Asymptotic Behavior of Bayesian Nonparametric Procedures

Yang Xing

Faculty of Forest Sciences

Centre of Biostochastics, Department of Forest Economics

Umeå

Doctoral Thesis

Swedish University of Agricultural Sciences

Umeå 2009

Acta Universitatis agriculturae Sueciae
2009:35

ISSN 1652-6880
ISBN 978-91-86195-82-3

Print: Arkitektkopia 2009

Asymptotic Behavior of Bayesian Nonparametric Procedures

Abstract

Asymptotics plays a crucial role in statistics. The theory of asymptotic consistency of Bayesian nonparametric procedures has been developed by many authors, including Schwartz (1965), Barron, Schervish and Wasserman (1999), Ghosal, Ghosh and Ramamoorthi (1999), Ghosal, Ghosh and van der Vaart (2000), Shen and Wasserman (2001), Walker and Hjort (2001), Walker (2004), Ghosal and van der Vaart (2007) and Walker, Lijoi and Prunster (2007). This theory is mainly based on existence of uniformly exponentially consistent tests, computation of a metric entropy and measure of a prior concentration around the true value of parameter. However, both the test condition and the metric entropy condition depend on models but not on prior distributions. Because a posterior distribution depends on the complexity of the model only through its prior distribution, it is therefore natural to explore appropriate conditions which incorporate prior distributions. In this thesis we introduce Hausdorff α-entropy and an integration condition, both of which incorporate prior distributions and moreover are weaker than the metric entropy condition and the test condition, respectively. Furthermore, we provide an improved method to measure the prior concentration. By means of these new quantities, we derive several types of general posterior consistency theorems and general posterior convergence rate theorems for i.i.d. and non-i.i.d. models, which lead to improvements in a number of currently known theorems and their applications. We also study rate adaptation for density estimation within the Bayesian framework and particularly obtain that the Bayesian procedure with hierarchical prior distributions for log spline densities and a finite number of models achieves the optimal minimax rate when the true density is Hölder-continuous. This result disconfirms a conjecture given by Ghosal, Lember and van der Vaart (2003). Finally, we find a new both necessary and sufficient condition on Bayesian exponential consistency for prior distributions with the Kullback-Leibler support property.

Keywords: Adaptation, consistency, density function, Hausdorff entropy, Hellinger metric, Kullback-Leibler divergence, log spline density, Markov chain, nonparametrics, posterior distribution, rate of convergence, sieve.

Author's address: Yang Xing, Department of Forest Economics, SLU, S-901 83 Umeå, Sweden.
E-mail : yang.xing@sekon.slu.se

List of Papers

This thesis is based on the work contained in the following papers, referred to by Roman numerals in the text:

I. Xing, Y. (2008). Convergence rates of posterior distributions for observations without the iid structure, available at www.arxiv.org: 0811.4677, Submitted.

II. Xing, Y. (2008). Convergence rates of nonparametric posterior distributions, available at www.arxiv.org: 0804.2733, Submitted.

III. Xing, Y. On Adaptive Bayesian Inference, *Electronic J. Statist.* Vol. 2 (2008), 848-862.

IV. Xing, Y and Ranneby, B. Sufficient conditions for Bayesian consistency, *J. Statist. Plann. Inference.* **139** (2009), 2479-2489.

V. Xing, Y and Ranneby, B. (2008). Both necessary and sufficient conditions for Bayesian exponential consistency, available at www.arxiv.org: 0812.1084, Submitted.

VI. Xing, Y and Ranneby, B. (2008). On Bayesian consistency, Submitted.

Papers III and IV are reproduced with the permission of the publishers.

Summary of Papers

Asymptotics is known as one of the most important properties of statistical procedures. For a statistical procedure, lack of consistency might lead to seriously incorrect inferences, and this is true in a Bayesian setting as well. The asymptotic behavior of Bayesian nonparametric procedures has been the focus of a considerable amount of research during past three decades, see for instance Schwartz [11], Barron, Schervish and Wasserman [1], Ghosal, Ghosh and Ramamoorthi [5], Ghosal, Ghosh and van der Vaart [6], Ghosal and van der Vaart [10], Shen and Wasserman [12], Walker [13], Walker and Hjort [14] and Walker, Lijoi and Prunster [15]. Let $\big(\mathfrak{X}^{(n)}, \mathcal{A}^{(n)}, P_\theta^{(n)} : \theta \in \Theta\big)$ for $n = 1, 2, \ldots$ be a sequence of statistical experiments with observations $X^{(n)}$, where the parameter set Θ is assumed to be independent of the index n. Suppose that the distribution $P_\theta^{(n)}$ for each $\theta \in \Theta$ admits a density $p_\theta^{(n)}$ relative to some σ-finite measure $\mu^{(n)}$ on $\mathfrak{X}^{(n)}$. Denote by θ_0 the true value of parameter of the distributions which generate the observations $X^{(n)}$ for all n. For each statistical experiment $\big(\mathfrak{X}^{(n)}, \mathcal{A}^{(n)}, P_\theta^{(n)} : \theta \in \Theta\big)$, we let Π_n stand for a prior distribution on Θ and let $\Pi_n(\,\cdot \,|\, X^{(n)})$ stand for the corresponding posterior distribution given by the Bayes theorem

