

On Bayesian Adaptive Regression Function Estimation

Sulla Stima Bayesiana Adattiva della Funzione di Regressione

Catia Scricciolo^(*)

Istituto di Metodi Quantitativi, Università “L. Bocconi”, Milano

e-mail: catia.scricciolo@uni-bocconi.it

Riassunto: Il problema della stima non parametrica della funzione di regressione in un modello con disegno aleatorio e distribuzione gaussiana dell'errore viene considerato adottando un'impostazione bayesiana. Se è noto che la funzione di regressione è derivabile, ma non si dispone di informazioni sul grado, si pone anche il problema della scelta della classe di funzioni su cui definire la distribuzione iniziale. Il problema della scelta del modello può essere affrontato in modo naturale incorporando nella legge iniziale anche l'incertezza sul grado di regolarità della funzione. Nel presente lavoro, vengono fornite delle condizioni sulla distribuzione iniziale affinché lo stimatore bayesiano della funzione di regressione sia *adattivo*, converga cioè con la velocità ottimale relativa alla classe, qualunque sia tra quelle considerate, cui la funzione di regressione appartiene.

Keywords: Bayesian adaptive estimation, nonparametric regression, rate of convergence.

1. Introduction

The problem of nonparametric regression consists in estimating the conditional mean function which models the relationship of stochastic dependence of a response variable Y on a covariate X when such function is assumed to belong to some functional class. But, choosing a proper parameter space from a list of possible models is a difficult task. Both problems of estimation and model selection in the framework of a *random design Gaussian regression model* can be encompassed following a Bayesian approach.

Consider the problem of estimating a regression function $r_0(\cdot)$ based on a random sample $Z_1 = (X_1, Y_1), \dots, Z_n = (X_n, Y_n)$ with the same distribution as $Z = (X, Y)$ satisfying

$$Y = r_0(X) + \xi,$$

for independent random variables (r.v.'s) X and ξ taking values in $(\mathbb{X}, \mathcal{B}(\mathbb{X}))$, with $\mathbb{X} \subset \mathbb{R}$, and $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, respectively. Let P_X be the distribution of X with density f_X with respect to (w.r.t.) some σ -finite measure ν . Throughout the paper the assumption that f_X is bounded and bounded away from zero will be in force. Let ξ have distribution $N(0, \sigma^2)$, with σ^2 known. Conditionally on X , the response r.v. Y is Gaussian with mean $r_0(X)$ and variance σ^2 . The joint density of (X, Y) w.r.t. $\nu \otimes \lambda$, where λ denotes Lebesgue measure on $\mathbb{R}$, is

$$f_{0,\sigma}(x, y) = f_{Y|X}(y; r_0, \sigma) f_X(x) = (2\pi\sigma^2)^{-1/2} \exp\{-(y-r_0(x))^2/2\sigma^2\} f_X(x).$$

^(*) Address for correspondence: Catia Scricciolo, Istituto di Metodi Quantitativi, Università “L. Bocconi”, Viale Isonzo 25, I-20135 Milano, Italy.

The corresponding probability measure generating the observations $Z_1, \dots, Z_n$ that take values in $\mathbb{Z} = \mathbb{X} \times \mathbb{R}$ will be denoted by P_0 . Since the regression function is a random element, a prior distribution need to be defined on some class $\mathcal{R} \subseteq L_2(P_X)$ of measurable functions $r : \mathbb{X} \rightarrow \mathbb{R}$. First, suppose that Y is known to be a p -smooth function of the explanatory variable X plus some random perturbation. Then, the regression function can be estimated with a variety of methods, either frequentist (kernel, truncated series, wavelets, etc.) or Bayesian with optimal rate of convergence $\epsilon_{n,p} = n^{-p/(1+2p)}$ under $L_2(P_X)$ -loss in a minimax sense. Suppose now that no prior knowledge on the degree of smoothness is available so that the conditional mean is a smooth, but otherwise unknown, function. Then, rather than a single model we have a list of models and are faced also with the problem of selecting a model: a choice among elements in a list has to be made based on the observations. In a frequentist perspective, the rate-adaptive estimators select the right rate among a collection of possible values as when only one model is considered. In a Bayesian framework, uncertainty in the model can be incorporated in the prior measure by combining a prior on the smoothness parameter indexing the models with the priors on the single models: first the model, say $\mathcal{R}_q$, with index q is chosen by a prior putting weights $w_q \geq 0$ ($\sum_{q \in \mathbb{Q}} w_q = 1$) on the index set $\mathbb{Q}$ and next the regression function is selected from the chosen model using a prior Π_q on it. This corresponds to using the compound prior $\Pi = \sum_{q \in \mathbb{Q}} w_q \Pi_q$ on the union $\mathcal{R} = \cup_{q \in \mathbb{Q}} \mathcal{R}_q$. If the regression function belongs to the model with unknown index p and the rate of convergence corresponding to this model pertains, whatever $p \in \mathbb{Q}$, we then have *full rate adaptation*, namely, the obtained rate is the same as when p is known and only the model $\mathcal{R}_p$ with index p is considered. There are still few instances on Bayesian adaptation in the literature: Belitser and Ghosal (2003) study adaptive estimation for the mean of an infinite-dimensional normal distribution, Ghosal *et al.* (2003) investigate adaptive log-spline density estimation, Huang (2004) provides general theorems on rates of convergence for posterior distributions that can be used for adaptive density or regression function estimation.

