

Frequentist and Bayesian nonparametrics for boundary models

Markus Reiß

Institut für Mathematik
Humboldt-Universität zu Berlin

www.mathematik.hu-berlin.de/~mreiss

Leiden, 17 June 2019

German fans want to hire ...



2005: HSV-Fans fordern die Verpflichtung von van der Vaart. © imago

Outline

Parametrics

Nonparametrics

Semiparametrics

Bayes

Outlook

The taxi problem

Observations: $X_1, \dots, X_n \sim U([0, \vartheta])$ i.i.d. for $\vartheta > 0$

MLE: $M_n := \max(X_1, \dots, X_n)$, $n \frac{\vartheta - M_n}{\vartheta} \rightarrow \text{Exp}(1)$

bias-corrected MLE: $\hat{\vartheta}_n = \frac{n+1}{n} M_n$ is UMVU (Lehmann-Scheffé)

Bayesian estimators are asymptotically efficient.

(Ibragimov, Has'minskii 1979; Chernozhoukov, Hong 2004; Kleijn, Knapik 2012)

Irregular model: convergence rate n^{-1} , non-Gaussian limit, LAE

The taxi problem

Observations: $X_1, \dots, X_n \sim U([0, \vartheta])$ i.i.d. for $\vartheta > 0$

MLE: $M_n := \max(X_1, \dots, X_n)$, $n \frac{\vartheta - M_n}{\vartheta} \rightarrow \text{Exp}(1)$

bias-corrected MLE: $\hat{\vartheta}_n = \frac{n+1}{n} M_n$ is UMVU (Lehmann-Scheffé)

Bayesian estimators are asymptotically efficient.

(Ibragimov, Has'minskii 1979; Chernozhoukov, Hong 2004; Kleijn, Knapik 2012)

Irregular model: convergence rate n^{-1} , non-Gaussian limit, LAE

The taxi problem

Observations: $X_1, \dots, X_n \sim U([0, \vartheta])$ i.i.d. for $\vartheta > 0$

MLE: $M_n := \max(X_1, \dots, X_n)$, $n \frac{\vartheta - M_n}{\vartheta} \rightarrow \text{Exp}(1)$

bias-corrected MLE: $\hat{\vartheta}_n = \frac{n+1}{n} M_n$ is UMVU (Lehmann-Scheffé)

Bayesian estimators are asymptotically efficient.

(Ibragimov, Has'minskii 1979; Chernozhoukov, Hong 2004; Kleijn, Knapik 2012)

Irregular model: convergence rate n^{-1} , non-Gaussian limit, LAE

The taxi problem

Observations: $X_1, \dots, X_n \sim U([0, \vartheta])$ i.i.d. for $\vartheta > 0$

MLE: $M_n := \max(X_1, \dots, X_n)$, $n^{\frac{\vartheta - M_n}{\vartheta}} \rightarrow \text{Exp}(1)$

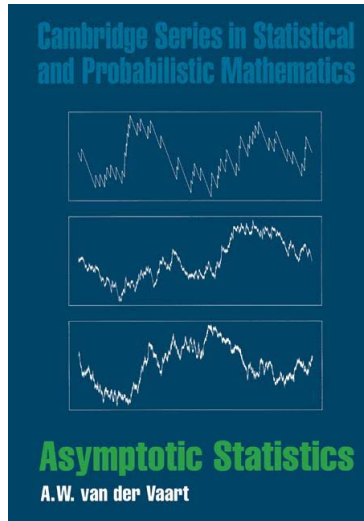
bias-corrected MLE: $\hat{\vartheta}_n = \frac{n+1}{n} M_n$ is UMVU (Lehmann-Scheffé)

Bayesian estimators are asymptotically efficient.

(Ibragimov, Has'minskii 1979; Chernozhoukov, Hong 2004; Kleijn, Knapik 2012)

Irregular model: convergence rate n^{-1} , non-Gaussian limit, *LAE*

Let's check



9.4 Uniform Distribution

The model consisting of the uniform distributions on $[0, \theta]$ is not differentiable in quadratic mean (see Example 7.9.) In this case an asymptotically normal approximation is impossible. Instead, we have convergence to an exponential experiment.