$$\Pi_n\big(A \,\big|\, X^{(n)}\big) = \frac{\int_A p_\theta^{(n)}(X^{(n)})\, \Pi_n(d\theta)}{\int_\Theta p_\theta^{(n)}(X^{(n)})\, \Pi_n(d\theta)}$$

for each measurable subset A in Θ. Observe that one can of course take a common prior Π on Θ for all n, but to increase the scope of applicability, we assume here that the prior Π_n may depend on n. In the case that $X^{(n)}$ is a random vector $(X_1, X_2, \ldots, X_n)$ with independent variables X_i, where n is the sample size and each X_i is generated from a density $p_{\theta,i}$ relative to some σ-finite measure μ_i on $(\mathfrak{X}_i, \mathcal{A}_i)$, then $P_\theta^{(n)}$ is usually chosen to be the joint distribution with the product density $p_\theta^{(n)}(X^{(n)}) = \prod_{i=1}^n p_{\theta,i}(x_i)$ relative to the direct product measure $\mu^{(n)} = \mu_1 \times \mu_2 \times \cdots \times \mu_n$ on $\mathfrak{X}^{(n)} = \mathfrak{X}_1 \times \mathfrak{X}_2 \times \cdots \times \mathfrak{X}_n$. The posterior distribution $\Pi_n(\,\cdot \,|\, X^{(n)})$ is said to be consistent almost surely (respectively, in probability) at the true parameter θ_0 if for any open neighborhood A of θ_0, $\Pi_n(\,A \,|\, X^{(n)})$ tends to one as $n \to \infty$ almost surely (respectively, in probability) under the distribution governed by θ_0. For a consistent posterior distribution, the rate of convergence is measured by the size of the smallest shrinking balls around the true parameter on which the posterior masses tend to one as the index n increases to infinity. An early work on consistency of posterior distributions is due to Doob [3], who proved under a very mild condition that for any given prior distribution, the posterior distribution is consistent at each true parameter except possibly on a set with zero prior mass. Since the exceptional set for the true parameter can be dense everywhere on the parameter space, Doob's result actually cannot produce any information on posterior consistency at a given parameter. In fact, Freedman [4] and Diaconis and Freedman [2] have demonstrated that a prior distribution having positive mass on all weak neighborhoods of the true parameter cannot imply the weak consistency of the posterior distribution, in which the open neighborhood A of the true parameter is induced by weak topology. A well known sufficient condition on consistency of posterior distributions was found by Schwartz [11]. Her result implies that if the true parameter is in the Kullback-Leibler support of a prior distribution in the setup of i.i.d. observations then the posterior distribution accumulates in any given weak neighborhood of the true parameter when the sample size is large enough. However, in many applications like density estimation it is natural to ask for the almost sure consistency of Bayesian procedures with respect to some important metrics such as the Hellinger metric and the L_p-norm. Unfortunately, it is known that the condition on the Kullback-Leibler support of the true parameter relative to prior distribution is not enough to guarantee the almost sure consistency of the posterior distribution with respect to general topologies. Some additional restrictions must be needed to obtain a positive result. Many people, including Barron, Schervish and Wasserman [1], Ghosal, Ghosh and Ramamoorthi [5], Ghosal and van der Vaart [10], Schwartz [11], Shen and Wasserman [12] and Walker [13], have made great contributions in this direction. The main currently known approach in the study of consistency and convergence rates of Bayesian nonparametric procedures consists of (1) computing a metric entropy which in fact puts appropriate size restrictions on the model, (2) checking the existence of uniformly exponentially consistent tests for testing the true parameter, against the intersection of some suitable sieve and any small neighborhood of each another parameter, and (3) measuring one type of prior concentration which depends on the prior mass assigned to suitable neighborhoods of the true parameter. Both the metric entropy condition and the test condition depend on models but not on prior distributions. Since a posterior distribution depends on the complexity of the model only through its prior distribution, it is therefore natural to explore conditions which incorporate prior distributions.

This thesis aims at developing a better approach to handle asymptotic behavior of Bayesian nonparametric procedures. Our results imply that both the metric entropy condition and the test condition can be improved by means of our weaker and prior-incorporating conditions. We also provide an improved method to measure prior concentration around the true parameter. These lead to improvements in a number of general posterior consistency theorems and general posterior convergence rate theorems for i.i.d. and non-i.i.d. models and their applications. Our main tools are the Hausdorff α-entropy, an integration condition and an improved method to measure the prior concentration, which are introduced and studied in Papers IV, I and II respectively. Throughout this thesis we shall adopt the convention that $\log 0 = -\infty$ and $0/0 = 0$. We introduce

The Hausdorff α-Entropy (Paper IV). *Let $\alpha \geq 0$ and $\Theta_n \subset \Theta$. For $\delta > 0$, the Hausdorff α-entropy $J(\delta, \Theta_n, \alpha, \Pi_n, d_n)$ of Θ_n relative to a given prior Π_n and a semimetric d_n on Θ is defined as*

$$J(\delta, \Theta_n, \alpha, \Pi_n, d_n) = \log \inf \sum_{j=1}^{N} \Pi(B_j)^\alpha,$$

where the infimum is taken over all coverings $\{B_1, B_2, \ldots, B_N\}$ of Θ_n and N may take ∞ such that each B_j is contained in some d_n-ball

$$\{\theta \in \Theta : d_n(\theta, \theta_j) < \delta\}$$

of radius δ and center at θ_j.