In this paper we show that the posterior expected regression function entailed by a mixture prior over a list of models for different smoothness degrees, achieves the optimal rate of convergence in the frequentist sense. The main result is stated in the next section where also an illustration is presented. Final remarks are deferred to Section 3.

2. Main Result

We begin to study the pointwise behavior of the posterior distribution. We assume that the optimal rate of convergence relative to each model $\mathcal{R}_q$ is $\epsilon_{n,q}$, with $n\epsilon_{n,q}^2 \rightarrow \infty$ as $n \rightarrow \infty$.

Theorem 1 *Let $\mathbb{Q}$ be a finite or countable set of positive integers. Let $\mathcal{R} = \cup_{q \in \mathbb{Q}} \mathcal{R}_q$ be a class of uniformly bounded functions such that $r_0 \in \mathcal{R}_p$, for some $p \in \mathbb{Q}$. Let $\Pi = \sum_{q \in \mathbb{Q}} w_q \Pi_q$, with $w_q > 0$ for all $q \in \mathbb{Q}$. If for a constant $c > 0$,*

$$\Pi_p(\{r \in \mathcal{R}_p : \|r - r_0\|_2 \leq c\epsilon_{n,p}\}) \geq c_1 e^{-Cn\epsilon_{n,p}^2}, \quad (1)$$

with $c_1, C > 0$ possibly depending on r_0 , then for a constant $M > 0$ large enough,

$$\Pi(\{r : \|r - r_0\|_2 \geq M\epsilon_{n,p}\} | Z_1, \dots, Z_n) \rightarrow 0$$

as $n \rightarrow \infty$, P_0^∞ -almost surely, where P_0^∞ denotes the infinite product measure of P_0 .

Proof. For ease of notation, let $B(r_0; \epsilon_{n,p}) = \{r : \|r - r_0\|_2 \geq M\epsilon_{n,p}\}$ and $R_n(f_{r,\sigma}) = \prod_{i=1}^n (f_{r,\sigma}(Z_i)/f_{0,\sigma}(Z_i))$, where $f_{r,\sigma}(\cdot, \cdot) = f_{Y|X}(\cdot; r, \sigma)f_X(\cdot)$. The probability in the last display can be written as

$$\Pi(B(r_0; \epsilon_{n,p}) | Z_1, \dots, Z_n) = \frac{\sum_{q \in \mathbb{Q}} w_q \int_{\mathcal{R}_q \cap B(r_0; \epsilon_{n,p})} R_n(f_{r,\sigma}) d\Pi_q(r)}{\int R_n(f_{r,\sigma}) d\Pi(r)} \equiv \frac{U_n}{L_n}.$$

