Statistica Sinica **8**(1998), 337-357

ON THE METHOD OF PENALIZATION

Xiaotong Shen

Ohio State University

Abstract: In this article, we study convergence properties of the method of penalization and related estimates. A penalized estimate is defined as an optimizer of a scaled criterion with a penalty that penalizes undesirable properties of the parameters. We develop some exponential probability bounds for the penalized likelihood ratios with a general penalty. Based on these inequalities, rates of convergence of the penalized estimates can be quantified. When convergence is measured by the Hellinger distance, the rate of convergence of the penalized maximum likelihood estimate depends only on the size of the parameter space and the penalization coefficient. We also explore the role of penalty in the penalization process, especially its relationship with the convergence properties and its connection with Bayesian analysis. We illustrate the theory by several examples.

Key words and phrases: Convergence properties, exponential bound, penalization, posterior distribution.

1. Introduction

A statistical procedure such as maximum likelihood method is often based on optimizing a criterion over a parameter space. When the parameter space is large, the optimization becomes difficult and the resulting estimates may have undesirable properties such as inconsistency and non-smoothness. To overcome these difficulties, the criterion is restricted with a penalty measuring such properties of the estimate. The optimization can then be carried out based on the penalized criterion. This procedure is called the method of penalization (see e.g., Wahba (1990) for references). In recent years, convergence properties of nonparametric and semi-parametric procedures has received considerable interest. Despite much research work on the method of penalization, there is no systematic study on the convergence properties of this method. In this article, to study the convergence properties of the method of penalization with a general penalty, we develop a general theory. We also investigate the role of the penalty in the penalization process and its connection with Bayesian analysis.

We now briefly review the most relevant literature. Convergence properties of the penalized estimates have been studied by many authors in specific models with variants of L_2-penalty (see e.g., Silverman (1982), Wahba (1990), Van De Geer (1990), Chen (1991), Gu and Qiu (1993) for references). To our knowledge,

the paper that gives a fairly general treatment for rates of convergence of the penalized estimates is Cox and O'Sullivan (1990). However, the existing results are restricted to the cases in which the parameter space is $W^{m,2}$ measured by L_2-smoothness, where $W^{m,2}$ is a Sobolev space and m is the parameter associated with the degree of smoothness of the functions (see Devore and Lorentz (1991) about Sobolev spaces). Still, it is not clear how the convergence properties of the method of penalization depends on the size of the parameter space and the penalty. Some recent developments on conditional quantile regression (see e.g., Koenker, Portnoy and Ng (1994)) suggested that some non-L_2 penalties also lead to solutions similar to the smoothing splines resulting from the L_2 penalty. Unfortunately, the convergence properties of the penalized estimates there are unavailable. Furthermore, Nemirovskii et al. (1985) showed that in non-parametric regression, linear estimates can not achieve the optimal rate of convergence when the regression function belongs to Sobolev spaces $W^{m,p}$ measured by L_p-smoothness for $1 \leq p < 2$. This means that in such cases, the penalized procedure with the L_2-penalty can not perform well in $W^{m,p}$ for $1 \leq p < 2$. The question is whether the penalized procedure can achieve the optimal rate of convergence. Moreover, what role does the penalty play in the penalization process? The investigation in this paper is expected to provide insights into the structure of the method of penalization and thus provide guidance for using this method in estimation, testing and discriminant analysis, etc.

To address the above issues, we develop some exponential bounds for the penalized likelihood ratios. Based on these inequalities, we establish a general theory on the convergence properties of the method of penalization. The theory relates the (local) size of the parameter space, and the magnitude of the penalization coefficient to the best possible rate of convergence. We show that the problem of penalization is essentially equivalent to a certain constrained optimization problem associated with the penalty. We also construct a prior which links the method of penalization to the convergence properties of the posterior distribution. To illustrate the theory, we examine a number of examples in non-parametric and semi-parametric models. In some of these examples, the obtained rates agree with the known optimal rates. In addition, we show that in density estimation and non-parametric regression, the optimal rate of convergence based on $W^{m,p}$ can be achieved with an L_p penalty for $p \geq 1$ (see Nemirovskii, Polyak and Tsybakov (1985)).

Let $Y_1 = (X_1, Z_1), \ldots, Y_n = (X_n, Z_n)$ be independently distributed according to density $p_i(\theta_0, y)$. We estimate $\theta \in \Theta$, where Θ is the parameter space. Let $l(\theta, Y_i)$ be the criterion function and $\tilde{l}(\theta, y) = l(\theta, y) - \lambda_n J(\theta)$ be the penalized criterion function, where $J(\theta)$ is a non-negative penalty and λ_n is the penalization

coefficient (the degree of penalization). The scaled penalized criterion to be optimized is defined as $\tilde{L}_n(\theta) = n^{-1}\sum_{i=1}^n \tilde{l}(\theta, Y_i)$. An approximate maximizer of the penalized criterion, denoted as $\hat{\theta}_n$, is called an approximate penalized estimate, in the sense that

$$\tilde{L}_n(\hat{\theta}_n) \geq \sup_{\theta\in\Theta} \tilde{L}_n(\theta) - a_n, \tag{1.1}$$

where $a_n \to 0$ as $n \to \infty$. For the exact estimate, $a_n = 0$. The procedure is called the penalized maximum likelihood (PML) estimation when l is a log-likelihood, and is the penalized least square regression when l is $-(y-\theta(x))^2$.

The paper is organized in six parts. Section 2 presents some examples and illustrative conclusions from the general theory. Section 3 discusses the convergence properties of the penalized maximum likelihood estimates (PMLEs) under the Hellinger distance and of the penalized estimates under the square root of the Kullback-Leiber information. Section 4 discusses the convergence properties of the penalty and its relationship with the convergence properties of the posterior distribution. Section 5 illustrates the main results by several examples: density estimation, the proportional odds model, non-parametric regression, and non-parametric conditional quantile regression. Section 6 is devoted to technical proofs.

2. Examples

In this section, we present some examples and convergence properties of the penalized estimates with penalty $J(\theta) = (\int_a^b |\theta^{(m)}(x)|^p dx)^{1/p}$ for some $p > 0$ and $0 < a < b$. The results for the penalized estimates with $J(\theta) = \int_a^b |\theta^{(m)}(x)|^p dx$ can also be obtained similarly.

Example 1. Density estimation

In this example, we study the rates of convergence of the PMLE and its derivatives in Hellinger distance. Although the problem of density estimation has been studied by many authors, we consider this problem in a general setting with L_p-measures of smoothness for $p \geq 1$. Let $Y_1, \ldots, Y_n$ be independently and identically distributed according to a density θ_0^2. We estimate $\theta \in \Theta$ using the method of penalization. In our formulation, the often assumed condition that the underlying density is uniformly bounded below is not used. To our knowledge, the result on the PMLE with L_p-smoothness measures and that on the fractional derivatives in Case 2 are not yet available in the existing literature. In addition, Case 3 provides an insight into the relationship between the rate of convergence of the penalized estimate and that of its derivative.

Case 1. For $m \geq 1$ and $min(p,2)m > d$, let $\Theta = W^{m,p}[0,1]^d$, where $W^{m,p}[0,1]^d$ is a Sobolev space with the degree of smoothness m measured by L_p, and $J(\theta) = \|\theta^{([m])}\|_p + [\iint_{[0,1]^d\times[0,1]^d}(|\theta^{([m])}(x) - \theta^{([m])}(y)|/|x-y|^{m-[m]})^p dxdy]^{1/p}$. Here $[m]$ is the integer part of m, $\|\cdot\|_p$ is the usual L_p norm. The Sobolev space $W^{m,p}[0,1]^d$ is defined as: $W^{m,p}[0,1]^d = \{\theta \in L_p[0,1]^d : \|\theta\|_{W^{m,p}} \leq \infty\}$, where $\|\theta\|_{W^{m,p}} = \sum_{k\leq[m]} \|\theta^{(k)}\|_p + (\iint_{[0,1]^d\times[0,1]^d} \frac{|\theta^{([m])}(x)-\theta^{([m])}(y)|^p}{|x-y|^{(m-[m])p}} dxdy)^{1/p}$.

Case 2. For $0 < m < 1$, let $\Theta = \{\theta \in C[0,1] : \theta \geq 0,\ \theta(0) = \theta(1) = 0, J(\theta) < \infty\}$ with $J(\theta) = \|\theta\|_H$, where $\|\theta\|_H = \sup_{x,y} |\theta(x)-\theta(y)|/|x-y|^m$ is the Hölder norm, and $C[0,1]$ is the space of all continuous functions.

Case 3. Now consider convergence properties of derivatives of the PMLE in terms of $\|\hat\theta_n^{(k)} - \theta_0^{(k)}\|_q$, where $q > 0$ is a real number and $0 \leq k < m$ satisfying: $(m-k)/2 + k/p \geq m/q$. In the following discussion, we will restrict our attention to the case in which $d = 1$ and m is an integer. The result can be extended to the case in which m is a fraction (see Adams (1975) for a definition of the norm with a fraction order).