9.6 Theorem. Let P_θ^n be the distribution of a random sample of size n from a uniform distribution on $[0, \theta]$. Then the sequence of experiments $(P_{\theta-h_0/n}^n; h \in \mathbb{R})$ converges for each fixed $\theta > 0$ to the experiment consisting of observing one observation from the shifted exponential density $z \mapsto e^{-(z-h)/\theta} 1\{z > h\}/\theta$.¹

Proof. If Z is distributed according to the given exponential density, then

$$\frac{dP_\theta^n}{dP_{h_0}^n}(Z) = \frac{e^{-(Z-h)/\theta} 1\{Z > h\}/\theta}{e^{-(Z-h_0)/\theta} 1\{Z > h_0\}/\theta} = e^{(h-h_0)/\theta} 1\{Z > h\},$$

almost surely under h_0 , because the indicator $1\{Z > h_0\}$ in the denominator equals 1 almost surely if h_0 is the true parameter.

The joint density of a random sample $X_1, \dots, X_n$ from the uniform $[0, \theta]$ distribution can be written in the form $(1/\theta)^n 1\{X_{(n)} \leq \theta\}$. The likelihood ratios take the form

$$\frac{dP_{\theta-h_0/n}^n}{dP_{\theta-h_0/n}^n}(X_1, \dots, X_n) = \frac{(\theta - h/n)^{-n} 1\{X_{(n)} \leq \theta - h/n\}}{(\theta - h_0/n)^{-n} 1\{X_{(n)} \leq \theta - h_0/n\}}.$$

Under the parameter $\theta - h_0/n$, the maximum of the observations is certainly bounded above by $\theta - h_0/n$ and the indicator in the denominator equals 1. Thus, with probability 1 under $\theta - h_0/n$, the likelihood ratio in the preceding display can be written

$$(e^{(h-h_0)/\theta} + o(1)) 1\{-n(X_{(n)} - \theta) \geq h\}.$$

By direct calculation, $-n(X_{(n)} - \theta) \xrightarrow{\text{d}} Z$. By the continuous-mapping theorem and Slutsky's lemma, the sequence of likelihood processes converges under $\theta - h_0/n$ marginally in distribution to the likelihood process of the exponential experiment. ■

Along the same lines it may be proved that in the case of uniform distributions with both endpoints unknown a limit experiment based on observation of two independent exponential variables pertains. These types of experiments are completely determined by the discontinuities of the underlying densities at their left and right endpoints. It can be shown more generally that exponential limit experiments are obtained for any densities that have jumps at one or both of their endpoints and are smooth in between. For densities with discontinuities in the middle, or weaker singularities, other limit experiments pertain.

The convergence to a limit experiment combined with the asymptotic representation theorem, Theorem 9.3, allows one to obtain asymptotic lower bounds for sequences of estimators, much as in the locally asymptotically normal case in Chapter 8. We give only one concrete statement.

¹ Define P_θ arbitrarily for $\theta < 0$.

9.7 Corollary. Let T_n be estimators based on a sample $X_1, \dots, X_n$ from the uniform distribution on $[0, \theta]$ such that the sequence $n(T_n - \theta)$ converges under θ in distribution to a limit L_θ , for every θ . Then for Lebesgue almost-every θ we have $\int |x| dL_\theta(x) \geq E|Z - \text{med } Z|$ and $\int x^2 dL_\theta(x) \geq E(Z - EZ)^2$ for the random variable Z exponentially distributed with mean θ .

