Journal of the Korean
Statistical Society
Vol. 23, No. 2, 1994

Asymptotic Theory for Multi-Dimensional Mode Estimator[†]

JeanKyung Kim[1]

ABSTRACT

In this paper we extend Kim and Pollard's cube root asymptotics to other rates of convergence, to establish an asymptotic theory for a multidimensional mode estimator based on uniform kernel with shrinking bandwidths. We obtain rates of convergence depending on shrinking rates of bandwidth and non-normal limit distributions. Optimal decreasing rates of bandwidth are discussed.

KEYWORDS: Empirical process, Shrinking bandwidth, Kernel mode estimators, Functional central limit theorem.

1. INTRODUCTION

Chernoff(1964) suggested a mode estimator of a distribution on the real line based on a uniform kernel $K_a(\cdot,\theta) = \frac{1}{2a}[\theta - a, \theta + a]$. (We adopt set notation for indicator functions and use linear functional notation for expectation.) He showed that θ_n maximizing $\frac{1}{n}\sum_{i=1}^{n} K_a(\xi_i,\theta)$, ξ_i's are data points, had a

[†] This research was supported by Inha University grant 1992–1993.

[1] Department of Statistics, Inha University, Incheon, 402–751, Korea.

$n^{-1/3}$ rate of convergence and a limit distribution determined by a functional on a gaussian process, using one dimensional techniques based on a Markov property and semi-martingale inequalities. He also considered the case where the bandwidth, a_n, decreased with sample size, showing the estimator had rate of convergence depending on a_n.

Using empirical process technique, Kim and Pollard (1990) established an asymptotic theory for multidimensional estimators defined, in a similar way to Chernoff's estimator, by maximization of stochastic processes based on empirical processes. As an example they showed that a multidimensional analogue of Chernoff's mode estimator had a $n^{-1/3}$ rate of convergence and non-normal limit distribution. However, since they considered only the case where $O_p(n^{-1/3})$ asymptotics obtained, we could not apply their method for the problems like mode estimators based on kernels with decreasing bandwidth, where the estimators had other rates of convergence.

In this paper, with a slight modification of Kim and Pollard's argument, we establish an asymptotic theory for multidimensional mode estimators based on uniform kernels with shrinking bandwidths. As a possible 'uniform kernel' function in $\Re^d$, we adopt the indicator function of a ball with decreasing volume α_n in examples through this paper. Any convex set of fixed shape may be used instead of a ball, however, the symmetry property of a ball make it simpler to use.

Let $\{\xi_i\}$ be a sequence of independent observations from a fixed distribution P with a unique mode θ_0 on $\Re^d$. For a decreasing sequence of real numbers $\{\alpha_n\}$, define a kernel mode estimator $\hat{\theta}_n$ as a maximizing value of

$$\frac{1}{cn\alpha_n}\sum_{i=1}^{n} K\Big(\frac{\xi_i-\theta}{\alpha_n^{1/d}}\Big), \tag{1.1}$$

where the function K is an indicator function of a ball with radius 1 centered at origin (the ball can be replaced by any fixed convex set) in $\Re^d$, and c is a constant making sure that $\int \frac{1}{c}K = 1$.

Let θ_n be a maximizing value of the expected value of (1.1). Then $(\hat{\theta}_n - \theta_0)$ can be decomposed into two parts: $(\hat{\theta}_n - \theta_n)$ which is a probabilistic part and contributes to the variance of the mean square error of $\hat{\theta}_n$; $(\theta_n - \theta_0)$ which is a non-random, bias part. The next two sections give limit theorems for $(\hat{\theta}_n - \theta_n)$ using inequalities already available in the empirical process literature. We also discuss on optimal rates of α_n. In Section 4 we examine two two-dimensional mode estimators. One is a symmetric case where $\theta_n = \theta_0$, using the standard bivariate normal as a underlying distribution. The other is an asymmetric case where we should also deal with the non-random part carefully.

2. RATE OF CONVERGENCE

In this section we establish consistency and a rate of convergence result for $\hat{\theta}_n$, which are defined by the location of the maximizing value of stochastic processes derived from empirical processes. As an appropriate space for the sample paths of the stochastic processes, we define $\mathcal{X}$ to be the set of all locally bounded functions on $\Re^d$ equiped with the topology of uniform convergence on compacta. Then $\hat{\theta}_n$ can be represented by means of the *argmax* functional on $\mathcal{X}$, which assigns the location of the maximizing value of each element of $\mathcal{X}$. However, there are some difficulties involved in the definition of the *argmax* functional for processes whose paths do not achieve their supremum, or do achieve a maximum at several points in $\Re^d$. As illustrated in Kim and Pollard(1990), we avoid them by proving limit theorems for random elements of $\Re^d$ that come close enough to maximizing processes with paths in $\mathcal{X}$.