The Hausdorff α-entropy $J(\delta, \Theta_n, \alpha, \Pi_n, d_n)$ is a decreasing function of α in $[0, \infty)$. We have that $J(\delta, \Theta_n, 1, \Pi_n, d_n) = \Pi_n(\Theta_n)$ and $J(\delta, \Theta_n, 0, \Pi_n, d_n)$ is precisely equal to the metric entropy with respect to d_n. The metric entropy does not incorporate the prior Π_n and has been widely used by many authors in their study of asymptotics of Bayesian nonparametric procedures. Clearly, the Hausdorff α-entropy with $0 < \alpha < 1$ is much smaller than the metric entropy. For convenience we call the constant $C(\delta, \Theta_n, \alpha, \Pi_n, d_n) := e^{J(\delta, \Theta_n, \alpha, \Pi_n, d_n)}$ for the Hausdorff α-constant of the set Θ_n. Given $0 \leq \alpha \leq 1$ and $\Theta_n \subset \Theta$, we obtain

$$C(\delta, \Theta_n, \alpha, \Pi_n, d_n) \leq \Pi_n(\Theta_n)^\alpha \, N(\delta, \Theta_n, d_n)^{1-\alpha},$$

where $N(\delta, \Theta_n, d_n)$ stands for the minimal number of balls of d_n-radius δ needed to cover Θ_n and is usually called for the covering number of Θ_n. Our results show that one can replace the metric entropy by the Hausdorff α-entropy (or equivalently, the covering number is replaced by the Hausdorff α-constant) to study Bayesian asymptotics.

Denote by $R_\theta^{(n)}(X^{(n)}) = p_\theta^{(n)}(X^{(n)})/p_{\theta_0}^{(n)}(X^{(n)})$ the likelihood ratio of observations $X^{(n)}$ from the statistical experiment $(\mathfrak{X}^{(n)}, \mathcal{A}^{(n)}, P_\theta^{(n)} : \theta \in \Theta)$. For i.i.d. observations $X^{(n)} = (X_1, X_2, \ldots, X_n)$ generating from the density f_{θ_0}, one has a simple expression $R_\theta^{(n)}(X^{(n)}) = \prod_{i=1}^n \big\{ f_\theta(X_i)/f_{\theta_0}(X_i) \big\}$. Now we define

An Integration Condition (Paper I). *Let $\{d_n\}$ and $\{e_n\}$ be two sequences of semimetrics on Θ. For some $\alpha \in (0,1)$ there exist constants $K_1 > 0$, $K_2 > 0$ and $K_3 \geq 0$ such that the inequality*

$$P_{\theta_0}^{(n)} \Big(\int_{\theta \in \Theta_1 :\, d_n(\theta, \theta_0) > \varepsilon} R_\theta^{(n)}(X^{(n)})\, \Pi_n(d\theta) \Big)^\alpha$$

$$\leq K_1\, e^{-K_2 n \varepsilon^2} C(\varepsilon, \{\theta \in \Theta_1 : \, d_n(\theta, \theta_0) > \varepsilon\}, \alpha, \Pi_n, e_n)^{K_3}$$

holds for any $\varepsilon > 0$, $\Theta_1 \subset \Theta$ and for all n large enough.

The integration condition is weaker than the classical condition on existence of uniformly exponentially consistent tests, which states that for every n, $\varepsilon > 0$ and $\theta_1 \in \Theta$ with $d_n(\theta_1, \theta_0) > \varepsilon$ there exists a test ϕ_n such that

$$P_{\theta_0}^{(n)} \phi_n \leq e^{-n\varepsilon^2} \quad \text{and} \quad \inf_{\theta \in \Theta :\, e_n(\theta, \theta_1) < \varepsilon} P_\theta^{(n)} \phi_n \geq 1 - e^{-n\varepsilon^2}.$$

Another advantage of this integration condition is that it holds automatically for certain interesting classes of metrics and can moreover be helpful in constructing priors which lead to good properties of the posteriors. We found that the classical test condition can be replaced by the integration condition in study of Bayesian asymptotics.

Denote by $||g||_p = \big(\int |g|^p\big)^{1/p}$ the standard L_p-norm of g. The Hellinger distance $H(f_0, f) = ||\sqrt{f_0} - \sqrt{f}||_2$ times the supremum norm of the ratio f_0/f was suggested by Ghosal, Ghosh and van der Vaart [6] and adopted by many people to measure the prior concentration in order to ensure the almost sure convergence of posterior distributions. In Paper II we instead use the following modification of the Hellinger distance.