Proposition 1. *In Case* 1, *let* $\eta_n = n^{-\frac{m}{2m+d}}$. *In Case* 2, *let* $\eta_n = n^{-\frac{m}{2m+1}}$ *when* $m > 1/2$; $\eta_n = n^{-1/4}(\log n)^{1/2}$ *when* $m = 1/2$; *and* $\eta_n = n^{-m/2}$ *when* $m < 1/2$. *In Case* 3, $\eta_n = n^{-\frac{m-k}{2m+1}\frac{m(m-k-1/p+1/q)}{(m-k)(m-1/p+1/2)}}$. *In Cases* 1-3, *let* $\lambda_n \sim \eta_n^2$. *Then, in Cases* 1-2, *we have* $P(\|\hat\theta_n - \theta_0\|_2 \geq \eta_n) \leq 7\exp(-c_5 n\eta_n^2)$. *In Case* 3, *we have* $P(\|\hat\theta_n^{(k)} - \theta_0^{(k)}\|_q \geq \eta_n) \leq 7\exp(-c_5 n\eta_n^2)$. *Here* $\hat\theta_n$ *is the penalized MLE, and* $c_5 > 0$ *is a constant.*

In Cases 1-2, the rates are optimal. In Case 3, note that $\frac{m(m-k-1/p+1/q)}{(m-k)(m-1/p+1/2)} \leq 1$. Clearly, when $p = q = 2$, the rate of convergence becomes $n^{-\frac{m-k}{2m+1}}$, which is believed to be optimal (see Stone (1982) for the case of non-parametric regression).

It is interesting to note in this case that although the space $W^{m,p}[0,1]$ is strictly larger than the space $C^m[0,1]$, the rate of convergence of the PMLE based on $W^{m,p}[0,1]$ is the same as that based on $C^m[0,1]$ since the effective sizes (metric entropy) which determine the rates are the same. In non-parametric regression, Nemirovskii, Polyak and Tsybakov (1985) showed that linear estimates can not achieve the optimal rate of convergence when the underlying function belongs to some Sobolev space $W^{m,p}[0,1]$ when $1 \leq p < 2$. This example shows that in density estimation, the optimal rate can be achieved by the PMLE with the L_p penalty for $p \geq 1$ and $m > 1/2$. This aspect may offer an insight on how to choose penalty in penalization estimation. (Also see Example 3.)

Example 2. Proportional odds model (Case 2 interval censoring)

In this example, we consider the convergence properties of the method of penalization in a semi-parametric model. In survival analysis, it is important

to study the relationship between a failure time $T \in (0, U]$ of an event and a covariate X. Suppose for the ith individual, $i = 1, \ldots, n$, there exist a failure time T_i and a p-dimensional covariate vector X_i. Under "Case 2" interval censoring, T_i is not fully observed. In this case, the observations consist of $Z_i = (Y_{i1}, Y_{i2}, \delta_i^{(1)}, \delta_i^{(2)}, X_i)$, where the exact time T_i is replaced by $(Y_{i1}, Y_{i2}, \delta_i^{(1)}, \delta_i^{(2)})$. Here $\delta_i^{(1)} = I(T_i \le Y_{i1})$ and $\delta_i^{(2)} = I(Y_{i1} < T_i \le Y_{i2})$ are indicators of whether the event T_i occurred before the monitoring time Y_{i1}, within the monitoring time interval $(Y_{i1}, Y_{i2}]$, or after the monitoring time Y_{i2}. Under the assumption of independent and identically distributed observations, the log-likelihood $L_n(Z, \beta, B)$ can be written as

$$n^{-1} \sum_{i=1}^{n} \Big[\delta_i^{(1)} \log F(Y_{i1}|X_i) + \delta_i^{(2)} \log(F(Y_{i2}|X_i) - F(Y_{i1}|X_i)) \\ +(1 - \delta_i^{(1)} - \delta_i^{(2)}) \log(1 - F(Y_{i2}|X_i)) \Big],$$

where $F(t|x) = P(T \le t|x) = \frac{\exp(B(t) - \beta^T x)}{1 + \exp(B(t) - \beta^T x)}$ is the probability of failure time given covariate values x and $B(t)$ is a baseline function. We estimate the regression parameter $\beta = (\beta_1, \ldots, \beta_p)$, where $B(t)$ is a nuisance parameter. Assume that β belongs to a compact set of R^p, $B(t) \in W^{m,q}[0, U]$ and $J(\theta) = (\int_0^U |B^{(m)}(t)|^p dt)^{1/p}$ for some $min(p, 2)m > 1$. In addition, assume that $F(Y_{\cdot 2}|X) - F(Y_{\cdot 1}|X) \ge c > 0$ for a small constant $c > 0$.

Proposition 2. *Under the assumptions, we have for the penalized estimate* $\hat{\theta}_n = (\hat{\beta}_n, \hat{B}_n)$ *with* $\lambda_n \sim n^{-\frac{2m}{2m+1}}$,

$$P\Big([\int (\hat{B}_n(y) - B_0(y))^2 dy]^{1/2} \ge \eta_n \Big) \le 7 \exp(-c_5 n \eta_n^2),$$

where $\eta_n = n^{-\frac{m}{2m+1}}$ *and* $c_5 > 0$ *is a constant.*

Example 3. Non-parametric regression

Consider the regression model $Y_i = \theta(x_i) + e_i$, where $\{x_i\}$ are fixed, and $\{e_i\}$ are independently identically distributed with mean 0 and a known variance σ^2. We assume that $E \exp(t_0|e_1|) < \infty$ for a constant $t_0 > 0$. Additionally, the smallest positive eigenvalue of matrix $n^{-1} H^T H$ is bounded below by a constant $g_0 > 0$, where

$$H = \begin{pmatrix} 1 & \cdots & x_1 & \cdots & x_1^m \\ 1 & \cdots & \cdots & \cdots & \cdots \\ 1 & \cdots & \cdots & \cdots & \cdots \\ 1 & \cdots & x_n & \cdots & x_n^m \end{pmatrix}$$

Case 1: Smooth function. Let $\theta \in \Theta = W^{m,p}[0,1]$, where $W^{m,p}[0,1]$ is a Sobolev space. We estimate θ using the least square criterion $l(\theta, y) = -(y-\theta)^2$ with a penalty $J(\theta) = [\int_0^1 |\theta^{(m)}(x)|^{(p)} dx]^{1/p}$ with $min(p,1)m > 1$ and $p \geq 1$.

Case 2: Monotone function. $\Theta = \{\theta \in C^1[0,1] : \theta(x)$ is monotone$\}$ and $J(\theta) = \int_0^1 (\theta^{(1)}(x))^2 dx$.

Proposition 3: *Let* $\rho(\theta_0, \theta) = (n^{-1}\sum_{i=1}^n (\theta(x_i) - \theta_0(x_i))^2)^{1/2}$. *In Case* 1, *let* $\eta_n = n^{-\frac{m}{2m+1}}$ *with* $\lambda_n \sim n^{-\frac{2m}{2m+1}}$. *In Case* 2, *let* $\eta_n = n^{-1/3}$ *with* $\lambda_n \sim n^{-2/3}$. *Then* $P(\rho(\theta_0, \hat{\theta}_n) \geq \eta_n) \leq 7\exp(-d_8 n\eta_n^2)$, *where* $\hat{\theta}_n$ *is the penalized estimate, and* $d_8 > 0$ *is a constant.*

In this case, the rate η_n is optimal (see Stone (1982)). Note that for the rate of convergence, the assumption that $E\exp(t_0|e_1|) < \infty$ of e_1 is not necessary. However, such an assumption yields the exponential probability bound. In fact, if the assumption on the moment of e_i is made, then the same result as above can be obtained (see Shen and Wong (1994)). In contrast to the result of Van De Geer (1990), our result yields an exponential bound and is valid for a more general penalization such as the L_p penalty for $p > 1$.

The rate of convergence of the penalized estimate with an L_2 penalty has been studied by many authors. As mentioned in the introduction, the result of Nemirovskii, Polyak and Tsybakov (1985) implies that the smoothing spline based on the L_2 penalty can not achieve the optimal rate of convergence in Sobolev spaces $W^{m,p}$ for $1 \leq p < 2$. The above result says that the penalized estimate can achieve this rate with a L_p penalty. Choosing an appropriate penalty is important in this situation.

Case 3: Posterior distribution. Let $\theta(x)$ be $\sum_{i=1}^\infty (\alpha_i cos(2\pi ix) + \beta_i sin(2\pi ix))$, where α_i and β_i are independently distributed as $N(0, (2\pi i)^{-2d})$ with $\sum_{i=0}^\infty i^{2(m-d)} < \infty$. Let $\breve{m}(\theta)$ be the probability measure induced by θ. Under this prior, by the three series theorem, the series of the squared random variables converges if and only if the series of its second moments converges. Consequently, the sample paths of θ have mth derivatives $\theta^{(m)}$ in L_2 if only if $\sum_{j=1}^\infty j^{2(m-d)} < \infty$ (see Adams (1975) for a definition of the derivative with a fraction order). The corresponding measure is a measure on Sobolev space $W^{2,m}[0,1]$ with the Sobolev inner product, see Kuo (1975) for a reference of measures on Banach spaces.