Proof (Sketch). By Lemma 8.10, the estimator sequence T_n is automatically almost regular in the sense that $n(T_n - \theta + h/n)$ converges under $\theta - h/n$ in distribution to L_θ for Lebesgue almost every θ and h , at least along a subsequence. Thus, it is matched in the limit experiment by an equivariant-in-law estimator for almost every θ . More precisely, for almost every θ there exists a randomized statistic T_θ such that the law of $T_\theta(Z + h, U) - h$ does not depend on h (if Z is exponentially distributed with mean θ). By classical statistical decision theory the given lower bounds are the (constant) risks of the best equivariant-in-law estimators in the exponential limit experiment in terms of absolute error and mean-square error loss functions, respectively. ■

In view of this lemma, the maximum likelihood estimator $X_{(n)}$ is asymptotically inefficient. This is not surprising given its bias downwards, but it is encouraging for the present approach that the small bias, which is of the order $1/n$, is visible in the "first-order" asymptotics. The bias can be corrected by a multiplicative factor, which, unfortunately, must depend on the loss function. The sequences of estimators

$$\frac{n + \log 2}{n} X_{(n)} \quad \text{and} \quad \frac{n + 1}{n} X_{(n)}$$

are asymptotically efficient in terms of absolute value and quadratic loss, respectively.

9.5 Pareto Distribution

The Pareto distributions are a two-parameter family of distributions on the real line with parameters $\alpha > 0$ and $\mu > 0$ and density

$$x \mapsto \frac{\alpha \mu^\alpha}{x^{\alpha+1}} 1\{x > \mu\}.$$

This density is smooth in α , but it resembles a uniform distribution as discussed in the preceding section in its dependence on μ . The limit experiment consists of a combination of a normal experiment and an exponential experiment.

The likelihood ratios for a sample of size n from the Pareto distributions with parameters $(\alpha + g/\sqrt{n}, \mu + h/n)$ and $(\alpha + g_0/\sqrt{n}, \mu + h_0/n)$, respectively, is equal to

$$\begin{aligned} & \left(\frac{\alpha + g/\sqrt{n}}{\alpha + g_0/\sqrt{n}} \right)^n \left(\frac{(\mu + h/n)^{\alpha + g/\sqrt{n}}}{(\mu + h_0/n)^{\alpha + g_0/\sqrt{n}}} \left(\prod_{i=1}^n X_i \right)^{-(\alpha + g/\sqrt{n})} \right) 1\{X_{(1)} > \mu + \frac{h}{n}\} \\ & = \exp \left((g - g_0) \Delta_\alpha - \frac{1}{2} \frac{g^2 + g_0^2}{\alpha^2} + o(1) \right) \left(e^{(h-h_0)/\mu + o(1)} 1\{Z_\alpha > h\} \right). \end{aligned}$$

Nonparametric boundary models

Nonparametric regression

$$Y_i = f(i/n) + \varepsilon_i, \quad i = 1, \dots, n$$

Regular noise: Asymptotic Le Cam equivalence with

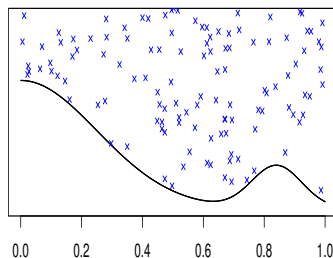
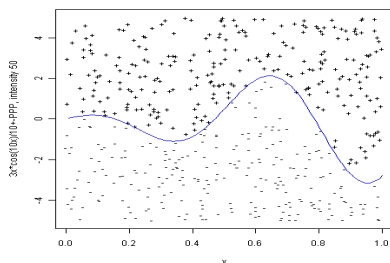
$$dY(t) = f(t)dt + \frac{\sigma}{\sqrt{n}}dW_t, \quad t \in [0, 1]$$

(Brown, Low 1996, Grama, Nussbaum 2002)

What is the continuous limit model for $\varepsilon_i \sim U([-a, a])$?

One picture says it all...

PPP boundary models



Theorem. (Meister, MR 2013) There is asymptotic equivalence:

- for $\varepsilon_i \sim U([-a, a])$ with observing two independent Poisson point processes (PPP) with intensity na^{-1} on the epigraph $\{(x, y) : y \geq f(x)\}$ and hypograph $\{(x, y) : y \leq f(x)\}$;
- for $\varepsilon_i \sim \text{Exp}(\lambda)$ with observing one PPP with intensity $n\lambda$ on the epigraph $\{(x, y) : y \geq f(x)\}$.