We bound U_n . Let $B > 0$ be that constant for which $\sup_{r \in \mathcal{R}} \|r\|_\infty \leq B$, where $\|\cdot\|_\infty$ denotes the sup-norm. It is easy to see that the squared Hellinger distance $d_H^2(f_{0,\sigma}, f_{r,\sigma}) = \int (\sqrt{f_{0,\sigma}} - \sqrt{f_{r,\sigma}})^2$ can be thus upper and lower bounded

$$d_H^2(f_{0,\tau}, f_{r,\tau}) \geq d_H^2(f_{0,\sigma}, f_{r,\sigma}) \geq C_B^2 \|r - r_0\|_2^2,$$

where $C_B^2 = (1 - \exp\{-B^2/2\sigma^2\})/2B^2$ and $\tau^2 = \sigma^2/2$. Let $K = C_B M$. For any $q \in \mathbb{Q}$, define $H_q^c = \{r \in \mathcal{R}_q : d_H(f_{0,\tau}, f_{r,\tau}) > K\epsilon_{n,p}\}$. Clearly, $\mathcal{R}_q \cap B(r_0; \epsilon_{n,p}) \subseteq H_q^c$. Also, note that $R_n(f_{r,\sigma}) = R_n^{1/2}(f_{r,\tau})$. Reasoning as in Walker and Hjort (2001), page 814,

$$\mathbb{E}_0^n[U_n] \leq \sum_{q \in \mathbb{Q}} w_q \mathbb{E}_0^n \left[\int_{H_q^c} R_n^{1/2}(f_{r,\tau}) d\Pi_q(r) \right] \leq \exp(-2nc_n), \quad (2)$$

with $c_n = K^2 \epsilon_{n,p}^2 / 4$. Let $z^\infty = (z_1, z_2, \dots) \in \mathbb{Z}^\infty$. Defined the set $\mathbb{Z}_1^{(n)} = \{z^\infty : U_n(z^\infty) \geq e^{-nc_n}\}$, by Markov's inequality and (2), $P_0^\infty(\mathbb{Z}_1^{(n)}) \leq \exp(-nc_n)$.

Clearly, $L_n \geq w_p \int_{\mathcal{R}_p} R_n(f_{r,\sigma}) d\Pi_p(r)$ because $w_q > 0$ for all $q \in \mathbb{Q}$. From Lemma 2 by Shen and Wasserman (2001), page 691, it is known that $P_0^\infty(\mathbb{Z}_2^{(n)}) \leq 2e^{-n\epsilon_{n,p}^2/2}$, with $\mathbb{Z}_2^{(n)} = \{z^\infty : (\int_{\mathcal{R}_p} R_n(f_{r,\sigma}) d\Pi_p(r))(z^\infty) \leq \frac{1}{2} \Pi_p(S_n) e^{-2n\epsilon_{n,p}^2}\}$. If (1) is satisfied (cfr. Scricciolo, 2005, page 416), by combining previous bounds, for constants $d, D > 0$,

$$\mathbb{E}_0^n \left[\frac{U_n}{L_n} \right] \leq e^{-nc_n} + \mathbb{E}_0^n \left[\frac{U_n}{L_n} \mathbb{I}_{(\mathbb{Z}_1^{(n)} \cup \mathbb{Z}_2^{(n)})^c} \right] + 2e^{-n\epsilon_{n,p}^2/2} \leq de^{-Dn\epsilon_{n,p}^2}.$$

For suitable $d_1, E, F > 0$, by Markov's inequality, $P_0^\infty(\{z^\infty : (U_n/L_n)(z^\infty) \geq e^{-En\epsilon_{n,p}^2}\}) \leq d_1 e^{-Fn\epsilon_{n,p}^2}$. The assertion follows from the first Borel-Cantelli lemma. $\square$

Corollary 1 Let $\hat{r}_n(\cdot) = \int r(\cdot) d\Pi(r | Z_1, \dots, Z_n)$ be the Bayes' regression estimator. Under conditions of Theorem 1,

$$\|\hat{r}_n - r_0\|_2^2 = O_P(\epsilon_{n,p}^2).$$

Proof. It suffices to appeal to Theorem 5 in Shen and Wasserman (2001), page 694. $\square$

A few remarks are in order. The proof of Theorem 1 avoids the need for a sieve as it is usually the case when studying rates of convergence for posterior distributions, see, e.g., Ghosal *et al.* (2000). The result is based on $\int_{H_q^c} R_n^{1/2}(f_{r,\tau}) d\Pi_q(r) \leq \exp(-nc_n)$ almost surely for all large n , which holds true in this set-up because the functions in $\mathcal{R}$ are uniformly bounded and the likelihood ratio is related to the pseudo-likelihood ratio.