The following result then holds.

$$P(\rho(\theta_0, \theta) \geq \eta_n, \theta \in \Theta | Y_1, \ldots, Y_n) \leq d_{11}\exp(-O_p(n\eta_n^2)), \text{ in } P.$$

where $\eta_n = \max(n^{-\frac{m}{2m+1}}, n^{-\frac{n}{2d}}) = n^{-\frac{m}{2d}}$, and $d_{11} > 0$ is a constant. It appears that $n^{-\frac{m}{2d}}$ is slightly slower than $n^{-\frac{2m}{2m+1}}$ (optimal rate). We need to point out

that η_n is a rate for all $\theta_0 \in \Theta = W^{m,2}[0,1]$. Of course, the rates for some $\theta_0 \in \Theta$ are $n^{-\frac{m}{2m+1}}$ which is faster. For instance, the rate for $\theta_0 = 1$ is $n^{-\frac{m}{2m+1}}$ by Theorem 5 with $\tau = 1/(2d-1)$. Indeed, as shown by Shen (1994), $n^{-\frac{m}{2d}}$ is essentially the rate attainable by certain $\theta_0 \in \Theta$. This is because $\pi(\theta)$ assigns a small probability in any neighborhood of θ in the non-parametric setting. (See Example 3 and Shen (1994) for more discussions about this phenomenon.)

Example 4. Non-parametric estimation of conditional quantile

In this example, we answer the question raised in the introduction concerning the convergence properties of the penalized estimate in non-parametric quantile regression. Let $Y_i = \theta(X_i) + e_i,\ i = 1, \ldots, n$ where X_i can be fixed or random. Suppose the errors $\{e_i\}$ are independent of $\{X_i\}$ when $\{X_i\}$ are random, and e_i has a distribution with 0 as the τth quantile, i.e., $P(e_i \leq 0) = \tau$ for $0 < \tau < 1$. Furthermore, there exists $\delta^* > 0$ such that for any $0 < t \leq \delta^*$ and some constant $g_0 > 0$, $P(|e_i| \leq t) \geq g_0 t$. We estimate the conditional quantile of Y_i given X_i by maximizing $n^{-1}\sum_{i=1}^n l(y_i, \theta(x_i)) - \lambda_n J(\theta)$ over the parameter space $\Theta = W^{m,p}[a,b]$ (Sobolev space), where $l(y,\theta) = (I(y < \theta) - \tau)(y - \theta)$ (the Czech function, see e.g., Koenker, Portnoy and Ng (1994)), a and b are fixed, and $J(\theta) = \|\theta^{([m])}\|_p + [\int_a^b \int_a^b (|\theta^{([m])}(x) - \theta^{([m])}(y)|/|x-y|^{m-[m]})^p dxdy]^{1/p}$.

Proposition 4. *Let $\rho(\theta_0, \theta) = (E(\theta - \theta_0)^2)^{1/2}$. Under the assumptions, for the penalized estimate, we have $P(\rho(\theta_0, \hat{\theta}_n) \geq \eta_n) \leq 7\exp(-d_7 n\eta_n^2)$, where $\eta_n = n^{-\frac{m}{2m+1}}$ with $\lambda_n \sim n^{-\frac{2m}{2m+1}}$, and $d_7 > 0$ is a constant.*

3. Convergence Properties

In this section, we present some exponential inequalities for the supremum of the penalized likelihood ratios. The inequalities are developed by appropriately controlling the expectations, the variances of the log-likelihood ratios, and the penalty. Based on the inequalities, the consistency and the rate of convergence of the PMLEs can be established under simple conditions on the size of the parameter space.

Let P_i be a probability measure on a measurable space $\mathcal{Y}_i$ induced by the density $p_i(\theta_0, y)$. Define $P = n^{-1}\sum_{i=1}^n P_i$. Here and in the sequel, the expectation E and the variance Var, and the expectation E_i are evaluated under P and P_i, respectively.

3.1. Penalized MLEs

We first consider the case in which the criterion function is a likelihood and convergence is measured by the Hellinger distance. To quantify the size of a space, we briefly discuss the metric entropy. Suppose $f : \Theta \times \mathcal{Y}_i \to \mathcal{R}$ with

$E_i f^2(\theta, Y_i) < +\infty$ for all $\theta \in \Theta$. Let $\mathcal{F} = \{f(\theta, \cdot) : \theta \in \Theta\}$, and $S(\varepsilon, m) = \{f_1^l, f_1^u, \ldots, f_m^l, f_m^u\} \subset \mathcal{L}_2$ satisfying $\max_{1\le j\le m} \|f_j^u - f_j^l\|_2 \le \varepsilon$. If for any $f \in \mathcal{F}$, there exists a j such that $f_j^l \le f \le f_j^u$, *a.e.* P, then $H(\varepsilon, \mathcal{F}) = \log N(\varepsilon, \mathcal{F}) = \log(\min\{m : S(\varepsilon, m)\})$ is called the Hellinger metric entropy with bracketing when f is the square root density, and is the L_2 metric entropy with bracketing when f is replaced by a criterion function. (For more discussions about metric entropy of this type, see e.g., Kolmogorov and Tihomirov (1959), and Birman and Solomjak (1967).)

Let $h(\theta_0, \theta) = n^{-1}\sum_{i=1}^n [\int (p_i^{1/2}(\theta, y) - p_i^{1/2}(\theta_0, y))^2 dy]^{1/2}$ be the Hellinger distance. For any real number $k \ge 1$, let $\mathcal{F}_1(k) = \{p_i^{1/2}(\theta, \cdot) : \theta \in A(k)\}$ with $A(k) = \{\theta \in \Theta : J(\theta) \le k\}$.

Assumption A. There exist some constants $c_i > 0$ $(i = 1, 2)$ such that for $\varepsilon > 0$,

$$\sup_{\{k\ge 1\}} \psi_1(\varepsilon, k) \le c_2 n^{1/2}, \tag{3.1}$$

where $\psi_1(\varepsilon, k) = \int_L^{L^{1/2}} H^{1/2}(u, \mathcal{F}_1(k))du/L$ with $L = (c_1\varepsilon^2 + \lambda_n(k-1))$.

Theorem 1. *In addition to Assumption* A, *suppose* $\max(J(\theta_0), 1)\lambda_n \le c_3\varepsilon^2$ *for a constant* $0 < c_3 < 1/2$. *Then there exist* $c_i > 0$ $(i = 4, 5)$ *such that for any* $\varepsilon > 0$ *satisfying* (3.1),

$$P^*\Big(\sup_{\{h(\theta_0,\theta)\ge\varepsilon, \theta\in\Theta\}} \prod_{i=1}^n \frac{p(\theta, Y_i)\exp(-\lambda_n J(\theta))}{p(\theta_0, Y_i)\exp(-\lambda_n J(\theta_0))} \ge \exp(-c_4 n\varepsilon^2)\Big) \le 7\exp(-c_5 n\varepsilon^2),$$

where P^* is the outer measure (see Pollard (1984)).

Theorem 1 says that the probability of the supremum of the penalized likelihood ratios outside an ε Hellinger-neighborhood of θ_0 is exponentially small for any sample size n and $\varepsilon > 0$ satisfying (3.1) which is determined by the metric entropy equation. Such an inequality is useful, especially in obtaining rates of convergence of the penalized estimates and the posterior distribution (see Sections 3 and 4).

Corollary 1. *Suppose Assumption* A *holds. Then for the PMLE* $\hat\theta_n$ *defined in* (1.1) *with* $a_n = o(\varepsilon_n^2)$,

$$P\left(h(\theta_0, \hat\theta_n) \ge \eta_n\right) \le 7\exp(-c_5 n\eta_n^2), \tag{3.2}$$

where $\eta_n = \max(\varepsilon_n, \lambda_n^{1/2})$ *with* ε_n *the smallest* ε *satisfying* (3.1).

The best possible rate of convergence is governed by the smallest ε_n satisfying (3.1) with $\lambda_n \sim \varepsilon_n^2$. The trade-off phenomenon between the best rate of

convergence for the PMLE and the magnitude of the penalization coefficient λ_n can be considered as a generalization of the familiar bias/variance trade off in non-parametric regression and density estimation.

3.2. Penalized estimates

We generalize the results in Section 3.1 to a more general setting in which the criterion may not be a likelihood and the measure of convergence is defined by $\rho(\cdot,\cdot) = K^{1/2}(\cdot,\cdot)$ which is used for measuring distance between two parameter points, where $K(\theta_0,\theta) = E(l(\theta_0,Y) - l(\theta,Y))$. Here $K(\theta_0,\theta)$ is required to be non-negative. The quantity $K(\cdot,\cdot)$ is the Kullback-Leiber information when l is a log-likelihood. Note that the Kullback-Leiber information $K(\cdot,\cdot)$ usually dominates the commonly used measures such as the Hellinger distance. Therefore, the convergence under $K(\cdot,\cdot)$ is a strong mode.

We now introduce some notation. Let $V(\theta_0,\theta) = \mathrm{Var}\,(l(\theta,Y) - l(\theta_0,Y))$. For any $k_i > 0$, let $A(k_1,k_2) = \{\theta \in \Theta : k_1 \le \rho(\theta_0,\theta) \le 2k_1, J(\theta) \le k_2\}$ and $\mathcal{F}_2(k_1,k_2) = \{l(\theta,\cdot) - l(\theta_0,\cdot) : \theta \in A(k_1,k_2)\}$. In the following, $t_0 > 0$ and $d_i > 0$ are constants.

Assumption B. For some $0 \le \beta < 1$, $\sup_{A(k_1,k_2)} V(\theta_0,\theta) \le d_1 k_1^2[1 + (k_1^2 + k_2)^\beta]$.