An Improved Prior Concentration Rate (Paper II). *A sequence $\{\varepsilon_n\}_1^\infty$ of numbers decreasing to zero is called for a prior concentration rate for the density space $\mathbb{F}$ if there exists a constant $c > 0$ such that the prior mass of the subset $\{f \in \mathbb{F} : H_*(f_0, f) \leq \varepsilon_n\}$ exceeds $e^{-cn\varepsilon_n^2}$ for all n, where $H_*(f_0, f)$ stands for $\big|\big|(\sqrt{f_0} - \sqrt{f})\big(\frac{2}{3}\sqrt{\frac{f_0}{f}} + \frac{1}{3}\big)^{1/2}\big|\big|_2$.*

Since the inequalities $H_*(f,g) \leq \big|\big|f/g\big|\big|_\infty^{1/4} H(f,g) \leq \big|\big|f/g\big|\big|_\infty H(f,g)$ hold for all density functions f and g, it would be better to adopt the quantity H_* instead of that suggested by Ghosal, Ghosh and van der Vaart [6]. Particularly, the supremum norm of such a ratio may become infinite in some cases, whereas the quantity H_* works very well.

The first three papers of this thesis deal with rates of convergence of posterior distributions, and the last three ones deal with consistency of posterior distributions in the setup of i.i.d. observations. Let me now briefly summarize the main results of these papers.

Paper I

In Paper I we supply the prior-dependent integration condition to establish general posterior convergence rate theorems for observations which may not be independent and identically distributed. The classical condition on the existence of uniformly exponentially consistent tests for testing the true parameter against each small neighborhood of other parameters has been widely adopted in the study of asymptotics of Bayesian nonparametric procedures. Note that a posterior depends on the complexity of the model only through its prior. As far as the Bayesian approach is concerned, it would be natural to explore alternative and appropriate conditions which incorporate priors. In this paper we use an integration condition together with the Hausdorff α-entropy, introduced in Paper IV, to study convergence rates of posteriors. The integration condition and the Hausdorff α-entropy both incorporate priors. Moreover, the integration condition is weaker than the existence of uniformly exponentially consistent tests and holds automatically for certain classes of metrics used to describe rates of convergence. The integration condition is also useful in construction of priors which yield the optimal rates of convergence. These lead to improvements of several existing theorems. For instance, we establish the following posterior convergence rate theorem.

Theorem 1. *Suppose that the integration condition holds and that* $\varepsilon_n > 0$, $n\,\varepsilon_n^2 \geq c_0 \log n$ *for all large* n *and some fixed constant* $c_0 > 0$. *Suppose that there exist a constant* $c_1 < K_2$ *and a sequence of subsets* Θ_n *on* Θ *such that*

$$C(j\varepsilon_n, \{\theta \in \Theta_n : j\varepsilon_n < d_n(\theta,\theta_0) \leq 2j\varepsilon_n\}, \alpha, \Pi_n, e_n)^{K_3} \leq e^{c_1 j^2 n\varepsilon_n^2}\, \Pi_n\big(W_n(\theta_0,\varepsilon_n)\big)^\alpha$$

for all sufficiently large integers j *and* n, *where*

$$W_n(\theta_0,\varepsilon_n) = \big\{\theta \in \Theta :\, H_*(p_{\theta_0}^{(n)}, p_\theta^{(n)}) \leq \sqrt{n}\varepsilon_n\big\}.$$

Then for each r *large enough we have that*

$$\Pi_n\big(\theta \in \Theta_n : \, d_n(\theta, \theta_0) \geq r\, \varepsilon_n | X^{(n)}\big) \longrightarrow 0$$

almost surely as $n \to \infty$. If furthermore there exists $c_2 > \frac{1}{c_0}$ such that

$$\sum_{n=1}^{\infty} \frac{e^{n\, \varepsilon_n^2\,(3+2c_2)}\, \Pi_n(\Theta \setminus \Theta_n)}{\Pi_n\big(W_n(\theta_0, \varepsilon_n)\big)} < \infty,$$

then there exists a constant $b > 0$ such that for each large r and all large n,

$$\Pi_n\big(\theta \in \Theta : \, d_n(\theta, \theta_0) \geq r\, \varepsilon_n | X^{(n)}\big) \leq e^{-bn\varepsilon_n^2} \qquad \textit{almost surely}$$

which tends to zero as $n \to \infty$.

Under an appropriately weaker condition we obtain an analogue of Theorem 1 for the in-probability posterior convergence. We also establish a posterior convergence rate theorem for general Markov processes, which is an extension of a theorem for stationary α-mixing Markov chains given by Ghosal and van der Vaart [10]. As an application we improve on the posterior rate of convergence for a nonlinear autoregressive model given by Ghosal and van der Vaart [10]. Many authors have studied Bayesian convergence rates for the Gaussian white noise model and constructed some interesting priors to get the optimal convergence rates of posteriors. Now our theorems is applied to extend their results to multi-normally distributed observations which may not be independent.