Working with the PPP model

Observation: PPP $(X_i, Y_i)_{i \geq 1}$ of intensity n on the epigraph of f .

Constant model: For $f(x) = \vartheta$ is $\min_i Y_i \sim \vartheta + \text{Exp}(n)$ MLE.

Locally constant: $\hat{f}_{n,h}(x) = \min\{Y_i \mid |X_i - x| \leq h\}$ estimates $f(x)$.

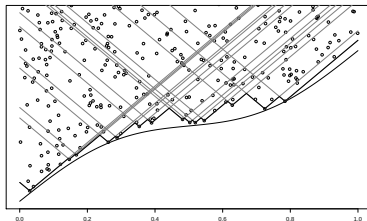
Minimax rates: On Hölder balls $C^\alpha(R)$ is $n^{-\alpha/(\alpha+1)}$ optimal.

(Hall, van Keilegom 2009, Jirak, Meister, MR 2014)

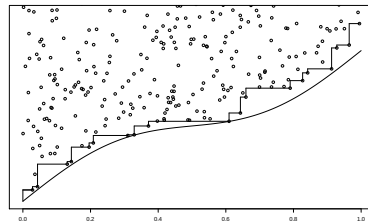
Nonparametric MLE

Likelihood: For the law P_f of the PPP on the epigraph of $f \in L^1([0, 1])$ and $f \geq f_0$

$$\frac{dP_f}{dP_{f_0}} = \exp \left(n \int_0^1 (f - f_0)(x) dx \right) \mathbf{1}(\forall i : f(X_i) \leq Y_i).$$



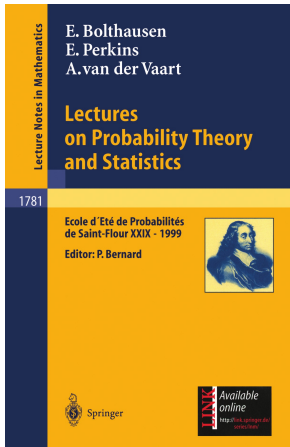
Lipschitz-MLE



monotone MLE

The nonparametric MLEs on Hölder and monotone balls achieve the optimal nonparametric rates, but biased.

What about linear functionals?



give an abstract definition of “likelihood”, but shall describe “likelihoods that work” for a number of examples to set the stage. We denote the likelihood for the parameter P given one observation x by $\text{lik}(P)(x)$ or $\text{lik}(\theta, \eta)$ if $P = P_{\theta, \eta}$.

Given a measure P , write $P\{x\}$ for the measure of the one-point set $\{x\}$. The function $x \rightarrow P\{x\}$ may be considered the density of P , or its absolutely continuous part, with respect to counting measure. The *empirical likelihood* of a sample $X_1, \dots, X_n$ is the function,

$$P \rightarrow \prod_{i=1}^n P\{X_i\}.$$

Given a model $\mathcal{P}$, a maximum likelihood estimator could be defined as the distribution $\hat{P}$ that maximizes the empirical likelihood over $\mathcal{P}$. Such an estimator may or may not exist.

Example 9.1 (Empirical distribution). Let $\mathcal{P}$ be the set of all probability distributions on the measurable space $(X, \mathcal{A})$ (in which one-point sets are measurable). Then, for n fixed different values $x_1, \dots, x_n$, the vector $(P\{x_1\}, \dots, P\{x_n\})$ ranges over all vectors $p \geq 0$ such that $\sum p_i \leq 1$ when P ranges over $\mathcal{P}$. To maximize $p \rightarrow \prod p_i$, it is clearly best to choose p maximal: $\sum p_i = 1$. Then, by symmetry, the maximizer must be $p = (1/n, \dots, 1/n)$. Thus, the empirical distribution $\mathbb{P}_n = n^{-1} \sum \delta_{X_i}$ maximizes the empirical likelihood over the nonparametric model, whence it is referred to as the *nonparametric maximum likelihood estimator*.

If there are ties in the observations, this argument must be adapted, but the result is the same.