Assumption C. There exists a random variable $W(Z_i)$ such that $|l(\theta,Y_i) - l(\theta_0,Y_i)| \le |\theta(X_i) - \theta_0(X_i)|W(Z_i)$, where $\{X_i\}$ and $\{Z_i\}$ are independent, $\sup_i E_i \exp(t_0 W(Z_i)) < \infty$ and $E(\theta(X) - \theta_0(X))^2 \le d_3 V(\theta_0,\theta)$. Additionally, $\sup_{A(k_1,k_2)} \|\theta - \theta_0\|_{\sup} \le d_2(k_1^2 + k_2)^\gamma$ for $0 \le \gamma < 1$.

Assumption D. Assume that

$$\sup_{\{k_1 \ge 1, k_2 \ge 1\}} \psi_2(k_1,k_2) \le d_6 n^{1/2}, \tag{3.3}$$

where $\psi_2(k_1,k_2) = \int_L^U H^{1/2}(u, \mathcal{F}_2(k_1,k_2))du/L$ with $U = d_4\varepsilon(k_1^2+k_2)^{(1+max(\beta,\gamma))/2}$ and $L = d_5\lambda_n(k_1^2 + k_2)$.

Theorem 2. *In addition to Assumptions* B-D, *suppose* $\max(J(\theta_0),1)\lambda_n \le d_7\varepsilon^2$. *Then there exists a constant* $d_8 > 0$ *such that for any* $\varepsilon > 0$ *satisfying* (3.3),

$$\begin{aligned} &P^*\Big(\sup_{\{\rho(\theta_0,\theta)\ge\varepsilon, \theta\in\Theta\}} n^{-1}\sum_{i=1}^n (\tilde{l}(\theta,Y_i) - \tilde{l}(\theta_0,Y_i)) \ge -\varepsilon^2/2\Big) \\ &\le 7\exp(-d_8 n \min(\lambda_n^2/\varepsilon^2, \lambda_n)). \end{aligned}$$

Corollary 2. *Suppose Assumptions* B-D *hold. Then for the penalized estimate defined in* (1.1) *with* $a_n = o(\varepsilon_n^2)$,

$$P(\rho(\theta_0,\hat{\theta}_n) \ge \eta_n) \le 7\exp(-d_8 n\eta_n^2), \tag{3.4}$$

where $\eta_n = \max(\varepsilon_n, \lambda_n^{1/2})$ *with* ε_n *the smallest* ε *satisfying* (3.3). *The best possible rate for the penalized estimate is governed by the smallest* ε_n *satisfying* (3.3) *with* $\lambda_n \sim \varepsilon_n^2$.

Assumption B specifies the local and global behaviors of the criterion difference. Locally, $\sup_{A(k_1,k_2)} V(\theta_0, \theta) \sim k_1^2$, whereas globally $\sup_{A(k_1,k_2)} V(\theta_0, \theta) \sim (k_1^2 + k_2)^\beta$ (see Examples 3 and 4). Assumption C is based on the moment generating function. Assumption D characterizes the size of the parameter space.

In the classical theory of maximum likelihood estimation, it is known that consistency and rate of convergence require quite different conditions on the likelihood function because the local behavior of the likelihood function differs from its global behavior (see Example 4). In Theorem 2, two issues have been dealt with simultaneously under the assumption of existence of the moment generating function as specified in Assumption C. Theorem 2 says that the rate of convergence η_n is essentially determined by the integral equation in (3.3) related to the local behavior of the likelihood function, and the consistency is mainly determined by the global behavior of the likelihood function. Although the assumption on the moment generating function is not necessary for obtaining rate of convergence, such an assumption yields an exponential bound. For the convergence properties, based on some moment conditions, an alternative condition can be established using a truncation argument (see Shen and Wong (1994)). The above results are valid as stated for any restricted parameter space.

4. Penalty and its Connection with Bayesian Analysis

4.1. Penalty

As discussed in the introduction, the penalty is used to penalize some undesirable properties of the estimate in a large parameter space which typically is not compact. We will show that the penalized estimate falls into a set $\{\theta \in \Theta : J(\theta) \le (1 + o(1))J(\theta_0)\}$. This means that the penalty function forces maximization to be carried out in a compact parameter space. Moreover, the penalization problem specified in (1.1) is essentially equivalent to a problem of the constrained optimization of the criterion (without penalty) over $\{\theta \in \Theta : J(\theta) \le J(\theta_0)\}$. Consequently, the penalized procedure may be viewed as an automatic procedure for estimating $J(\theta_0)$. (Also see Gu (1994).)

Theorem 3. *Under the assumptions in Theorem* 1, *for any* $0 < \delta < 1/4$, *if* $(1 - \delta)c_3\varepsilon_n^2 \le \max(J(\theta_0), 1)\lambda_n$, *then for the PMLE defined in* (1.1) *with* $a_n = o(\varepsilon_n^2)$, $P(J(\hat{\theta}_n) \ge (1 + O(\delta)) \max(J(\theta_0), 1)) \le 10 \exp(-c_5 n\varepsilon_n^2)$.

Theorem 3, in conjunction with the results in Theorem 2, can be used for obtaining the convergence properties of derivatives of the PMLEs when a

smoothness-related penalty is used. (See Example 1 for more detailed discussions.)

Theorem 4. *Under the assumptions in Theorem* 2, *for any* $0 < \delta < 1/4$, *if* $(1-\delta)d_6\varepsilon_n^2 \le \lambda_n$, *then for the penalized estimate defined in* (1.1) *with* $a_n = o(\varepsilon_n^2)$, $P(J(\hat{\theta}_n) \ge (1+O(\delta))\max(J(\theta_0),1)) \le 10\exp(-d_8 n\varepsilon_n^2)$.

4.2. Posterior distribution

In statistics, using penalty may be interpreted as formulating prior knowledge about the unknown parameters into the model (see e.g, Wahba (1990) and Cox (1993) for discussions). As discussed in Section 3, the penalty forces the maximizer being in a compact set of the parameter space. From the Bayesian prospective, this may be viewed as appropriately constructing a prior such that the posterior distribution is supported on a compact set of the parameter space with a very large probability.

We now formulate the problem from a Bayesian point of view. For simplicity, consider the case of independent and identically distributed observations. Let $\mathcal{Y}$ denote the sample space of a single observation Y, $\mathcal{B}$ be a Borel σ-field, and Θ be a subset of some separable Banach space. Let $\breve{m}(\theta)$ be a probability distribution on Θ (see Kuo (1975) for a reference of measures on Banach spaces). Now define the prior probability distribution $\pi(\theta \in A)$ as $\int_A \exp(-n\lambda_n J(\theta))d\breve{m}(\theta)$ for $A \in \mathcal{B}$, the Borel σ-field. The posterior probability of θ given $(Y_1,\ldots,Y_n)$, according to the Bayes rule, is

$$P(\theta \in A|Y_1,\ldots,Y_n) = \frac{\int_{\theta\in A}\prod_{i=1}^n p(\theta,Y_i)d\pi(\theta)}{\int\prod_{i=1}^n p(\theta,Y_i)d\pi(\theta)}. \tag{4.1}$$

The inference concerning θ can be made based on (4.1). We now formulate some conditions on $\breve{m}(\theta)$.

Assumption E. (Local behavior of prior) There exists constants $\tau > 0$ and $d_i > 0$ $(i = 9, 10)$ such that for any small $t > 0$, $\breve{m}(\rho(\theta_0,\theta) \le t, J(\theta) \le \max(J(\theta_0),1)), \theta \in \Theta) \ge d_9\exp(-d_{10}t^{-2\tau})$.

Theorem 5. *In addition to Assumptions* B-E, *suppose* $n\lambda_n \to \infty$. *Then there exists a constant* $d_{11} > 0$ *such that for* n *sufficiently large,* $P(\rho(\theta_0,\theta) \ge \eta_n, \theta \in \Theta|Y_1,\ldots,Y_n) \le d_{11}\exp(-O_p(n\eta_n^2))$, *in* P, *where* $\eta_n = \max(\varepsilon_n, \lambda_n^{1/2}, n^{-\frac{1}{2(1+\tau)}})$, *and* ε_n *is the smallest* ε *satisfying* (3.3).

Theorem 5 says that under the conditions that are used for formulating the convergence properties of the method of penalization as in Section 3, the posterior distribution is concentrated in an η_n-neighborhood of θ_0 with a large probability.

The rate η_n is governed by ε_n which is related to the penalized likelihood ratios and by $n^{-\frac{1}{2(1+\tau)}}$ which is related to the prior assignment in a neighborhood of θ_0. The prior probability $\pi(\theta)$ assigns a small probability to the parameters with large values of $J(\theta)$, which plays a role similar to that of the penalty in penalization.

5. Applications

In this section, we apply the general theory to obtain the results (Propositions 1-4) presented in Section 2.

Example 1. Density estimation

Case 1. We now verify Assumption A. By the norm equivalence Theorem, $H(u, \mathcal{F}_1(k)) \leq c(k/u)^{d/m}$ for some constant $c > 0$ (see e.g. Theorem 7.48 and Corollary 4.16, Adams (1975) and Theorem 5.2 of Birman and Solomjak (1967)). Note that the result in Kolmogorov and Tihomirov (1959) can not be used in this case. Assumption A is satisfied with $\psi_1(\varepsilon, k) = \varepsilon^{-\frac{2m+d}{2m}}(c_1 + c_3(k-1))^{-\frac{2m-d}{4m}}$, when $\max(J(\theta_0), 1)\lambda_n \leq c_3\varepsilon^2$.

