_{00}$ is the projection of f on the "approximation space", and $Q_j(f)=\sum_{k=0}^{2^j-1}\beta_{jk}\psi_{jk}$ is the projection of f onto the "detail space".

In terms of the coefficients in the wavelet expansion, the Besov norm can be equivalently defined by

$$\|f\|_{B^s_{p,q}}\asymp\|\beta\|_{B^s_{p,q}}=|\alpha_{00}|+\left\{\sum_{j=0}^{\infty}2^{j(s+1/2-1/p)q}\|\beta_{j.}\|_p^q\right\}^{1/q}.$$

Note that for cases where $q=\infty$ the usual change to the sup norm is required. We also define $P_J\beta$ to be the sequence β' such that $\beta'_{jk}=\beta_{jk}$ when $j\le J$ and $\beta'_{jk}=0$ when $j>J$.

After wavelet transformation for (1), we get the Gaussian sequence model:

$$X_{00}=\alpha^0_{00}+z_{00}/\sqrt{n}$$
$$X_{jk}=\beta^0_{jk}+z_{jk}/\sqrt{n},\quad j\ge 0,\ \ k=0,1,\ldots,2^j-1,$$

where the superscript 0 indicates the true parameter.

Using a Bayesian approach for Gaussian sequence estimation, we put a prior on each β_{jk} independently:

$$\beta_{jk}\sim\pi_jN(0,a_j^2)+(1-\pi_j)\delta_0, \tag{3}$$

with parameters $a_j^2\asymp 2^{-2\alpha j},\pi_j\asymp 2^{-\gamma j}$, for some $\alpha\ge 0$, and $\gamma\ge 0$. Here α is the ratio of the log-inverse-variance over the resolution level, which is the key tuning parameter in the prior for optimal convergence rates. γ also has some effects but