The empirical likelihood is appropriate for the nonparametric model. For instance, in the case of a Euclidean space, even if the model would be restricted to distributions with a continuous Lebesgue density p , then we still could not use the map $p \rightarrow \prod_{i=1}^n p(X_i)$ as a likelihood. The supremum of this “likelihood” is infinite, for we could choose p to have an arbitrarily high, very thin peak at some observation.

Open Problem 9.2. Suppose we use $p \rightarrow \prod_{i=1}^n p(X_i)$ as a likelihood, restricted to a Hölder ball of densities $p: [0, 1] \rightarrow \mathbb{R}$, e.g. all densities which are twice continuously differentiable with second derivative bounded by 1 and which are themselves bounded by some fixed number. Is it true that $\int h(x) \hat{p}(x) dx$ is an asymptotically efficient estimator for $\psi(P) = \int h(x) p(x) dx$ for every reasonable function h ?

Example 9.3 (Cox model). We already discussed the problem of finding a likelihood for the Cox model in Lecture 5. There we settled on using the function

$$\text{lik}(\theta, A)(y, \delta, x) = \left(e^{\theta x} A\{y\} e^{-e^{\theta x} A(y)} \right)^{\delta} \left(e^{-e^{\theta x} A(y)} \right)^{1-\delta}.$$

What about linear functionals?

Semiparametric target: $\psi(f) = \int_0^1 f(x)w(x)dx$

for some weight function $w \in L^2$

Plug-in: Is $\hat{\psi} = \int_0^1 \hat{f}(x)w(x)dx$ efficient?

Answer: No, not even rate-optimal;
both, for nonparametric MLE or best local polynomial $\hat{f}$.

Reason: the inherent bias!

Question: Is there a nonparametric bias correction?

What about linear functionals?

Semiparametric target: $\psi(f) = \int_0^1 f(x)w(x)dx$

for some weight function $w \in L^2$

Plug-in: Is $\hat{\psi} = \int_0^1 \hat{f}(x)w(x)dx$ efficient?

Answer: No, not even rate-optimal;
both, for nonparametric MLE or best local polynomial $\hat{f}$.

Reason: the inherent bias!

Question: Is there a nonparametric bias correction?

What about linear functionals?

Semiparametric target: $\psi(f) = \int_0^1 f(x)w(x)dx$

for some weight function $w \in L^2$

Plug-in: Is $\hat{\psi} = \int_0^1 \hat{f}(x)w(x)dx$ efficient?

Answer: No, not even rate-optimal;
both, for nonparametric MLE or best local polynomial $\hat{f}$.

Reason: the inherent bias!

Question: Is there a nonparametric bias correction?

Unbiased functional estimation

Unbiased estimator:

$$\hat{\psi} = \int_0^1 \hat{f}^{MLE}(x) w(x) dx - \frac{1}{n} \sum_{i \geq 1} \mathbf{1}(\hat{f}^{MLE}(X_i) = Y_i) w(X_i)$$

Theorem.(MR, Selk 2017) $\hat{\psi}$ is for each n **UMVU** over $\mathcal{C}^\alpha(R)$, $\alpha \in (0, 1]$, $R > 0$, with

$$\begin{aligned} \text{Var}(\hat{\psi}) &= \frac{1}{n} \int_0^1 \mathbb{E}[\hat{f}^{MLE}(x) - f(x)] w(x)^2 dx \\ &\lesssim R^{1/(\alpha+1)} n^{-(2\alpha+1)/(\alpha+1)} \|w\|_{L^2}^2. \end{aligned}$$

The rate $n^{-(\alpha+1/2)/(\alpha+1)}$ is minimax optimal over $\mathcal{C}^\alpha(R)$.
Also for monotone functions $\hat{\psi}$ is UMVU with rate $n^{-3/4}$.

cf. Korostelev, Tsybakov (1993), Gayraud (1997)

Nonlinear functionals and testing

L^p -functional: $\psi(f) = \|f\|_{L^p}^p = \int_0^1 |f(x)|^p dx$, $p \geq 1$.