Consequently, the rate of convergence of the PMLE under the Hellinger distance is $\eta_n = \max(\varepsilon_n, \lambda_n^{1/2})$, where $\varepsilon_n = n^{-\frac{m}{2m+d}}$ is the solution of the equation: $\psi_1(\varepsilon_n, 1) = c_2 n^{1/2}$. By Corollary 1, (2.2) holds with the best rate of convergence $\eta_n = n^{-\frac{m}{2m+d}}$ with $\lambda_n \sim \varepsilon_n^2 = n^{-\frac{2m}{2m+d}}$.

Case 2. By Lemma 7 of Shen and Wong (1994), $\sup_{\{\theta \in A(k)\}} \|\theta\|_{\sup} \leq c' k^{\frac{2m}{2m+1}}$ for some constant $c' > 0$. Hence for any small $u > 0$, $H(u, \mathcal{F}_1(k)) \leq \exp(c' \log(k^{\frac{2m}{2m+1}}/u) + c' \log((k/u)^{1/m})) \leq c'(k/u)^{1/m}$ (see the proof of Theorem 15 of Kolmogorov and Tihomirov (1959) for the corresponding constants with the sup-entropy). When $\max(J(\theta_0), 1)\lambda_n \leq c_3\varepsilon^2$, Assumption A is satisfied with $\psi_1(\varepsilon, k) = \varepsilon^{-\frac{2m+1}{2m}}(c_1 + c_3(k-1))^{-\frac{2m-1}{4m}}$ for $m > 1/2$, $\psi_1(\varepsilon, k) = \varepsilon^{-1}[\log(\varepsilon(c_1 + c_3(k-1))]^{-1}$ for $m = 1/2$, and $\psi_1(\varepsilon, k) = \varepsilon^{-\frac{1}{m}}$ for $m < 1/2$. By Corollary 1, (2.2) holds with $\eta_n = n^{-\frac{m}{2m+1}}$ when $m > 1/2$, with $\eta_n = n^{-1/4}(\log n)^{1/2}$ when $m = 1/2$, and with $\eta_n = n^{-m/2}$ when $m < 1/2$. Here $\lambda_n \sim \eta_n^2$.

Case 3. By the triangular inequality, $J(\hat{\theta}_n - \theta_0) \leq J(\hat{\theta}_n) + J(\theta_0)$. Thus, by Theorem 3 and Lemma 2, $\|\hat{\theta}_n^{(k)} - \theta_0^{(k)}\|_q \leq [J(\hat{\theta}_n - \theta_0)]^{(1-r)}\|\hat{\theta}_n - \theta_0\|_2^r$, where $r = \frac{m-k-1/p+1/q}{m-1/p+1/2}$ and $(m-k)/2 + k/p \geq m/q$. Consequently, $P(\|\hat{\theta}_n^{(k)} - \theta_0^{(k)}\|_q \geq \eta_n) \leq 7\exp(-c_5 n\eta_n^2)$, where $\eta_n = n^{-\frac{m-k}{2m+1}\frac{m(m-k-1/p+1/q)}{(m-k)(m-1/p+1/2)}}$.

Example 2. Proportional odds model (Case 2 interval censoring)

We now verify Assumption A. Here the density is

$$p(\theta, y_1, y_2) = F(\theta, y_1|x_i)^{\delta^{(1)}}(F(\theta, y_2|x) - F(\theta, y_1|x))^{\delta^{(2)}}(1 - F(\theta, y_2|x))^{(1-\delta^{(1)}-\delta^{(2)})},$$

where $\theta = (B, \beta)$. Note that

$$\begin{aligned}
&|p^{1/2}(\theta_1, y_1, y_2) - p^{1/2}(\theta_2, y_1, y_2)| \\
&\le c(|F^{1/2}(\theta_1, y_1|x) - F^{1/2}(\theta_2, y_1|x)| + |F^{1/2}(\theta_1, y_2|x) - F^{1/2}(\theta_2, y_2|x)|) \\
&\quad + |(1 - F(\theta_1, y_1|x))^{1/2} - (1 - F(\theta_2, y_1|x))^{1/2}| + |(1 - F(\theta_1, y_1|x))^{1/2} \\
&\quad - (1 - F(\theta_2, y_1|x)|))^{1/2}|) \\
&\le c(|B_1(y_1) - B_2(y_1)| + |B_1(y_2) - B_2(y_2)| + |(\beta_1 - \beta_2)^T x|).
\end{aligned}$$

In the above calculations, the fact that $|q(x)| \le 1$ has been used, where $q(x) = 1/(1 + \exp(x))^{1/2}$ for $x \ge 0$. By the norm equivalence Theorem, $H(u, \mathcal{F}_1(k)) \le c[(k/u)^{1/m} + p \log(1/u)]$ for a constant $c > 0$. When $\max(J(\theta_0), 1)\lambda_n \le c_3\varepsilon^2$, with $\psi_1(\varepsilon, k) = \varepsilon^{-\frac{2m+1}{2m}}(c_1 + c_3(k-1))^{-\frac{2m-1}{4m}}$, Assumption A is satisfied. Consequently, the rate of convergence of the PMLE under the Hellinger distance is $\max(\varepsilon_n, \lambda_n^{1/2})$, where $\varepsilon_n = n^{-\frac{m}{2m+1}}$ is the solution of the equation: $\psi_1(\varepsilon_n, 1) = c_2 n^{1/2}$. By Corollary 1, $P(h(\theta_0, \hat{\theta}_n) \ge \eta_n) \le 7\exp(-c_5 n\eta_n^2)$ with the best rate of convergence $\eta_n = n^{-\frac{m}{2m+1}}$ when $\lambda_n \sim n^{-\frac{2m}{2m+1}}$. This implies that $\max(\int (F^{1/2}(\hat{\theta}_n, y|x) - F^{1/2}(\theta_0, y|x))^2 dy, \int ((1 - F(\hat{\theta}_n, y|x))^{1/2} - (1 - F(\theta_0, y|x))^{1/2})^2 dy)$ is bounded by η_n^2 in probability. After some calculations, we conclude that

$$P\Big([\int ((\hat{B}_n(y) - B_0(y)) + (\hat{\beta}_n^T - \beta^T)x)^2 dy]^{1/2} \ge \eta_n\Big) \le 7\exp(-c_5 n\eta_n^2),$$

which implies that $P([\int (\hat{B}_n(y) - B_0(y))^2 dy]^{1/2} \ge \eta_n) \le 7\exp(-c_5 n\eta_n^2)$.

Example 3. Non-parametric regression

Case 1: Smooth function. Note that

$$l(\theta(x_i), y_i) - l(\theta_0(x_i), y_i) - E_i[l(\theta(x_i), Y_i) - l(\theta_0(x_i), Y_i)] = 2(y_i - \theta_0)(\theta(x_i) - \theta_0(x_i)),$$

$K(\theta_0, \theta) = \rho^2(\theta_0, \theta)$, and $V(\theta_0, \theta) = \sigma^2\rho^2(\theta_0, \theta)$. By Lemma 2, $\sup_{A(k_1,k_2)} \|\theta - \theta_0\|_{\sup} \le k_1^r k_2^{1-r}$, where $\frac{m-1/p}{m-1/p+1/2}$. Note that $k_1^r k_2^{1-r} \le c(k_1^2 + k_2)^{1-r/2}$ for any $k_i > 0$. Hence, Assumptions B and C are satisfied with $\beta = 0$, and $\gamma = 1 - r/2$. We now verify Assumption D. Without loss of generality, assume that $\theta_0 = 0$ in the following. From approximation theory, we know that for any $\theta \in A(k_1, k_2) \subset W^{m,p}$, there exists a polynomial $h \in A(k_1, k_2)$ with degree $n-1$ that interpolates $(x_1, \ldots, x_n)$ such that $\|\theta\|_{\sup} \le \|h\|_{\sup} + \|\theta - h\|_{\sup} \le \|h\|_{\sup} + k_2$. Since the form of h is available depending on the representation of $\theta \in \Theta$. It follows after some calculations that $\|h\|_{\sup} \le g_1\rho(\theta_0, h)/g_0$. Hence, by Theorem 5.2 of Birman and Solomjak (1967), for any small $u > 0$,

$$H(u, \mathcal{F}_2(k_1, k_2)) \le H(u/g_0, A(k_1, k_2), \|\cdot\|_{\sup}) \le c' u^{-1/m}[k_1 + k_2]^{1/m},$$

where $H(u, A(k_1,k_2), \|\cdot\|_{\sup})$ is the metric entropy with sup-norm. Therefore, Assumption D is satisfied with $\psi_2(k_1,k_2) = \frac{\varepsilon^{(1-1/2m)}}{\lambda_n} \frac{(k_1^2+k_2)^{(1+\gamma)(1-1/2m)/2}(k_1+k_2)^{1/2m}}{k_1^2+k_2}$. By Corollary 2, (3.4) holds, i.e., $P(\rho(\theta_0,\hat{\theta}_n) \geq \eta_n) \leq 7\exp(-d_8 n\eta_n^2)$, where $\eta_n = \max(\varepsilon_n, \lambda_n^{1/2}) = n^{-\frac{m}{2m+1}}$ with the penalization coefficient $\lambda_n \sim n^{-\frac{2m}{2m+1}}$.