Define the *empirical measure* P_n as the random probability measure that puts a mass $\frac{1}{n}$ on each of $\xi_1, \ldots, \xi_n$ from P; define the *empirical process* as the signed measure $\nu_n = \sqrt{n}(P_n - P)$.

Let $f_n(\cdot, \theta) = K\Big(\frac{\cdot - \theta}{\alpha_n^{1/d}}\Big)$, then $\hat{\theta}_n$ maximazes $P_n f_n(\cdot, \theta) = \frac{1}{n} \sum f_n(\xi_i, \theta)$ and θ_n maximizes $P f_n(\cdot, \theta) = \int f_n(\cdot, \theta) dP$. To prove $(\hat{\theta}_n - \theta_n)$ converge to 0 we need to show that $P_n f_n(\cdot, \theta)$ is close to $P f_n(\cdot, \theta)$ uniformly on θ. Define a class of functions indexed by a subset Θ of $\Re^d$,

$$\mathcal{F}_n = \Big(f_n(\cdot, \theta) : \theta \in \Theta\Big). \tag{2.1}$$

For consistency of $\hat{\theta}_n$ we need the following Lemma (a modified version of Theorem II.37 in Pollard(1984)) which gives a uniform bound for $P_n f_n(\cdot, \theta) - P f_n(\cdot, \theta)$ over Θ.

Lemma 2.1. For each n, let $\mathcal{F}_n$ be a VC subgraph class of bounded functions. Let $\{\epsilon_n\}$ be a non-increasing sequence of positive numbers for which $\alpha_n \epsilon_n^2 \gg n^{-1} \log n$. If $P f_n^2 \leq \alpha_n$ for each f_n in $\mathcal{F}_n$ then

$$\sup_{\mathcal{F}_n} \Big| P_n f_n - P f_n \Big| \leq \epsilon_n \alpha_n \quad \text{almost surely.}$$

The symbol $\gg$ means that the left hand side is of bigger order than the right hand side. That is, $a_n \gg b_n$ implies $\frac{b_n}{a_n} = o(1)$.

The VC subgraph class is one of several VC related classes defined by Dudley (1987) as variations of the VC class of sets introduced in Vapnik and Cervonenkis (1971). Since the definition of VC subgraph class (see Dudley(1987)) involves some combinatorial condition which is not quite important here, we omit them. Instead, we give several examples and state some useful properties of a VC subgraph class, which will be important for our proofs.

One simple example of VC subgraph class is a class of the indicator functions of balls in $\Re^d$. More generally, a class of indicator functions of sets of any VC class (class of cylinders, class of convex sets with same shape and so forth) is a VC subgraphclass. Also, any subset of a finite-dimensional vector space of real functions on a set X is a VC subgraph class.

For a class $\mathcal{F}$ of functions, define the (natural) envelope F as $F = \sup_{\mathcal{F}} |f|$. A VC subgraph class has the following properties:

- If $\mathcal{F}$ is a VC subgraph class, then so is $\Big\{|f| : f \in \mathcal{F}\Big\}$ with the same enveolpe.

- If $\mathcal{F}$, with envelope F, is VC subgraph class then so is $\Big\{f_1 - f_2 : f_i \in \mathcal{F}\Big\}$ with envelope $2F$.

The following Theorem proves a stronger one than consistency. To understand the reason why we do it this way, consider a simple case where f_n is an indicator funciton of a ball in $\Re^d$ with radius $\alpha_n^{1/d}$. Since f_n shrinks with diameter of order $\alpha_n^{1/d}$, it is natural to expect that $|\hat{\theta}_n - \theta_n|$ tends to stay within $\alpha_n^{1/d}$ neighborhood of origin. Let $\delta_n = \alpha_n^{1/d}$.

Theorem 2.2 (Consistency). For each n, let $\mathcal{F}_n$ defined (2.1) be a VC subgraph class of indicator functions. If there exists a non-increasing sequence $\{\epsilon_n\}$, for which $\alpha_n \epsilon_n^2 \gg n^{-1} \log n$, such that

$$Pf_n(\cdot, \theta_n) \; - \sup_{|\theta - \theta_n| > \delta_n} Pf_n(\cdot, \theta) \; > \; 2\epsilon_n \alpha_n, \quad \text{eventually}, \tag{2.2}$$

where $\delta_n = \alpha_n^{1/d}$, then

$$|\hat{\theta}_n - \theta_n| \; \leq \; \delta_n \quad \text{almost surely.}$$

Proof. Lemma 2.1 gives that

$$\sup_{\mathcal{F}_n} \Big| P_n f_n - P f_n \Big| \leq \epsilon_n \alpha_n \quad \text{almost surely.}$$

Together with (2.2) this implies that with probability one it is eventually true that

$$\begin{aligned} P_n f_n(\cdot, \hat{\theta}_n) \; &\geq \; P_n f_n(\cdot, \theta_n) \\ &\geq \; P f_n(\cdot, \theta_n) - \epsilon_n \alpha_n \\ &> \; \sup_{|\theta - \theta_n| > \delta_n} P f_n(\cdot, \theta) + \epsilon_n \alpha_n \\ &\geq \; \sup_{|\theta - \theta_n| > \delta_n} P_n f_n(\cdot, \theta). \end{aligned}$$

It follows that with probability one,

$$|\hat{\theta}_n - \theta_n| \; \leq \; \delta_n, \quad \text{eventually.}$$

Before we give a rigorous proof of the rate of convergence, it is quite instructive to start with a heuristic argument that suggests the rate of convergence of $\hat{\theta}_n$.