Unbiased estimator:

$$\hat{\psi} = \int_0^1 |\hat{f}^{MLE}(x)|^p dx - \frac{1}{n} \sum_{i \geq 1} \mathbf{1}(\hat{f}^{MLE}(X_i) = Y_i) p |Y_i|^{p-1} \operatorname{sgn}(Y_i)$$

Theorem. (MR, Wahl 2019) $\hat{\psi}$ is for each n UMVU over $C^\alpha(R)$, $\alpha \in (0, 1]$, $R > 0$, with

$$\operatorname{Var}(\hat{\psi}) \lesssim \max \left(n^{-(2\alpha+1)/(\alpha+1)} \int_0^1 |g(x)|^{2p-2} dx, n^{-(2\alpha p+1)/(\alpha+1)} \right).$$

Testing $H_0 : f = 0$ versus $H_{1,n} : f \in C^\alpha(R)$, $\|f\|_{L^p} \geq r_n$ is possible for the optimal separation rate $r_n = n^{-(\alpha+1/(2p))/(\alpha+1)}$.

cf. Lepski, Nemirovski, Spokoiny (1999), Ingster, Suslina (2003)

Bayesian nonparametrics

The Annals of Statistics
2006, Vol. 34, No. 5, 1435–1463
DOI: 10.1214/009053607000000613
© Institute of Mathematical Statistics, 2006

RATES OF CONTRACTION OF POSTERIOR DISTRIBUTIONS BASED ON GAUSSIAN PROCESS PRIORS

BY A. W. VAN DER VAART AND J. H. VAN ZANTEN¹

Vrije Universiteit Amsterdam

We derive rates of contraction of posterior distributions on nonparametric or semiparametric models based on Gaussian processes. The rate of contraction is shown to depend on the position of the true parameter relative to the reproducing kernel Hilbert space of the Gaussian process and the small ball probabilities of the Gaussian process. We determine these quantities for a range of examples of Gaussian priors and in several statistical settings. For instance, we consider the rate of contraction of the posterior distribution based on sampling from a smooth density model when the prior models the log density as a (fractionally integrated) Brownian motion. We also consider regression with Gaussian errors and smooth classification under a logistic or probit link function combined with various priors.

1. Introduction. Gaussian processes have been adopted as building blocks for constructing prior distributions on infinite-dimensional statistical models in several settings. For instance, in the setting of nonparametric density estimation, a prior distribution on a collection of probability densities (relative to a measure ν) can be defined structurally as the random density

$$(1.1) \quad \frac{e^{W_x}}{\int e^{W_s} d\nu(x)},$$

where $(W_x : x \in \mathcal{X})$ is a Gaussian process indexed by the sample space $\mathcal{X}$ of the observations. The Gaussian process is exponentiated to force the prior to charge only nonnegative functions, and is next renormalized to integrate to unity. Several other constructions have also been considered in the literature, for density estimation as well as other statistical problems; see Section 3 and [4, 5, 9, 12, 15, 19–22, 26, 29, 34]. The book [27] makes a connection to machine learning and the website <http://www.gaussianprocess.org> lists additional references.

Given a prior and observations, Bayes' rule yields a posterior distribution on the parameter space. In the frequentist set-up, in which the data are sampled from a fixed "true" distribution and the amount of information in the data increases indefinitely, the corresponding posterior distributions often contract to the fixed

1436

A. W. VAN DER VAART AND J. H. VAN ZANTEN

true distribution, which is referred to as *posterior consistency*. In this paper, we study the *rate of contraction* of the posterior distribution relative to global metrics on the parameters.