Case 2: Monotone function. Assumptions B-D can be verified similarly. By Corollary 2, (3.2) holds with $\eta_n = \max(\varepsilon_n, \lambda_n^{1/2})$, where ε_n is the solution of the equation: $\int_{d_5\varepsilon^2}^{d_4\varepsilon} u^{-1/2}du = d_6 n^{1/2}$. Therefore, by Corollary 2, (3.4) holds with $\eta_n = n^{-1/3}$ and the penalization coefficient $\lambda_n \sim n^{-2/3}$.

Case 3: Posterior distribution. To obtain the rate of convergence for the posterior distribution, we only need to verify Assumption E. By Lemma 2 of Shen (1994), $\tau = \max((d-m)/m, 1/(2d-1))$. By Theorem 5, $\eta_n = (n^{-\frac{m}{2m+1}}, n^{-\frac{m}{2d}})$ with $\lambda_n \sim n^{-\frac{2m}{2m+1}}$.

Example 4. Non-parametric estimation of conditional quantile

We now verify Assumptions B-D. Note that

$$\begin{aligned} K(\theta_0,\theta) &= 2E(I(\theta-\theta_0 \geq 0)\int_0^{|\theta-\theta_0|}(\theta-\theta_0-y)dF(y) \\ &\quad + I(\theta-\theta_0 \leq 0)\int_{-|\theta-\theta_0|}^{0}(y+\theta-\theta_0)dF(y)) \\ &= 2\int_0^{|\theta-\theta_0|} P(|e_i| \leq y)dy, \end{aligned}$$

and $V(\theta_0,\theta) \leq E(\theta-\theta_0)^2$. By Theorem 2 of Gabushin (1967) and the argument in Example 3, we have $\sup_{A(k_1,k_2)} \|\theta-\theta_0\|_{\sup} \leq k_1^r k_2^{1-r} \leq c(k_1^2+k_2)^{1-r/2}$, where $r = \frac{m-1/p}{m-1/p+1/2}$. It can be seen that $K(\theta_0,\theta) \geq c'((k_1^2+k_2)^{1-r/2})^{-1}V(\theta_0,\theta)$ when $\theta \in A(k_1,k_2)$. Assumptions B and C are satisfied with $\beta = \gamma = 1-r/2$ and $W(y) = 1$. Finally, by the norm equivalence Theorem (Adams (1975)) and Theorem 5.2 of Birman and Solomjak (1967), $H(u, \mathcal{F}_2(k_1,k_2)) \leq c'' u^{-1/m}(k_1+k_2)^{1/m}$ for a constant $c'' > 0$. Hence, Assumption D is satisfied with $\psi_2(k_1,k_2) = \frac{\varepsilon^{(1-1/2m)}}{\lambda_n} \frac{(k_1^2+k_2)^{(1+\beta)(1-1/2m)/2}(k_1+k_2)^{1/2m}}{k_1^2+k_2}$. Consequently, by Corollary 2, $P(\rho(\theta_0,\hat{\theta}_n) \geq \eta_n) \leq 7\exp(-d_7 n\eta_n^2)$, where $\eta_n = \max(\varepsilon_n, \lambda_n^{1/2})$, and where ε_n is the solution of the equation: $\psi_2(\varepsilon_n, 1) = d_6 n^{1/2}$. The best rate is $\eta_n = n^{-\frac{m}{2m+1}}$ with $\lambda_n \sim n^{-\frac{2m}{2m+1}}$.

The convergence results for the penalized estimate with $J(\theta)=\int |\theta^{([m])}(x)|^p dx$ for $p \geq 1$ can also be obtained as above.

6. Appendix

Before proceeding, we introduce some notation to be used in the following

proofs. Let

$$\nu_n(\tilde{l}(\theta,Y)-\tilde{l}(\theta_0,Y)) = n^{-1}\sum_{i=1}^{n}(\tilde{l}(\theta,Y_i)-\tilde{l}(\theta_0,Y_i)-E(\tilde{l}(\theta,Y_i)-\tilde{l}(\theta_0,Y_i))) = \nu_n(l(\theta,Y)-l(\theta_0,Y)).$$

For $i=1,2,\ldots$, $j=0,1,\ldots$, let

$$A_{i,j} = \{\theta\in\Theta : 2^{i-1}\varepsilon \le h(\theta_0,\theta) < 2^i\varepsilon, 2^{j-1}\max(J(\theta_0),1) \le J(\theta) < 2^j\max(J(\theta_0),1)\}.$$

For $i=1,2,\ldots$, let $A_{i,0}=\{\theta\in\Theta : 2^{i-1}\varepsilon \le h(\theta_0,\theta) < 2^i\varepsilon, J(\theta) < \max(J(\theta_0),1)\}$. In the proof of Theorem 2, $h(\theta_0,\theta)$ will be replaced by $\rho(\theta_0,\theta)$.

Proof of Theorem 1. The proof relies heavily on a large deviation inequality of Wong and Shen (1995) and the left-truncation argument for the log-likelihood ratios. Without loss of generality, we assume that $J(\theta_0)\ge 1$. For any $0\le\tau<\infty$, let $p^{(\tau)}(\theta,y)$ be the left truncation version of $p(\theta,y)$, i.e.,

$$p^{(\tau)}(\theta,y) = \begin{cases} \exp(-\tau)p(\theta_0,y), & \text{if } p(\theta,y) < \exp(-\tau)p(\theta_0,y),\\ p(\theta,y), & \text{otherwise.}\end{cases}$$

It can be easily seen that the results in Wong and Shen (1995) developed under independently and identically distributed observations can generally apply to the case of independently but non-identically distributed observations if the corresponding quantities there are replaced by the average quantities on the basis of each observation.

We first control means of the truncated log-likelihood ratios. By Lemma 4 of Wong and Shen (1995), we have $-n^{-1}\sum_{i=1}^{n}E\log(p^{(\tau)}(\theta,Y_i)/p(\theta_0,Y_i)) \ge (1-T)h^2(\theta_0,\theta)$, where $T=2\exp(-\tau/2)/(1-\exp(-\tau/2))^2$. Now choose τ such that $1-T-c_4=c_1>0$. Hence for any $i,j\ge 1$,

$$\inf_{A_{i,j}}[(1-T)h^2(\theta_0,\theta)+\lambda_n(J(\theta)-J(\theta_0))-c_4\varepsilon^2] \ge M(i,j),$$

where $M(i,j)=c_1(2^{i-1}\varepsilon)^2+\lambda_n(2^{j-1}-1)J(\theta_0)$. Note that $\max(J(\theta_0),1)\lambda_n \le c_3\varepsilon^2$. Then for any $i\ge 1$, $\inf_{A_{i,0}}[(1-T)h^2(\theta_0,\theta)+\lambda_n(J(\theta)-J(\theta_0))-c_4\varepsilon^2] \ge M(i)$, where $M(i)=c_1[(2^{i-1}\varepsilon)^2-c_3\varepsilon^2]$. Consequently,

$$I = P^*\Big(\sup_{\{h(\theta_0,\theta)\ge\varepsilon,\theta\in\Theta\}}\prod_{i=1}^{n}[p(\theta,Y_i)\exp(-\lambda_nJ(\theta))]/[p(\theta_0,Y_i)\exp(-\lambda_nJ(\theta_0))] \ge \exp(-c_4n\varepsilon^2)\Big)$$

$$\le P^*\Big(\sup_{\{h(\theta_0,\theta)\ge\varepsilon,\theta\in\Theta\}} n^{-1}\sum_{i=1}^{n}\log(p^{(\tau)}(\theta,Y_i)/p(\theta_0,Y_i)) \ge \lambda_n(J(\theta)-J(\theta_0))-c_4\varepsilon^2\Big)$$
$$\le \sum_{i,j=1}^{\infty} P^*\Big(\sup_{A_{i,j}}\nu_n(\log(p^{(\tau)}(\theta,Y)/p(\theta_0,Y))) \ge M(i,j)\Big)$$
$$+\sum_{i=1}^{\infty} P^*\Big(\sup_{A_{i,0}}\nu_n(\log(p^{(\tau)}(\theta,Y)/p(\theta_0,Y))) \ge M(i)\Big)$$
$$= I_1 + I_2.$$

We proceed to bound I_1 and I_2 separately. Here I_1 and I_2 are probabilities related to the global and the local behaviors of the penalized log-likelihood ratios. The rate of convergence of the PMLE is essentially determined by I_2.

We now control the variances of the truncated likelihood ratios. By Lemma 3 of Wong and Shen (1995), we obtain

$$\sup_{A_{i,j}}\mathrm{Var}\,(\log p^{(\tau)}(\theta,Y)/p(\theta_0,Y)) \le v^2(i,j)=4\exp(\tau)[(2^i\varepsilon)^2+\frac{2}{c_1}\lambda_n(2^{j-1}-1)J(\theta_0)].$$

We now verify the required conditions in Lemma 7 of Wong and Shen (1995). Condition (3.3) in that lemma follows from the fact that $M(i,j)/v^2(i,j) \le \frac{c_1}{4\exp(\tau)}$. In addition, it is easy to see that

$$\int_{aM(i,j)}^{v(i,j)} H^{1/2}(u,\mathcal{F}_1(j))du/M(i,j) \le \int_{aM(1,j)}^{v(1,j)} H^{1/2}(u,\mathcal{F}_1(j))du/M(1,j).$$

Thus Assumption A implies (3.4) in that lemma. Therefore there exists $c_5 > 0$ such that

$$I_1 \le \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} 3\exp(-c_5 n[c_1(2^{i-1}\varepsilon)^2 + (2^{j-1}-1)\lambda_n J(\theta_0)])$$
$$\le 3\exp(-c_5 n\varepsilon^2)/(1-\exp(-c_5 n\varepsilon^2)).$$

Similarly, I_2 can be bounded using an argument similar to that for I_1.