Let $Y_n(\theta) = P_n f_n(\cdot,\theta) - P_n f_n(\cdot,\theta_n)$, then $\hat{\theta}_n$ maximizes Y_n. To find the rate of convergence we first sketch the limit behavior of Y_n. Decompose $Y_n(\theta)$ into two parts, a deterministic trend and a random perturbation, as follows:

$$P\Big(f_n(\cdot,\theta) - f_n(\cdot,\theta_n)\Big) + (P_n - P)\Big(f_n(\cdot,\theta) - f_n(\cdot,\theta_n)\Big). \tag{2.3}$$

Under a smoothness assumption on P that $P\Big(f_n(\cdot,\theta) - f_n(\cdot,\theta_n)\Big)$ is twice differentiable at θ_n and the second deriavative is approximately linear in α_n. For some positive constant c_1, the trend of Y_n, the first part of (2.3), is approximately

$$-c_1\alpha_n|\theta - \theta_n|^2.$$

For each θ near θ_n, the central limit theorem gives that the perturbation of Y_n, the second part of (2.3) is approximately normal distribution with mean zero and variance $n^{-1}P\Big(f_n(\cdot,\theta) - f_n(\cdot,\theta_n)\Big)^2$. The function $\Big(f_n(\cdot,\theta) - f_n(\cdot,\theta_n)\Big)^2$ is the indicator function of symmetric difference of two balls. When θ is close to θ_n, the volume of the symmetric difference can be obtained by a surface integration around the boundary of a ball with volume α_n. Since the magnitude of the boundary of the ball is of order $\alpha_n^{(d-1)/d}$, the volume of the symmetric difference is roughly $O\Big(\alpha_n^{(d-1)/d}|\theta - \theta_n|\Big)$. Let $\beta_n = \alpha_n^{(d-1)/d}$, then $P\Big(f_n(\cdot,\theta) - f_n(\cdot,\theta_n)\Big)^2 = O\Big(\beta_n|\theta - \theta_n|\Big)$. So $Y_n(\theta)$ has a mean $-c_1\alpha_n|\theta-\theta_n|^2$, a parabolar maximized at θ_n, which is randomly perturbed by a process with a standard deviation $n^{-1/2}\beta_n^{1/2}|\theta - \theta_n|^{1/2}$. When $|\theta - \theta_n|$ gets big, the downward tendency of the parabola overwhelms the random perturbation. The maximum of Y_n should occur when two terms in (2.3) are of about the same order. That is, when $\alpha_n|\theta - \theta_n|^2$ is of the same order as $n^{-1/2}\beta_n^{1/2}|\theta - \theta_n|^{1/2}$. This implies that $|\theta - \theta_n|$ is of order $n^{-1/3}\alpha_n^{-2/3}\beta_n^{1/3}$.

Put $\gamma_n = n^{-1/3}\alpha_n^{-2/3}\beta_n^{1/3} = n^{-1/3}\alpha_n^{-(d+1)/3d}$. To prove γ_n rate of convergence, we need to give a uniform bound on the perturbation part of (2.3).

The following two maximal inequalities (Maximal Inequality 3.1 of Kim and Pollard) provide useful bounds for a stochastic process $(P_n - P)f$ indexed by a class of real-valued function $\mathcal{F}$. More general forms and their proofs can be found on Pollard(1989).

Maximal Inequalities 2.3. Let $\mathcal{F}$ be a VC subgraph class of functions with an envelope F, for which $PF^2 < \infty$. Suppose zero function is included in $\mathcal{F}$. Then there exists a function J, not depending on n, such that

$$\text{(i)}\ \sqrt{n}\mathbb{E}\sup_{\mathcal{F}}\Big|P_n f - Pf\Big| \leq \mathbb{E}\sqrt{P_n F^2} J\Big(\sup_{\mathcal{F}} P_n f^2/P_n F^2\Big) \leq J(1)\sqrt{PF^2},$$

$$\text{(ii)}\ n\mathbb{E}\sup_{\mathcal{F}}\Big|P_n f - Pf\Big|^2 \leq \mathbb{E}P_n F^2 J^2\Big(\sup_{\mathcal{F}} P_n f^2/P_n F^2\Big) \leq J(1)^2 PF^2.$$

The function J is continuous and increasing, with $J(0) = 0$ and $J(1) < \infty$.