Paper II

Paper II deals with convergence rates of posteriors for i.i.d. data. Our main tools are an improved method to measure prior concentration around the true density f_0 in the density space $\mathbb{F}$ and the Hausdorff α-entropy given in Paper IV. We present several types of general posterior convergence rate theorems. One of the main results is the following theorem.

Theorem 2. *Let $\{\bar\varepsilon_n\}_{n=1}^{\infty}$ and $\{\tilde\varepsilon_n\}_{n=1}^{\infty}$ be two positive sequences such that $n\,\min(\bar\varepsilon_n^2, \tilde\varepsilon_n^2) \to \infty$ as $n \to \infty$. Suppose that there exist constants $c_1 > 0$, $c_2 > 0$, $c_3 \geq 0$, $0 \leq \alpha < 1$ and a sequence $\{\mathcal{G}_n\}_{n=1}^{\infty}$ of subsets on $\mathbb{F}$ such that $\sum\limits_{n=1}^{\infty} e^{-n\,\tilde\varepsilon_n^2\, c_2} < \infty$ and*

(1) $\sum\limits_{n=1}^{\infty} C(\bar\varepsilon_n, \mathcal{G}_n, \alpha, \Pi, H)\, e^{-n\,\bar\varepsilon_n^2\, c_1} < \infty,$

(2) $\sum\limits_{n=1}^{\infty} e^{n\,\tilde\varepsilon_n^2\,(3+3c_2+c_3)}\, \Pi(\mathbb{F} \setminus \mathcal{G}_n) < \infty,$

(3) $\Pi\big(f \in \mathbb{F} : H_*(f_0, f) \le \tilde{\varepsilon}_n\ \big) \ge e^{-n\,\tilde{\varepsilon}_n^2\, c_3}$.

Then for $\varepsilon_n = \max(\bar{\varepsilon}_n, \tilde{\varepsilon}_n)$ and each $r > 2 + \sqrt{\frac{2(3\alpha+2\alpha c_2+\alpha c_3+c_1)}{1-\alpha}}$, we have

$$\Pi\big(f \in \mathbb{F} :\ H(f_0, f) \ge r\varepsilon_n \,\big|\, X_1, X_2, \dots, X_n\big) \longrightarrow 0$$

almost surely as $n \to \infty$.

Theorem 2 strengthens several theorems on posterior convergence rates in Ghosal, Ghosh and van der Vaart [6], Shen and Wasserman [12] and Walker, Lijor and Prunster [15]. We apply our results to some statistical models including Bernstein polynomial priors, priors based on uniform distribution, log spline models and finite-dimensional models. We obtain some improvements on known results for these models.

Paper III

Paper III deals with rate adaptation for density estimation within the Bayesian framework. Given a collection of models, from the Bayesian point of view it is natural to put a prior on model index and let the resulting posteriors determine a good single model. A rate-adaptive posterior achieves the rate of convergence provided by the best single model from the collection. We study convergence rates of Bayesian procedures for hierarchical priors, consisting of prior weights on a model index set and priors on each individual density model. More detailedly, for each positive integer n, denote by I_n a countable index set. Let $\Pi_{n,\gamma}$ be a probability measure on a subset $\mathcal{P}_{n,\gamma}$ of the density space for each $\gamma \in I_n$ and let $\{\lambda_{n,\gamma} : \gamma \in I_n\}$ be a discrete probability measure on I_n. We get an overall prior Π_n defined by

$$\Pi_n = \sum_{\gamma \in I_n} \lambda_{n,\gamma}\, \Pi_{n,\gamma}.$$

Assume that for a given true density f_0 there exists a best model $\mathcal{P}_{n,\beta_n}$ equipped with a prior Π_{n,β_n} such that the optimal posterior rate is ε_{n,β_n}. An interesting problem is to find conditions ensuring that the posterior distribution $\Pi_n\big(\cdot \,\big|\, X_1, X_2, \dots, X_n\big)$ achieves the same rate of convergence as we only use the best single model Π_{n,β_n} for this f_0. Ghosal, Lember and Van der Vaart [7][8] have proved in-probability results. When applying to log spline densities with a finite number of models, their result leads to adaptation up to a logarithmic factor and it was shown in [8] that the additional logarithmic factor in the convergence rate can be removed by choosing special prior weights $\lambda_{n,\gamma}$. Now we establish an almost sure assertion. Denote

$$W_{n,\gamma}(\varepsilon) = \big\{f \in \mathcal{P}_{n,\gamma} :\ H_*(f_0, f) \le \varepsilon\big\},$$

$$\begin{aligned} A_{n,\gamma}(\varepsilon) &= \big\{f \in \mathcal{P}_{n,\gamma} : d(f_0, f) \le \varepsilon\big\},\\ I_n^1 &= \{\gamma \in I_n : \varepsilon_{n,\gamma} \le \sqrt{H}\varepsilon_{n,\beta_n}\},\\ I_n^2 &= \{\gamma \in I_n : \varepsilon_{n,\gamma} > \sqrt{H}\varepsilon_{n,\beta_n}\}. \end{aligned}$$