In most cases, the Gaussian process can be viewed as a tight Borel measurable map in a Banach space, for instance, a space of continuous functions or an L_p -space. It is well known that the support of a centered (i.e., zero-mean) version of such a process (the smallest closed set having probability one under the induced measure) is equal to the closure of the *reproducing kernel Hilbert space* (RKHS) of the covariance kernel of the process. Because the posterior distribution necessarily puts all of its mass on the support of the prior, it follows that consistency can be valid only if the parameter w_0 defining the true distribution of the data belongs to this support. In the present paper, we prove that the rate of contraction in that case is expressible in terms of the function

$$(1.2) \quad \phi_{w_0}(\varepsilon) = \inf_{h \in \mathbb{H} : \|h - w_0\| < \varepsilon} \|h\|_{\mathbb{H}}^2 - \log \Pr(\|W\| < \varepsilon).$$

In this expression, $\|\cdot\|$ is the norm of the Banach space in which the Gaussian process W takes its values, $\mathbb{H}$ is the reproducing kernel Hilbert space of the process and $\|\cdot\|_{\mathbb{H}}$ the RKHS-norm. If the norm $\|\cdot\|$ on the sample space of the process "combines correctly" with the norm on the parameter space and n describes the informativeness of the data in the usual way, then the posterior contracts at the rate $\varepsilon_n \rightarrow 0$, satisfying

$$(1.3) \quad \phi_{w_0}(\varepsilon_n) \leq n\varepsilon_n^2.$$

This is the case, for instance, in density estimation on the unit interval with the log Gaussian process prior given in (1.1) and $\|\cdot\|$ the uniform norm given by $\|w\| = \sup\{|w(x)| : x \in \mathcal{X}\}$. This is also the case for regression and classification, with appropriate norms, as shown below. The rate of contraction ε_n thus depends on the position of the true parameter w_0 relative to the RKHS and the amount of mass $\Pr(\|W\| < \varepsilon)$ that the prior distribution puts in small balls around zero. In Section 4, we compute these quantities for a range of priors.

For instance, we prove that log Gaussian densities given in (1.1), combined with Brownian motion, yield a rate of contraction of $n^{-1/4}$ whenever the logarithm of the true density is α -smooth for some $\alpha \geq 1/2$ and yield the slower rate $n^{-\alpha/2}$ for $0 < \alpha \leq 1/2$. That higher smoothness (α large) does not improve the rate of contraction is disappointing, but perhaps not surprising, given that the Brownian paths themselves are $1/2$ -smooth: the data are not capable of smoothing out the prior roughness of the sample paths. Other, more smooth, Gaussian priors give better rates for smooth truths (depending on their RKHS), but worse rates for rough truths. In Section 4, we exhibit, for every possible smoothness level α , Gaussian priors that give the optimal rate of contraction if the true parameter possesses regularity α .

¹Received May 2006; revised March 2007.

²Supported by the Netherlands Organization for Scientific Research NWO.

AMS 2000 subject classifications. 60G15, 62G05.

Key words and phrases. Rate of convergence, Bayesian inference, nonparametric density estimation, nonparametric regression, classification.

Posterior contraction rates

Bayes formula:

$$\Pi(B|(X_i, Y_i)_{i \geq 1}) \propto \int_B e^{-n \int (f_0 - f)} + \frac{dP_{f \vee f_0}}{dP_{f_0}}((X_i, Y_i)_{i \geq 1}) d\Pi(f) \quad P_{f_0}\text{-a.s.}$$

Posterior contraction: (MR, Schmidt-Hieber 2018a)

For a Gaussian prior Π with RKHS H assume $\varepsilon_n \geq n^{-1}$ and

$$\inf_{h: \|h + 2\varepsilon_n - f_0\|_\infty \leq \varepsilon_n} \|h\|_H^2 - \log \Pi(\|X\|_\infty \leq \varepsilon_n) \leq n\varepsilon_n$$

then there exists a constant M such that for all n

$$E_{f_0} [\Pi(f : \|f - f_0\|_{L^1} \geq M\varepsilon_n | N)] \leq 3e^{-n\varepsilon_n}$$

Gaussian, series expansion and compound Poisson (CPP) priors can achieve posterior contraction rate $n^{-\alpha/(\alpha+1)}$ over $C^\alpha(R)$. CPP also achieve $n^{-1/2}$ for monotone f .