Ellipsoidal classes in $L_2[0, 1]$. Now we consider a mixture of ‘‘sieve’’ priors and use previous results to establish adaptation for the Bayes' estimator. Let $\mathbb{X} = [0, 1]$ and

$f_X(\cdot) = \mathbb{I}_{[0,1]}(\cdot)$. Let $\{\phi_j(\cdot)\}_{j \geq 1}$ be an orthonormal system of $L_2[0, 1]$ such that for some constant $L > 0$, $\sup_{j \geq 1} \|\phi_j\|_\infty \leq L$. For any $r \in L_2[0, 1]$, let $r_j = \int_0^1 r(x) \phi_j(x) dx$, $j \geq 1$, be the sequence of its Fourier coefficients. For any $q \geq 1$ and fixed real $Q > 0$, let $\mathcal{R}_q(Q) = \{r \in L_2[0, 1] : r(\cdot) = \sum_{j=1}^\infty r_j \phi_j(\cdot), \sum_{j=1}^\infty j^{2q} r_j^2 < Q\}$. A prior on $\mathcal{R}_q(Q)$ is induced by a prior for the coefficients of r . We first put a prior μ_q on the space ℓ_2 of all square summable sequences and next truncate it to the ellipsoid $\mathcal{E}_q(Q) = \{(r_1, r_2, \dots) \in \ell_2 : \sum_{j=1}^\infty j^{2q} r_j^2 < Q\}$. Let $\Pi_q = \mu_q \mathbb{I}_{\mathcal{E}_q(Q)} / \mu_q(\mathcal{E}_q(Q))$ be the restricted prior. We use a mixture $\mu_q = \sum_{k=1}^\infty \lambda_k \mu_{k,q}$, where, for each $k \geq 1$, $\mu_{k,q}$ is such that the r_j 's are independent and $r_j \sim N(0, j^{-(2q+1)})$ for $j = 1, \dots, k$, whereas r_j is degenerate at zero for all $j > k$. For constants $A, \gamma > 0$, let $\lambda_k \geq A e^{-\gamma k}$, for $k \geq 1$, such that $\sum_{k=1}^\infty \lambda_k = 1$. Let $\Pi = \sum_{q=1}^\infty w_q \Pi_q$, with $w_q > 0$ for all $q \geq 1$ and $\sum_{q=1}^\infty w_q = 1$. Shen and Wasserman (2001) show that condition (1) is satisfied, so by Corollary 1, $\hat{r}_n$ is adaptive. We point out that this result concerns the *pointwise* behavior of the Bayes' estimator as opposed to the minimax behavior studied in Scricciolo (2005), wherein it is shown that for a known degree of smoothness, say p , the Bayes' estimator resulting from the posterior of Π_p attains the minimax rate under L_2 -loss over ellipsoidal classes.

3. Final Remarks

We have considered the problem of estimating the regression function in a random design Gaussian regression model from the Bayesian perspective. If the unknown regression function is assumed to be smooth, but no prior information on the smoothness degree is available, a hierarchical prior on the union of competing models for different degrees of smoothness leads to the same optimal convergence rate for the posterior as when only a single model is specified. Thus, the Bayes' estimator based on the resulting posterior automatically adapts to the unknown degree of smoothness. We remark that our approach relies on the assumption that the functions in the parameter space are uniformly bounded. It would thus be interesting to investigate the same problem removing this assumption.

References

- Belitser E., Ghosal, S. (2003) Adaptive Bayesian inference on the mean of an infinite-dimensional normal distribution, *The Annals of Statistics*, 31, 536–559.
- Ghosal S., Ghosh J. K., van der Vaart A. W. (2000) Convergence rates of posterior distributions, *The Annals of Statistics*, 28, 500–531.
- Ghosal S., Lember J., van der Vaart A. W. (2003) On Bayesian adaptation, *Acta Applicandae Mathematicae*, 79, 165–175.
- Huang T.-M. (2004) Convergence rates for posterior distributions and adaptive estimation, *The Annals of Statistics*, 32, 1556–1593.
- Scricciolo C. (2005) On Bayesian nonparametric regression function estimation, in: *S.Co. 2005, Modelli Complessi e Metodi Computazionali Intensivi per la Stima e la Previsione*, Cleup Editrice - Padova, 413–418.
- Shen X., Wasserman L. (2001) Rates of convergence of posterior distributions, *The Annals of Statistics*, 29, 687–714.
- Walker S., Hjort N. L. (2001) On Bayesian consistency, *Journal of the Royal Statistical Society*, 63, 811–821.