Then we have

Theorem 3. *Suppose that there exist positive constants* $H \ge 1$, $K \ge 1$, E_γ, $\mu_{n,\gamma}$, B, G, J, L, C *and* $0 < \alpha < 1$ *such that* $1 - \alpha > 18\alpha(3B^2 + 2CB^2 + L)$, $n\varepsilon_{n,\beta_n}^2 \ge (1 + \frac{1}{C})\log n$, $\sup_{\gamma \in I_n^1} E_\gamma \varepsilon_{n,\gamma}^2 \le G\varepsilon_{n,\beta_n}^2$, $\sup_{\gamma \in I_n^2} E_\gamma \le G$ *and* $\sum_{\gamma \in I_n} \mu_{n,\gamma}^\alpha = \mathrm{O}(e^{Jn\varepsilon_{n,\beta_n}^2})$. *Let* r *be a constant with* $r \ge \frac{18(C+J+G)}{1-\alpha-18\alpha L-54\alpha B^2-36\alpha CB^2}$ $+\sqrt{H} + \frac{1}{B} + 1$ *such that*

$$(1) \quad N\big(\tfrac{\varepsilon}{3}, A_{n,\gamma}(2\varepsilon), d\big) \le e^{E_\gamma n\varepsilon_{n,\gamma}^2} \quad \textit{for all} \quad \gamma \in I_n \quad \textit{and} \quad \varepsilon \ge \varepsilon_{n,\gamma},$$

$$(2) \quad \frac{\lambda_{n,\gamma}\, \Pi_{n,\gamma}\big(A_{n,\gamma}(j\varepsilon_{n,\gamma})\big)}{\lambda_{n,\beta_n}\, \Pi_{n,\beta_n}\big(W_{n,\beta_n}(\varepsilon_{n,\beta_n})\big)} \le \mu_{n,\gamma}\, e^{Lj^2 n\varepsilon_{n,\gamma}^2} \quad \textit{for all } \gamma \in I_n^2 \textit{ and } j \ge r,$$

$$(3) \quad \frac{\Pi_{n,\gamma}\big(A_{n,\gamma}(j\varepsilon_{n,\beta_n})\big)}{\Pi_{n,\gamma}\big(W_{n,\gamma}(BK_{n,\gamma}\varepsilon_{n,\beta_n})\big)} \le \mu_{n,\gamma}\, e^{Lj^2 n\varepsilon_{n,\beta_n}^2} \quad \text{for all } \gamma \in I_n^1 \text{ and } j \ge K_{n,\gamma},$$

where $1+K_{n,\gamma}$ *stands for the least integer but larger than* r *such that* $A_{n,\gamma}\big((1 + K_{n,\gamma})\varepsilon_{n,\beta_n}\big) \ne \emptyset$,

$$(4) \quad \sum_{n=1}^{\infty} \sum_{\gamma \in I_n^2} \frac{\lambda_{n,\gamma}\, \Pi_{n,\gamma}\big(A_{n,\gamma}(r\varepsilon_{n,\gamma})\big)\, e^{(3+2C)n\varepsilon_{n,\beta_n}^2}}{\lambda_{n,\beta_n}\, \Pi_{n,\beta_n}\big(W_{n,\beta_n}(\varepsilon_{n,\beta_n})\big)} < \infty.$$

Then

$$\Pi_n\big(f : d(f, f_0) \ge r\varepsilon_{n,\beta_n} \,\big|\, X_1, X_2, \ldots, X_n\big) \longrightarrow 0$$

almost surely as $n \to \infty$.

As an application of Theorem 3 to log spline densities with finitely many models, we successfully take away the logarithmic factor without using any special prior weights $\lambda_{n,\gamma}$ and hence for a true density in the Hölder space $C^\gamma[0, 1]$ the posterior $\Pi_n\big(\cdot \,\big|\, X_1, X_2, \ldots, X_n\big)$ attains the optimal rate of convergence in the minimax sense, which is well known to be $n^{-\gamma/(2\gamma+1)}$. This disconfirms a conjecture given by Ghosal, Lember and van der Vaart [7]. We moreover study consistency of posteriors of the model index and give a sufficient condition ensuring that the posteriors concentrate their masses near the index of the best model.

Paper IV

In Paper IV we introduce the Hausdorff α-entropy to study the strong Hellinger consistency of posteriors for i.i.d. observations. By means of the Hausdorff α-entropy, we obtained

Theorem 4. *Let $\varepsilon > 0$. Suppose that the true density f_0 is in the Kullback-Leibler support of Π and suppose that there exist $0 \leq \alpha < 1$, $0 < \delta < \frac{\varepsilon\,(1-\alpha)}{7}$, $c_1, c_2 > 0$, $0 < \beta < \frac{\varepsilon^2}{4}$, and a sequence $\{\mathcal{G}_n\}_{n=1}^{\infty}$ of subsets of the density space $\mathbb{F}$ such that each $\mathcal{G}_n$ is contained in $\cup_{j=1}^{\infty}\mathcal{G}_{nj}$. If*

(i) $\Pi\big(\mathbb{F} \setminus \mathcal{G}_n\big) < c_1\, e^{-n\, c_2}$;

(ii) $\sum\limits_{j=1}^{\infty} N(\delta, \mathcal{G}_{nj}, H)^{1-\alpha}\, \Pi(\mathcal{G}_{nj})^{\alpha} < e^{n\,\beta}$ *for all large n,*

then

$$\Pi\big(f \in \mathbb{F} :\, H(f_0, f) \geq \varepsilon \,\big|\, X_1, X_2, \ldots, X_n\big) \longrightarrow 0$$

almost surely as $n \to \infty$.