Note that for a VC subgragh class $\mathcal{F}$, the function J is the same for every subclass of $\mathcal{F}$.

Let $g_n(\cdot,\theta) = f_n(\cdot,\theta) - f_n(\cdot,\theta_n)$, and $\mathcal{G}_n = \{g_n(\cdot,\theta) : \theta \in \Theta\}$. Using the Maximal Inequalities 2.3, outside of γ_n neighborhood of θ_n we establish a uniform bound on the $(P_n - P)g_n(\cdot,\theta)$. For each R, we define a subclass $\mathcal{G}_n(R)$ of $\mathcal{G}_n$ to be $\Big\{g_n(\cdot,\theta) : \theta \in \Theta, |\theta - \theta_n| < R\Big\}$, and its envelope $G_n(R) = \sup_{\mathcal{G}_n(R)} |g_n|$. For each R, assume that $G_n(R)$ is measurable.

Lemma 2.4. Suppose for each n, $\Big\{\mathcal{G}_n(R) : R \leq \delta_n\Big\}$ is a family of VC subgraph classes with their envelopes $G_n(R)$, for which

$$PG_n(R)^2 \leq CR\beta_n, \quad \text{for all } R \leq \delta_n$$

for a finite constant C and some decreasing sequences β_n and δ_n. Then for each $\epsilon > 0$ there exist random numbers $\{M_n\}$ of order $O_p(1)$ such that

$$\Big|P_n g_n(\cdot,\theta) - Pg_n(\cdot,\theta)\Big| \leq \epsilon\alpha_n|\theta - \theta_n|^2 + \alpha_n\gamma_n^2 M_n^2, \quad \text{for } |\theta - \theta_n| \leq \delta_n, \quad (2.4)$$

where $\gamma_n = n^{-1/3}\alpha_n^{-2/3}\beta_n^{1/3}$.

Proof. Define $M_n(\omega)$ as the infimum of those values for which the asserted uniform inequality holds. Define $A(n,j)$ to be the set of those θ in Θ for which $(j-1)\gamma_n \le |\theta - \theta_n| < j\gamma_n$. Then for constant m, the probability $\mathbb{P}\{M_n > m\}$ is less than or equal to

$$\mathbb{P}\Big\{\exists\theta : \Big|P_n g_n(\cdot,\theta) - P g_n(\cdot,\theta)\Big| > \epsilon\alpha_n|\theta-\theta_n|^2 + \alpha_n\gamma_n^2 m^2\Big\}$$

$$\le \sum_{j=1}^{\delta_n/\gamma_n} \mathbb{P}\Big\{\exists\theta \in A(n,j) : (\alpha_n\gamma_n^2)^{-1}\Big|P_n g_n(\cdot,\theta) - P g_n(\cdot,\theta)\Big| > \epsilon(j-1)^2 + m^2\Big\}.$$

To have the rate of convergence make some sense we may assume that $\delta_n/\gamma_n \to \infty$. The j-th summand is bounded by

$$(\alpha_n\gamma_n^2)^{-2}\mathbb{E}\sup_{\theta\in A(n,j)}\Big|P_n g_n(\cdot,\theta) - P g_n(\cdot,\theta)\Big|^2\Big/\Big(\epsilon(j-1)^2 + m^2\Big)^2.$$

By the part (ii) of the Maximal Inequalities 2.3 and the assumption about $PG_n(R)^2$, there is a finite constant C_1 such that the numerator of the last expression is less than $(\alpha_n\gamma_n^2)^{-2}n^{-1}C_1 j\gamma_n\beta_n = C_1 j$. We can therefore ensure that the sum is suitably small for all n by choosing m large enough.

With $\delta_n = \alpha_n^{1/d}$ and $\beta_n = \alpha_n^{(d-1)/d}$ in Lemma 2.4, the following Theorem gives $n^{-1/3}\alpha_n^{-(d+1)/3d}$ rate of convergence for $(\hat\theta_n - \theta_n)$.

Theorem 2.5 (Rate of convergence). If $Pg_n(\cdot,\theta)$ has a negative definite second derivative matrix, $-V_n$, at θ_n for which, $\lim_{n\to\infty}\alpha_n^{-1}V_n = V$, and if

(i) $\hat\theta_n = \theta_n + O_p(\delta_n)$,

(ii) $P_n g_n(\cdot,\hat\theta_n) \ge \sup_\theta P_n g_n(\cdot,\theta) - O_p(\alpha_n\gamma_n^2)$,

then under the condition of the Lemma 2.4,

$$\hat\theta_n = \theta_n + O_p(\gamma_n),$$

where $\delta_n = \alpha_n^{1/d}$, $\beta_n = \alpha_n^{(d-1)/d}$ and $\gamma_n = n^{-1/3}\alpha_n^{-2/3}\beta_n^{1/3} = n^{-1/3}\alpha_n^{-(d+1)/3d}$.