Finally, $I \le 6\exp(-c_5 n\varepsilon^2)/(1-\exp(-c_5 n\varepsilon^2))$. Therefore, $I \le (6+I)\exp(-c_5 n\varepsilon^2)$. The result then follows from the fact $I \le 1$.

Proof of Corollary 1. By (1.1), for any $\varepsilon_n > 0$ satisfying (3.1), there exists $c_4 > 0$ such that

$$P\Big(h(\theta_0,\hat\theta_n) \ge \varepsilon_n\Big) \le P^*\Big(\sup_{\{h(\theta_0,\theta)\ge\varepsilon_n,\theta\in\Theta\}}(\tilde L_n(\theta)-\tilde L_n(\theta_0)) \ge -a_n\Big)$$
$$\le P^*\Big(\sup_{\{h(\theta_0,\theta)\ge\varepsilon_n,\theta\in\Theta\}}(\tilde L_n(\theta)-\tilde L_n(\theta_0)) \ge -c_4\varepsilon_n^2\Big).$$

By Theorem 1, $h(\theta_0, \hat{\theta}_n) = O_p(\varepsilon_n)$ when $\max(J(\theta_0), 1)\lambda_n \le c_3\varepsilon_n^2$. Note that ε_n is the smallest ε satisfying (3.2). Hence $h(\theta_0, \hat{\theta}_n) = O_p(\lambda_n^{1/2})$ if ε_n is replaced by $\lambda_n^{1/2}$ when $\max(J(\theta_0), 1)\lambda_n > c_3\varepsilon_n^2$. The result then follows.

Lemma 1. *Suppose Assumption* C *holds. Let* $v^2 \ge \sup_{\theta\in\mathcal{F}} n^{-1}\sum_{i=1}^n V(\theta_0, \theta)$ *and* $B \ge \sup_{\theta\in\mathcal{F}} \|\theta - \theta_0\|_{sup}$. *Suppose further*

$$\int_L^U H^{1/2}(u, \mathcal{F})du \le n^{1/2}Ma^{3/2}/2^{10}, \tag{6.1}$$

where $U = H^-(\Psi(M, v), \mathcal{F})$ *and* $L = aM/2^8$ *(*$0 < a < 1$*). Let* $\Psi(M, v) = (1 - a)nM^2/[2(v^2 + BM/3)]$. *Then*

$$P^*\Big(\sup_{\theta\in\mathcal{F}} \nu_n(l(\theta, Y) - l(\theta_0, Y)) \ge M\Big) \le 3\exp(-\Psi(M, v)). \tag{6.2}$$

Inequality (6.2) *continues to hold for* $U \le L$ *with* 3 *replaced by* 1 *in* (6.2).

Remark. $\Psi(\cdot)$ satisfies the following inequality

$$\Psi(M, v) \ge \begin{cases} (1 - a)nM^2/4v^2 & \text{if} \quad MB/v^2 \le 3 \\ 3(1 - a)nM/4B & \text{if} \quad MB/v^2 > 3, \end{cases}$$

which will be used in the proof below.

Proof. The rest of proof follows the arguments similar to those in Theorem 3 of Shen and Wong (1994).

Proof of Theorem 2. The basic idea of the proof is similar to that in Theorem 1. However, the control of the means and variances of criterion differences is more complicated. Without loss of generality, we assume $\max(\lambda_n, \varepsilon) \le 1$. For any $i, j \ge 1$,

$$\inf_{A_{i,j}}[K(\theta_0, \theta) + \lambda_n(J(\theta) - J(\theta_0))] \ge (2^{i-1}\varepsilon)^2 + \lambda_n(2^{j-1} - 1)J(\theta_0),$$

and

$$\inf_{A_{i,0}}[K(\theta_0, \theta) + \lambda_n(J(\theta) - J(\theta_0))] \ge (2^{i-1}\varepsilon)^2 - \lambda_n J(\theta_0).$$

Since $\max(J(\theta_0), 1)\lambda_n \le d_7\varepsilon^2$, we have

$$\begin{aligned} I &= P^*\Big(\sup_{\{\rho(\theta_0,\theta)\ge\varepsilon, \theta\in\Theta\}} n^{-1}\sum_{i=1}^n(\tilde{l}(\theta, Y_i) - \tilde{l}(\theta_0, Y_i)) \ge -\varepsilon^2/2\Big) \\ &= \sum_{i,j=1}^{\infty} P^*(\sup_{A_{i,j}} \nu_n(l(\theta, Y) - l(\theta_0, Y)) \ge M(i, j)) \\ &\quad + \sum_{i=1}^{\infty} P^*(\sup_{A_{i,0}} \nu_n(l(\theta, Y) - l(\theta_0, Y)) \ge M(i)) \\ &= I_1 + I_2, \end{aligned}$$

where $M(i,j) = \frac{1}{2}\lambda_n[(2^{i-1})^2 + (2^{j-1}-1)J(\theta_0)]$ and $M(i) = d'(2^{i-1}\varepsilon)^2$.

To bound I_1, we verify the required conditions in Lemma 1. By Assumption B,

$$\sup_{A_{i,j}} V(\theta_0,\theta) \le v^2(i,j) = d_1(2^i\varepsilon)^2(1+((2^i)^2+2^jJ(\theta_0))^\beta).$$

When $MB/v^2 \le 3$, $\Psi(M,v) \ge (1-a)nM^2/4v^2$. It is easy to see that $U = H^-(\Psi(M,v),\mathcal{F}) \le v(i,j)$ Similarly, when $MB/v^2 > 3$, $U \le M^{1/2}(i,j)B^{1/2}(i,j)$. By Assumption D

$$\int_{aM(i,j)}^{\max(v(i,j),M^{1/2}(i,j)B^{1/2}(i,j))} H^{1/2}(u,\mathcal{F}_2(2^i\varepsilon,2^j))du/M(i,j) \le d_5n^{1/2}.$$

Hence (6.1) holds. By Lemma 1, we have

$$\begin{aligned} I_1 &\le 3\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\exp(-d_8n\min(M^2(i,j)/v^2(i,j), M(i,j)/B(i,j))) \\ &\le 3\sum_{i=1}^{\infty}\sum_{j=1}^{\infty}\exp(-d_8n\min((\lambda_n^2/\varepsilon^2)[(2^{i-1})^2+2^{j-1})]^{1-\beta}, \lambda_n[(2^{i-1})^2+2^{j-1})]^{1-\gamma})) \\ &\le 3\exp(-d_8n\min(\lambda_n^2/\varepsilon^2,\lambda_n))/[1-\exp(-d_8n\min(\lambda_n^2/\varepsilon^2,\lambda_n))], \end{aligned}$$

where d_8 may differ in each step. In the above derivation, the inequality $(a+b)^\beta \le a^\beta + b^\beta$ for $a,b>0$ and $0<\beta<1$ has been applied, and I_2 can be bounded with an argument similar to that for I_1. Finally, we have

$$\begin{aligned} I &\le 6\exp(-d_8n\min(\lambda_n^2/\varepsilon^2,\lambda_n))/[1-\exp(-d_8n\min(\lambda_n^2/\varepsilon^2,\lambda_n))] \\ &\le 7\exp(-d_8n\min(\lambda_n^2/\varepsilon^2,\lambda_n)). \end{aligned}$$

This completes the proof.

Proof of Corollary 2. Same arguments as in Corollary 1.

Proof of Theorem 3. The proof is essentially the same as that in Theorem 4 and thus is omitted.

Proof of Theorem 4. Without loss of generality, we assume that $J(\theta_0) \ge 1$. Otherwise, we replace it by 1. For $j = 1,\ldots,$ let $A_j = \{\theta \in \Theta : \rho(\theta_0,\theta) \le \varepsilon_n, 2^{j-1}J(\theta_0) \le J(\theta) < 2^jJ(\theta_0)\}$ and $A_0 = \{\theta\in\Theta : \rho(\theta_0,\theta)\le\varepsilon_n, J(\theta) < J(\theta_0)\}$. By (1.1),

$$\begin{aligned} &P\Big(J(\hat\theta_n) \ge J(\theta_0)\frac{\lambda_n+\delta\varepsilon_n^2}{\lambda_n-\delta\varepsilon_n^2}\Big) \\ &\le P(\lambda_n(J(\hat\theta_n)-J(\theta_0)) \ge \delta\varepsilon_n^2(J(\hat\theta_n)+J(\theta_0))) \end{aligned}$$

$$\begin{aligned}
&\le P(\nu_n(\tilde{l}(\hat{\theta}_n, Y) - \tilde{l}(\theta_0, Y))/(J(\hat{\theta}_n) + J(\theta_0)) \ge \delta\varepsilon_n^2 - a_n, \rho(\theta_0, \hat{\theta}_n) \le \varepsilon_n) \\
&\quad + P(\rho(\theta_0, \hat{\theta}_n) \ge \varepsilon_n) \\
&\le \sum_{j=0}^{\infty} P^*(\sup_{A_j} \nu_n(l(\theta, Y) - l(\theta_0, Y)) \ge \delta J(\theta_0) 2^{j-1}\varepsilon_n^2 - a_n) + P(\rho(\theta_0, \hat{\theta}_n) \ge \varepsilon_n) \\
&\le I + P(\rho(\theta_0, \hat{\theta}_n) \ge \varepsilon_n).
\end{aligned}$$

To bound I, let $M(j) = \delta J(\theta_0) 2^{j-1}\varepsilon_n^2(1 + o(1))$ and $v^2(j) = d_2(2^i\varepsilon_n)^2(1 + ((2^i)^2 + 2^j)^\beta))$. Then $M(j)/v^2(j) \le a$ for some $a > 0$. Furthermore, (6.1) is implied by Assumptions C and D. Consequently, applying Lemma 1, we obtain

$$I \le 3\sum_{j=0}^{\infty} \exp(-\Psi(M(j), v(j))) \le 3\exp(-d_9 n\varepsilon_n^2)/(1 - \exp(-d_9 n\varepsilon_n^2)).$$

By Theorem 2, with $\lambda_n J(\theta_0) \le d_7\varepsilon_n^2$, $P(\rho(\theta_0, \hat{\theta}_n) \ge \varepsilon_n) \le 7\exp(-d_9 n\varepsilon_n^2)$. The result then follows.