Theorem 4 contains several general posterior convergence theorems given by Barron, Schervish and Wasserman [1], Ghosal, Ghosh and Ramamoorthi [5] and Walker [13] as special cases. As applications we show that our theorem leads to improvements of some known results for some mixture models.

Paper V

Paper V deals with characterizations of exponential consistency of posteriors for i.i.d. observations. Recall that the true density is said to be in the Kullback-Leibler support of a prior if the prior mass on each Kullback-Leibler neighborhood of the true density is positive. It was proved in Freedman [4] and Diaconis and Freedman [2] that the Kullback-Leibler support condition cannot ensure the consistency of the posterior distribution. Many authors have obtained sufficient conditions for consistency of posteriors, see, for instance, Schwartz [11], Barron, Schervish and Wasserman [1], Ghosal, Ghosh and Ramamoorthi [5], Walker [13], Xing and Ranneby in Paper IV. The approaches of Barron et al. [1] and Ghosal et al. [5] are to construct suitable sieves and to compute metric entropies. Their works were discussed in great detail in the monograph of Ghosh and Ramamoorthi [9], see also the nice review of Wasserman [16]. Walker's result [13] relies upon summability of squareroots of prior probability of suitable coverings. Xing and Ranneby in Paper IV used the Hausdorff α-entropy to deal with the problem. All these results on posterior consistency are in fact to establish sufficient conditions on exponential consistency of posteriors under the Kullback-Leibler support condition. However, given the Kullback-Leibler support condition, much less is known about both necessary and sufficient conditions on exponential consistency of posteriors. We obtain one type of both necessary and sufficient conditions for Bayesian exponential consistency.

Theorem 5. *Let d be a semimetric on the density space $\mathbb{F}$ and let ε be a positive constant. If the true density f_0 is in the Kullback-Leibler support of Π, the following statements are equivalent.*

(i) *There exists a constant $\beta_1 > 0$ such that*

$$P_{f_0}^{\infty}\{\Pi\big(f \in \mathbb{F} : d(f, f_0) \geq \varepsilon \,\big|\, X_1, X_2, \ldots, X_n\big) > e^{-n\beta_1} \quad \textit{infinitely often}\ \} = 0.$$

(ii) *There exist constants $0 < \alpha_1 \leq 1$, $\beta_2 > 0$ and a sequence $\{D_n\}_1^{\infty}$ of sets $D_n \subset \mathbb{X}^n$ with $P_{f_0}^{\infty}(\limsup D_n) = 0$ such that*

$$E_{f_0}\left(1_{\mathbb{X}^n \setminus D_n} \int_{\{f \in \mathbb{F}: d(f,f_0) \geq \varepsilon\}} R_n(f)\, \Pi(df)\right)^{\alpha_1} \leq e^{-n\beta_2} \quad \textit{for all large } n,$$

where E_{f_0} stands for the expectation with respect to $X_1, X_2, \ldots, X_n$ and $1_{\mathbb{X}^n \setminus D_n}$ denotes the indicator function of $\mathbb{X}^n \setminus D_n$.

(iii) *For each $0 < \alpha \leq 1$ there exist a constant $\beta_\alpha > 0$ and a sequence $\{D_n\}_1^{\infty}$ of sets $D_n \subset \mathbb{X}^n$ with $P_{f_0}^{\infty}(\limsup D_n) = 0$ such that*

$$E_{f_0}\left(1_{\mathbb{X}^n \setminus D_n} \int_{\{f \in \mathbb{F}: d(f,f_0) \geq \varepsilon\}} R_n(f)\, \Pi(df)\right)^{\alpha} \leq e^{-n\beta_\alpha} \quad \textit{for all large } n.$$

(iv) *There exist a constant $\beta_3 > 0$ and a sequence $\{D_n\}_1^{\infty}$ of sets $D_n \subset \mathbb{X}^n$ such that $P_{f_0}^{\infty}(\limsup D_n) = 0$ and*

$$\int_{\{f \in \mathbb{F}: d(f,f_0) \geq \varepsilon\}} P_f^{\infty}(\mathbb{X}^n \setminus D_n)\, \Pi(df) \leq e^{-n\beta_3} \quad \textit{for all large } n.$$

An in-probability analogue of Theorem 5 is also established. As a consequence of Theorem 5 we derive a new sufficient condition on Bayesian consistency, which is weaker than the existing sufficient conditions. This makes it possible to obtain posterior consistency without any computation of metric entropy.