Proof. Choose ϵ so that $Pg_n(\cdot,\theta) \leq -2\epsilon\alpha_n|\theta-\theta_n|^2$ in a neighborhood of θ_n for which the assertion of Lemma 2.4 holds. Since $\hat{\theta}_n = \theta_n + O_p(\delta_n)$, by Lemma 2.4,

$$P_n g_n(\cdot,\hat{\theta}_n) \leq Pg_n(\cdot,\hat{\theta}_n) + \epsilon\alpha_n|\hat{\theta}_n-\theta_n|^2 + \alpha_n\gamma_n^2 M_n^2.$$

Condition (ii) implies that the left-hand side is bigger than $P_n g_n(\cdot,\theta_n) - O_p(\alpha_n\gamma_n^2) = -O_p(\alpha_n\gamma_n^2)$, the bound on $Pg_n(\cdot,\theta)$ forces

$$\epsilon\alpha_n|\hat{\theta}_n-\theta_n|^2 \;=\; O_p(\alpha_n\gamma_n^2),$$

from which the asserted rate of convergence follows.

3. WEAK CONVERGENCE

Results from Section 2 established the $O_p(\gamma_n) = O_p(n^{-1/3}\alpha_n^{-(d+1)/3d})$ rate of convergence for $(\hat{\theta}_n - \theta_n)$. The limit behavior of $\gamma_n(\hat{\theta}_n-\theta_n)$ will be deduced in this section, by an application of a slightly modified continuous mapping theorem for the rescaled process

$$X_n(t) = \begin{cases} \alpha_n^{-1}\gamma_n^{-2}P_n g_n(\cdot,\theta_n+t\gamma_n), & \text{if } \theta_n+t\gamma_n \in \Theta \\ -\infty, & \text{otherwise} \end{cases}, \tag{3.1}$$

where $g_n(\cdot,\theta_n+t\gamma_n) = f_n(\cdot,\theta_n+t\gamma_n) - f_n(\cdot,\theta_n)$. The process X_n can be decomposed into two parts: non-random mean part, $\alpha_n^{-1}\gamma_n^{-2}Pg_n(\cdot,\theta_n+t\gamma_n)$; random perturbation part, which is a centered process. Define the corresponding centered process

$$W_n(t) = \begin{cases} \alpha_n^{-1}\gamma_n^{-2}(P_n-P)g_n(\cdot,\theta_n+t\gamma_n), & \text{if } \theta_n+t\gamma_n \in \Theta \\ -\infty, & \text{otherwise} \end{cases}. \tag{3.2}$$

To prove weak convergence of the process $X_n(t)$, we adopt the definition weak convergence proposed by Hoffman-Jørgensen(1984) and disscused by Dudley(1985). The sufficient conditions for weak convergence consist of two parts:

convergence of finite dimensional distributions, which follow from the multivariate Central Limit Theorem; a uniform tightness condition, which will be deduced from the maximal inequalities in Section 2. The following Lemma (Theorem 2.3 in Kim and Pollard(1990)) provides those conditions easier to check. Recall that $\mathcal{X}$ is the set of all locally bounded functions on $\Re^d$ equipped with the topology of uniform convergence on compacta. We use a symbol $\rightsquigarrow$ for the convergence in distribution under the metric for uniform convergence on compacta.

Lemma 3.1. Let $\{Z_n(t)\}$ be a sequence of stochastic processes with sample paths in $\mathcal{X}$. Suppose:

(i) for each finite subset of S of $\Re^d$ there is a probability measure Q_s on $\mathcal{X}$ such that $\{Z_n(t) : t \in S\} \rightsquigarrow Q_s$;

(ii) for each $\epsilon > 0$, $\eta > 0$, and $M < \infty$, there is a $\delta > 0$ such that

$$\limsup \mathbb{P}\Big\{\sup \Big|Z_n(s) - Z_n(t)\Big| > \eta\Big\} < \epsilon,$$

where the supremum runs over all pairs of s, t with $\max(|s|,|t|) \leq M$ and $|s-t| < \delta$. Then there is a Borel probability measure Q with finite-dimensional projections Q_s, such that $Z_n \rightsquigarrow Q$ and Q concentrates of the separable set of all continuous functions in $\mathcal{X}$.

The following two Lemmas are designed to ensure that the process $\{W_n(t)\}$ satisfies conditions (i) and (ii) of Lemma 3.1.

Lemma 3.2 (Convergence of Finite Dimensional Distribution). Let $\{W_n(t)\}$ be a sequence of processes defined on (3.2). Suppose :

(i) $(\theta_n + t\gamma_n) \rightarrow$ interior point of Θ.

(ii) $H(s,t) = \lim_{n\to\infty} \alpha_n^{-2}\gamma_n^{-4}n^{-1}Pg_n(\cdot,\theta_n + s\gamma_n)g_n(\cdot,\theta_n + t\gamma_n)$ exists.