Proof of Theorem 5. From (3.1), we have

$$P(\rho(\theta_0, \theta) \ge \eta_n, \theta \in \Theta | Y_1, \ldots, Y_n) = \frac{\int_{\{\rho(\theta_0,\theta) \ge \eta_n, \theta \in \Theta\}} \exp(L_n(\theta) - L_n(\theta_0)) d\pi(\theta)}{\int \exp(L_n(\theta) - L_n(\theta_0)) d\pi(\theta)}.$$

We proceed to bound the numerator and the denominator of the last expression separately.

By the classical central limit theorem, for a fixed small $\delta > 0$

$$[(L_n(\eta) - L_n(\theta)) - E_\theta(L_n(\eta) - L_n(\theta))]/(\delta_n V^{1/2}(\theta, \eta)) \to Z > Z - \delta,$$

where Z is $N(0, 1)$ and $\delta_n = n^{-1/2}$. Let $S(t) = \{\theta \in \Theta : \rho(\theta_0, \theta) \le t, J(\theta) \le \max(J(\theta_0), 1)\}$. By Skorohod's representation theorem, Fatou's Lemma, the dominated convergence theorem and Assumption B, we have for some small $t_0 > 0$,

$$\begin{aligned}
&\liminf_{n\to\infty} \int \exp\left(n(L_n(\eta) - L_n(\theta))\right) d\pi(\eta) \\
&\ge \int \liminf_{n\to\infty} \exp(-n\delta_n V^{1/2}(\theta, \eta)(Z - \delta)) \exp(-nK(\theta, \eta)) d\pi(\eta) \\
&\ge \liminf_{n\to\infty} \exp(-n(\delta_n(t_0\eta_n) + (t_0\eta_n)^2))(|Z| + \delta)\pi(\eta \in S(t_0\eta_n)) \\
&\ge \liminf_{n\to\infty} \exp(-n(\delta_n(t_0\eta_n) + (t_0\eta_n)^2))(|Z| + \delta) \\
&\qquad\qquad \exp(-n\lambda_n \max(J(\theta_0), 1) - d_{10}(t_0\eta_n)^{-2\tau}).
\end{aligned}$$

By Assumption E,

$$
\begin{aligned}
I &= P\Big(\int_{\{\rho(\theta_0,\theta)\geq\eta_n,\theta\in\Theta\}} \exp(n(L_n(\theta)-L_n(\theta_0)))d\pi(\theta) \geq \exp(-n\eta_n^2/2)\Big) \\
&\leq \sum_{i,j=1}^{\infty} P^*\Big(\sup_{\{\theta\in A_{i,j}\}} \nu_n(l(\theta,Y)-l(\theta_0,Y)) \geq -\log\pi(A_{i,j})/n - \eta_n^2/2\Big) \\
&\leq \sum_{i,j=1}^{\infty} P^*\Big(\sup_{\{\theta\in A_{i,j}\}} \nu_n(l(\theta,Y)-l(\theta_0,Y)) \geq M(i,j)\Big),
\end{aligned}
$$

where $M(i,j) = d_3((2^{i-1}\eta_n)^2 + 2^{j-1}\lambda_n)$. Then $\sup_{A_{i,j}} V(\theta_0,\theta) \leq v^2(i,j) = d_1(2^i\eta_n)^2(1+((2^i\eta_n)^2+2^jJ(\theta_0))^\beta)$. Applying an argument as in the proof of Theorem 2, $I \leq 7\exp(-d_8 n\min(\lambda_n^2/\eta_n,\lambda_n))$. Consequently, by appropriately choosing t_0, we obtain $P(\rho(\theta_0,\theta)\geq\eta_n,\theta\in\Theta|Y_1,\ldots,Y_n) \leq d_{11}\exp(-O_p(n\eta_n^2))$. The result then follows.

Lemma 2. *Let $S = \{f\in W^{m,p}[a,b] : \|f\|_2 < L_1, \|f^{(m)}\|_p < L_2\}$, where a and b are fixed constants. Then $\|f^{(k)}\|_q \leq 2\|f\|_2^{\Delta}L_3^{1-\Delta}$, where $\Delta = \frac{m-k-1/p+1/q}{m-1/p+1/2}$, $(m-k)/2+k/p \geq m/q$, and $L_3>0$ depends on $b-a$ and L_i $(i=1,2)$.*

Proof. For any $f\in S$, by Theorem 1 of Gabushin (1967), we have $\|f^{(k)}\|_q \leq A(\delta^{-k-1/2+1/q}\|f\|_2 + \delta^{m-k-1/p+1/q}L_2)$, where $0<\delta\leq b-a$ and A is a positive constant. The result then follows by choosing L_3 large enough such that $\delta = (\|f\|_q/L_2)^{1/(m-1/p+1/q)} \leq (L_1/L_3)^{1/(m-1/p+1/q)} \leq b-a$. This completes the proof.

Acknowledgements

The author would like to thank the associate editor and the referees for helpful comments and suggestions. The research of the author is partly supported by the seeds grant of the research foundation at the Ohio State University.

References

Adams, R. A. (1975). *Sobolev Spaces.* Academic press, New York.

Birman, M. S. and Solomjak, M. Z. (1967). Piecewise-polynomial approximation of functions of the classes W_p. *Mat. Sb.* **73**, 295-317.

Chen, Z. (1991). Interaction spline models and their convergence rates. *Ann. Statist.* **19**, 1855-1868.

Cox, D. D. and O'Sullivan, F. (1990). Asymptotic analysis of penalized likelihood and related estimators. *Ann. Statist.* **18**, 1676-1695.

Cox, D. D. (1993). An analysis of Bayesian inference for non-parametric regression. *Ann. Statist.* **21**, 903-924.

Devore. R. and Lorentz, G. (1991). *Constructive Approximation.* Springer-Verlag.

Gabushin, V. N. (1967). Inequalities for norms of functions and their derivatives in the L_p metric. *Mat. Zametki* **1**, 291-298.

Gu, C. and Qiu, C. (1993). Smoothing spline density estimation: Theory. *Ann. Statist.* **21**, 217-234.

Gu, C. (1994). Model indexing and model selection in non-parametric function estimation. Purdue University, Technical report.

Koenker, R., Portnoy, S. and Ng, P. (1994). Quantile smoothing splines. *Biometrika* **81**, 673-680.

Kolmogorov, A. N. and Tihomirov, V. M. (1959). ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Mat. Nauk.* **14**, 3-86. (In Russian. English translation, *Ameri. Math. Soc. Transl.* **2**, **17**, 277-364. (1961))

Kuo, H. H. (1975). *Gaussian Measures on Banach Spaces.* Lecture notes in Math. **463**. Springer, Berlin.

Nemirovskii, A. S., Polyak, B. T. and Tsybakov, A. B. (1985). Rate of convergence of non-parametric estimates of maximum likelihood type. *Problems Inform. Transmission* **21**, 258-272.

Pollard, D. (1984). *Convergence of Stochastic Processes.* Springer-Verlag, New York.

Schoenberg, I. J. (1964). Spline functions and the problem of graduation. *Proc. Nat. Acad. Sci.* **52**, 947-950.

Shen, X. and Wong, W. H. (1994). Convergence rate of sieve estimates. *Ann. Statist.* **22**, 580-615.

Shen, X. (1994). On the properties of Bayes procedures in general parameter spaces. The Ohio State University, Department of Statistics, Technical Report No. 539.

Silverman, B. (1982). On the estimation of a probability function by the maximum penalized likelihood method. *Ann. Statist.* **10**, 795-810.

Stone, C. (1982). Optimal global rates of convergence for non-parametric regression. *Ann. Statist.* **10**, 1040-1053.

Van De Geer, S. (1990). Estimating a regression function. *Ann. Statist.* **18**, 907-924.

Wahba, G. (1990). *Spline Models for Observational Data,* CBMS-NSF Regional Conference Series, Philadelphia.

Wong, W. H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLEs. *Ann. Statist.* **23**, 339-362.

Department of Statistics, The Ohio State University, 1958 Neil Avenue, Columbus, OH 43210-1247, U.S.A.

E-mail: xshen@stat.ohio-state.edu

(Received March 1996; accepted June 1997)