Paper VI

Bayesian consistency for i.i.d. data relies not only on the prior distribution, but also on how the likelihood function behaves as the sample size increases. So the size of likelihood function plays a key role in determining posterior consistency. In Paper VI we study Bayesian consistency by raising the likelihood function to a constant power α. A remarkable result of Walker and Hjort [14] is that, if one adopts the square root of the likelihood function in the calculation of posterior distributions, then the almost sure Hellinger consistency of the resulting pseudoposterior distributions holds only under the Kullback-Leibler support condition. They also proved that, by instead using the likelihood function with some power $0 < \alpha < 1$, the almost sure consistency of the pseudoposterior distributions Q_n^α holds, but only with respect to a related metric $H_\alpha(f,g) = \big(1 - \int g^\alpha f^{1-\alpha}\big)^{1/2}$. We extend their result and obtain the almost sure consistency of Q_n^α with respect to Hellinger metric for all $0 < \alpha < 1$, that is, we have

Theorem 6. *Let H be the Hellinger metric on the density space $\mathbb{F}$. If the true density f_0 is in the Kullback-Leibler support of the prior Π, then for each $\varepsilon > 0$ we have that*

$$Q_n^\alpha\big(f \in \mathbb{F} : H(f, f_0) \geq \varepsilon\big) \longrightarrow 0$$

almost surely as $n \to \infty$.

We moreover establish a sufficient condition ensuring almost sure Hellinger consistency of posterior distributions. Our result implies a theorem in Barron, Schervish and Wasserman [1].

References

[1] Barron, A., Schervish, M. and Wasserman, L. (1999). The consistency of posterior distributions in nonparametric problems. Ann. Statist. **27**, 536-561.

[2] Diaconis, P. and Freedman, D. (1986). On the consistency of Bayes estimates. Ann. Statist. **14**, 1-26.

[3] Doob, J. L. (1948). Application of the theory of martingales. Coll. Int. du CNRS, Paris, 22-28.

[4] Freedman, D. (1963). On the asymptotic behavior of Bayes estimates in the discrete case. Ann. Math. Statist. **34**, 1386-1403.

[5] Ghosal, S., Ghosh, J. K. and Ramamoorthi, R. V. (1999). Posterior consistency of Dirichlet mixtures in density estimation. Ann. Statist. **27**, 143-158.

[6] Ghosal, S., Ghosh, J. K. and Van der Vaart, A. W. (2000). Convergence rates of posterior distributions. Ann. Statist. **28**, 500-531.

[7] Ghosal, S., Lember, J. and Van der Vaart, A. W. (2003). On Bayesian adaptation. Acta Applicandae Mathematicae, **79**, 165-175.

[8] Ghosal, S., Lember, J. and Van der Vaart, A. W. (2008). Nonparametric Bayesian model selection and averaging. Electronic J. Statist. **2**, 63-89.

[9] Ghosh, J. K. and Ramamoorthi, R. V. (2003). Bayesian nonparametrics. Springer-Verlag, New York.

[10] Ghosal, S. and van der Vaart, A. W. (2007). Convergence rates of posterior distributions for noniid observations. Ann. Statist. **35**, 192-223.

[11] Schwartz, L. (1965). On Bayes procedures Z. Wahr. verw. Geb. **4**, 10-26.

[12] Shen, X. and Wasserman, L. (2001). Rates of convergence of posterior distributions. Ann. Statist. **29**, 687-714.

[13] Walker, S. (2004). New approaches to Bayesian consistency. Ann. Statist. **32**, 2028-2043.

[14] Walker, S. G. and Hjort, N. L. (2001). On Bayesian consistency. J. R. Statist. Soc., B **63**, 811-821.

[15] Walker, S., Lijoi, A. and Prunster, I. (2007). On rates of convergence for posterior distributions in infinite-dimensional models. Ann. Statist. **35**, 738-746.

[16] Wasserman, L. (1998). Asymptotic properties of nonparametric Bayesian procedures. in Practical Nonparametric and Semiparametric Bayesian Statistics, eds. D. Dey, P. Müller, and D. Sinha, New York: Springer-Verlag, 293-304.

Acknowledgements

I wish to express my gratitude to my supervisor Professor Bo Ranneby for his great support, many stimulating suggestions and discussions which have helped me very much during this research. Thank you for opening the door to statistics for me.

I am grateful to Professors Dietrich von Rosen, Lennart Bondesson, Bengt Kriström, Yuri Belyaev and Associate Professors Jun Yu and Magnus Ekström for their constant support and helpful suggestions.

A special thanks to Professor Christer Kiselman for his broad knowledge, support and encouragement.

I will also thank Professor Urban Cegrell for all his support through the years and for sharing his ideas and knowledge in both mathematics and life.

In addition I will thank all my colleagues in department of forest economics at Swedish university of agricultural sciences for their kindness and generosity.

And finally I will acknowledge my wife Xiaohong, our lovely daughter Viktoria and son Chen. Without their support this thesis would never have come into existence.

Umeå, December 2008

Yang Xing