Then the finite dimensional projections of the process W_n converges in distribution to the finite dimensional projections of a process $W(t)$ which is a centered gaussian process with covariance kernel H.

Proof. For each fixed t, condition (i) ensures that for large n,

$$W_n(t) = \frac{1}{n}\sum_{i=1}^{n} \alpha_n^{-1}\gamma_n^{-2}\Big[g_n(\xi_i,\theta_n+t\gamma_n) - Pg_n(\xi_i,\theta_n+t\gamma_n)\Big].$$

For each subset $(t_1,\cdots,t_k)$ in $\Re^k$, we want to show that

$$\Big(W_n(t_1),\cdots,W_n(t_k)\Big) \rightsquigarrow \Big(W(t_1),\cdots,W(t_k)\Big).$$

Since a multivariate central limit theorem can be deduced from a central limit theorem for each linear combination of the random variables it is good enough to establish a central limit theorem for the triangular array $\{h_{ni}\}$,

$$h_{ni} = \alpha_n^{-1}\gamma_n^{-2}n^{-1}\sum_{j=1}^{k}\lambda_j\Big[g_n(\xi_i,\theta_n+t_j\gamma_n) - Pg_n(\cdot,\theta_n+t_j\gamma_n)\Big],$$

for each choice of the constant $\{\lambda_j : j = 1,\ldots,k\}$. These random variables have zero means and their variance satisfying

$$\begin{aligned}\sum_{i=1}^{n} Var(h_{ni}) &= \alpha_n^{-2}\gamma_n^{-4}n^{-1}\sum_{j=1}^{k}\sum_{l=1}^{k}\lambda_j\lambda_l\Big[Pg_n(\cdot,\theta_n+t_j\gamma_n)g_n(\cdot,\theta_n+t_l\gamma_n) \\ &\qquad -Pg_n(\cdot,\theta_n+t_j\gamma_n)Pg_n(\cdot,\theta_n+t_l\gamma_n)\Big] \\ &\longrightarrow \sum_{j=1}^{k}\sum_{l=1}^{k}\lambda_j\lambda_l H(t_j,t_l),\end{aligned}$$

because $Pg_n(\cdot,\theta_n+t_j\gamma_n)Pg_n(\cdot,\theta_n+t_l\gamma_n)$ term has a smaller order of magnitude than $Pg_n(\cdot,\theta_n+t_j\gamma_n)g_n(\cdot,\theta_n+t_l\gamma_n)$. Since g_n is bounded, the triangular array, $\{h_{ni}\}$, satisfy the Lindeberg condition.

Lemma 3.3 (Uniform Tightness). For each n, let $\mathcal{G}_n$ be a VC subgraph class. If

(i) $P\Big|g_n(\cdot,\theta)-g_n(\cdot,\theta')\Big| = O\Big(|\theta-\theta'|\beta_n\Big)$ for θ, θ' near θ_n,

(ii) the subclasses $\mathcal{G}_n(R)$ of $\mathcal{G}_n$ have envelopes $G_n(R)$ such that for some finite constant C,

$$PG_n(R)^2 \le CR\beta_n \quad \text{as} \quad R \cdot\!\!\longrightarrow 0,$$

where $\beta_n = \alpha_n^{(d-1)/d}$, then for each $\epsilon > 0$, $\eta > 0$ and finite $M > 0$ there exist n_0 and $\delta > 0$ such that for $n \ge n_0$,

$$\mathbb{P}\Big\{\sup_{\Lambda(M,\delta)}\Big|W_n(t)-W_n(t')\Big| > \eta\Big\} < \epsilon, \tag{3.3}$$

where $\Lambda(M,\delta) = \Big\{(t,t') : |t| \le M, |t'| \le M, |t-t'| \le \delta\Big\}$.

Proof. Let $\mathcal{H}_n$ be the class of all functions of the form

$$g_n(\cdot,\theta_n+t\gamma_n) - g_n(\cdot,\theta_n+t'\gamma_n),$$

with $(t,t') \in \Lambda(M,\delta)$. Then $\mathcal{H}_n$ has an envelope H_n which is bounded by $2G_n(M\gamma_n)$; thus PH_n^2 decreases like $O(\gamma_n\beta_n)$. By the Maximal Inequalities 2.3, the probability in (3.3) is less than a constant multiple of

$$\mathbb{E}\sup_{\mathcal{H}_n}\alpha_n^{-1}\gamma_n^{-2}n^{-1/2}|\nu_n h_n| \;\le\; \alpha_n^{-1}\gamma_n^{-2}n^{-1/2}\mathbb{E}\left[(P_nH_n^2)^{1/2}J(\sup_{\mathcal{H}_n}P_nh_n^2/P_nH_n^2)\right].$$