Posterior contraction rates

Bayes formula:

$$\Pi(B|(X_i, Y_i)_{i \geq 1}) \propto \int_B e^{-n \int (f_0 - f)} \frac{dP_{f \vee f_0}}{dP_{f_0}}((X_i, Y_i)_{i \geq 1}) d\Pi(f) \quad P_{f_0}\text{-a.s.}$$

Posterior contraction: (MR, Schmidt-Hieber 2018a)

For a Gaussian prior Π with RKHS H assume $\varepsilon_n \geq n^{-1}$ and

$$\inf_{h: \|h + 2\varepsilon_n - f_0\|_\infty \leq \varepsilon_n} \|h\|_H^2 - \log \Pi(\|X\|_\infty \leq \varepsilon_n) \leq n\varepsilon_n$$

then there exists a constant M such that for all n

$$E_{f_0} [\Pi(f : \|f - f_0\|_{L^1} \geq M\varepsilon_n | N)] \leq 3e^{-n\varepsilon_n}$$

Gaussian, series expansion and compound Poisson (CPP) priors can achieve posterior contraction rate $n^{-\alpha/(\alpha+1)}$ over $C^\alpha(R)$. CPP also achieve $n^{-1/2}$ for monotone f .

Semiparametric Bernstein-von Mises?

Target: $\psi(f) = \int_0^1 f(x)dx$ for monotone f

Prior: compound Poisson with intensity λ and $\Gamma(2, 1)$ -jumps

Functions: piecewise constant monotone $f_0 \in M_S(K_n, R)$, i.e.

$$f_0 = \sum_{\ell=0}^{K_n} a_\ell \mathbf{1}(\bullet \geq t_\ell) : 0 \leq a_\ell \leq R, 0 \leq t_1 \leq \dots \leq t_{K_n} \leq 1$$

with $K_n \ll n^{1/2}$, $t_{\ell+1} - t_\ell \gg n^{-1/2}$, $a_\ell(t_{\ell+1} - t_\ell) \gg K_n n^{-1}$.

A semiparametric BvM result: (MR, Schmidt-Hieber 2018b)

$$\sup_{f_0 \in M_S(K_n, R)} E_{f_0}^n \left[\left\| \Pi(\psi \in \bullet | N) - \mathcal{N} \left(\int \hat{f}_{K_n}^{MLE} - \frac{2K_n+1}{n}, \frac{2K_n+1}{n^2} \right) \right\|_{KS} \right] \rightarrow 0.$$

But: no posterior coverage for $f_0(x) = x!$ (Castillo, Rousseau 2015)

Semiparametric Bernstein-von Mises?

Target: $\psi(f) = \int_0^1 f(x)dx$ for monotone f

Prior: compound Poisson with intensity λ and $\Gamma(2, 1)$ -jumps

Functions: piecewise constant monotone $f_0 \in M_S(K_n, R)$, i.e.

$$f_0 = \sum_{\ell=0}^{K_n} a_\ell \mathbf{1}(\bullet \geq t_\ell) : 0 \leq a_\ell \leq R, 0 \leq t_1 \leq \dots \leq t_{K_n} \leq 1$$

with $K_n \ll n^{1/2}$, $t_{\ell+1} - t_\ell \gg n^{-1/2}$, $a_\ell(t_{\ell+1} - t_\ell) \gg K_n n^{-1}$.

A semiparametric BvM result: (MR, Schmidt-Hieber 2018b)

$$\sup_{f_0 \in M_S(K_n, R)} E_{f_0}^n \left[\left\| \Pi(\psi \in \bullet | N) - \mathcal{N} \left(\int \hat{f}_{K_n}^{MLE} - \frac{2K_n+1}{n}, \frac{2K_n+1}{n^2} \right) \right\|_{KS} \right] \rightarrow 0.$$

But: no posterior coverage for $f_0(x) = x$! (Castillo, Rousseau 2015)

Summary: boundary models

- Parametrics:
 - Bias-corrected MLE and Bayesian estimators are efficient.
- Nonparametrics:
 - Asymptotic equivalence of regression and PPP models.
 - Rate optimality for local polynomials, MLE, Bayes.
- Semiparametrics:
 - Bias corrected MLE is UMVU for (non)linear functionals.
 - BvM for growing parameter spaces, but not generally.
- Same ideas: α -convex support set estimation

Thank you for your attention.