Recall that $\nu_n = \sqrt{n}(P_n - P)$. Decompose according to whether $P_nH_n^2 \le \delta_0\gamma_n\beta_n$ or not. Since $\sup\limits_{\mathcal{H}_n}P_nh_n^2 \le P_nH_n^2$, and $\gamma_n = n^{-1/3}\alpha_n^{-2/3}\beta_n^{1/3}$ the right-hand side is bounded by

$$\delta_0^{1/2}J(1) + \alpha_n^{-1}\gamma_n^{-2}n^{-1/2}\sqrt{PH_n^2}\sqrt{\mathbb{E}J^2(\min(1,\sup_{\mathcal{H}_n}P_nh_n^2/\delta_0\gamma_n\beta_n))},$$

using the Cauchy Schwarz inequality. Since $\alpha_n^{-1}\gamma_n^{-2}n^{-1/2}\sqrt{PH_n^2} = O(1)$, it suffices to show that $\sup\limits_{\mathcal{H}_n}P_nh_n^2 = o_p(\gamma_n\beta_n)$. Any function h_n in $\mathcal{H}_n$ is bounded by some constant K. Hence the expected value of $\sup_{\mathcal{H}_n} P_n h_n^2$ is less than or equal to

$$K\mathbb{E}\sup_{\mathcal{H}_n} P|h_n| + K\mathbb{E}\sup_{\mathcal{H}_n}\Big|P_n|h_n| - P|h_n|\Big|.$$

By condition (i) and the definition of $\mathcal{H}_n$, the first term is $o(\gamma_n\beta_n)$; the second is less than $Kn^{-1/2}J(1)\sqrt{PH_n^2} = O(n^{-1/2}\gamma_n\beta_n)$ by the Maximal Inequality 2.3 applied to the VC subgraph class $\Big\{|h_n| : h_n \in \mathcal{H}_n\Big\}$ with the same envelope H_n.

Theorem 3.4 (Weak Convergence of X_n). If $Pg_n(\cdot,\theta)$ has a strictly negative definite second derivative, $-V_n$, at $\theta = \theta_n$ such that

$$\lim_{n\to\infty} \alpha_n^{-1}V_n = V,$$

then under the condition of Lemmas 3.2 and 3.3, the process X_n defined in (3.1) converges in distribution to the process

$$X(t) = -\frac{1}{2}t'Vt + W(t),$$

where W is a centered gaussian process with continuous sample paths and covariance kernel

$$H(s,t) = \lim_{n\to\infty} \alpha_n^{-2}\gamma_n^{-4}n^{-1}Pg_n(\cdot,\theta_n + s\gamma_n)g_n(\cdot,\theta_n + t\gamma_n).$$

Proof. The assumption on $Pg_n(\cdot,\theta)$ implies that as n gets big

$$X_n(t) = \alpha_n^{-1}\gamma_n^{-2}Pg_n(\cdot,\theta_n + t\gamma_n) + W_n(t) = -\frac{1}{2}t'Vt + o(1) + W_n(t). \quad (3.4)$$

Lemmas 3.2 and 3.3 guarantee that the process W_n satisfies the two conditions of Lemma 3.1 for convergence in distribution of stochastic processes with paths in $\mathcal{X}$. Together with (3.4) this implies $X_n \rightsquigarrow X$, where X has the asserted limit distribution.

The limit distribution of $n^{-1/3}\alpha_n^{-(d+1)/3d}(\hat{\theta}_n - \theta_n)$ is obtained by a modified continuous mapping theorem. In Kim and Pollard (1990) Theorem 2.7

accompanied with Lemmas 2.5 and 2.6 provides a perfect form for our applications. The following Theorem combines those theorem and lemmas. Let $n^{-1/3}\alpha_n^{-(d+1)/3d}(\hat{\theta}_n - \theta_n) = \hat{t}_n$.

Theorem 3.5 (A Continuous Mapping Theorem). Let $\{Z_n\}$ be random maps into $\mathcal{X}$ such that

$$Z_n \quad \rightsquigarrow \quad Z = -\frac{1}{2}t'Vt + W(t),$$

where V is a fixed positive definite matrix and W is a centered gaussian process with continuous sample paths and covariance kernel H for which

$$H(kt, kt') = kH(t, t') \quad \text{for } k > 0 \text{ and } t, t' \in \Re^d.$$

Suppose $\{\hat{t}_n\}$ satisfies the following:

(i) $\hat{t}_n = O_p(1)$;

(ii) $Z_n(\hat{t}_n) \geq \sup_t Z_n(t) - u_n$ for random variables $\{u_n\}$ of order $o_p(1)$.

Then

$$\hat{t}_n \quad \rightsquigarrow \quad \arg\max(Z).$$

Note that the limit distribution of $n^{-1/3}\alpha_n^{-(d+1)/3d}(\hat{\theta}_n - \theta_n)$ has a similar form to that of the maximization estimators examined in Kim and Pollard (1990). The only difference is the rate of convergence which depends on α_n. Kim and Pollard (1990) anaylzed the mode estimator with $\alpha_n = 1$ showing that it had a $O_p(n^{-1/3})$ rate of convergence.

The optimal rate of α_n is determined by way of minimizing the mean square errors, which depend on the underlying distribution P. Suppose we use a symmetric kernel K. If the underlying distribution has a symmetric density, where θ_n corresponds to the true mode θ_0, the optimal rate of α_n is a constant. For the case where an asymmetric underlying density produces a non-zero bias term, $|\theta_n - \theta_0| = O(\alpha_n^{k/d})$ for some $k \geq 1$, the optimal rate of α_n is $n^{-d/(d+1+3k)}$.

When $d = 1$ and $k = 2$, which is the case Chernoff(1964) considered, our result, $\alpha_n = n^{-1/8}$, coincides with his.

4. APPLICATION

In following two Examples we use a symmetric kernel function. For each θ in $\Re^2$ and n, let $f_n(\cdot, \theta)$ be the indicator function of a closed ball, $U_n(\cdot, \theta)$, with a center θ and a radius $\sqrt{\alpha_n}$. Then the class, $\mathcal{F}_n = \{f_n(\cdot, \theta) : \theta \in \Re^2\}$, is a VC subgraph class.

Example 4.1. Let P denote the standard bivariate normal distribution, $N(0, I)$, and P_n be the empirical measure based on a sample $\{\xi_i\}$ from P. Since the underlying distribution has a symmetric density $p(\cdot)$ and $f_n(\cdot, \theta)$ is also symmetric, the maximizing value θ_n of $P_n f_n(\cdot, \theta)$ is $\theta_0 = 0$. So the optimal rate of α_n is constant, which leads to the cube root asymptotics.

With $\alpha_n = 1$ Kim(1988) analysed this mode estimator, giving a cube root rate of convergence and a limit distribution as a argmax functional of a gaussian process with mean $-\frac{1}{2}e^{-1/2}$ (I correct the wrong mean value of Kim(1988)) and covariance kernel $H(s,t) = \frac{1}{\pi}e^{-1/2}(|s| + |t| - |s - t|)$.

Example 4.2. As an asymmetric case in the real line, Chernoff(1964) considered a case where a underlying density function is defined for some positive constants $c_2 > 0$ and $c_3 > 0$, $p(x) = c_0 - c_2x^2 + c_3x^3 + o(x^3)$, for x near zero. He noted that with $\alpha_n = O(n^{-1/8})$, the mode estimator $\hat{\theta}_n$ converged to the true mode θ_0 at the $O_p(n^{-1/4})$ rate of convergence.

As a two-dimensional analogue we consider a case where a underlying distribution has unique mode $\theta_0 = 0$ with a density $p(x, y) = c_0 - c_2(x^2 + y^2) + c_3(x^3 + y^3) + o(x^3 + y^3)$ for $|(x, y)| \to 0$. Assume that $c_2 > 0$ and $c_3 > 0$. The density function is symmetric with respect to the line $y = x$ in $\Re^2$ not with respect to the origin, and decrease more rapidly (or slowly) on quater plane, $x < 0$, $y < 0$ (or $x > 0$, $y > 0$). We assume whatever conditions on the density function outside the small neighborhood of origin to ensure that $|\hat{\theta}_n| = o_p(1)$.

To prove limit theorems, we first calculate $Pf_n(\cdot,\theta)$. Let $\Theta = \{\theta : |\theta| = o(1)\}$. For each $\theta = (s,t)$ in Θ, $Pf_n(\cdot,\theta)$ can be written as

$$\int\int\left\{\left|(x-s,y-t)\right| \leq \sqrt{\alpha_n}\right\}p(x,y)dxdy = \int\int\left\{\left|(x,y)\right| \leq \sqrt{\alpha_n}\right\}p(x+s,y+t)dxdy.$$

Using polar coordinates, the change of variables give

$$\begin{aligned} Pf_n(\cdot,\theta) = Pf_n(\cdot,s,t) \; &= c_0\pi\alpha_n - \frac{c_2}{2}\pi\alpha_n^2 + \frac{3c_3}{4}\pi\alpha_n^2(s+t) - c_2\pi\alpha_n(s^2+t^2) \\ &\quad + c_3\pi\alpha_n(s^3+t^3) + o(\alpha_n(s^3+t^3)). \end{aligned} \tag{4.1}$$

Frome the equation (4.1) we can see that $Pf_n(\cdot,\theta)$ is maximized at $s = t > 0$. More precisely the differenciation of the righthand-side of (4.1) gives that $Pf_n(\cdot,\theta)$ achieves its maximum at $\theta_n = (s_n,t_n)$ for which $s_n = t_n = (3c_3/8c_2)\alpha_n + o(\alpha_n)$. Since $|\theta_n| = O(\alpha_n) = o(\sqrt{\alpha_n})$, it is eventually true that

$$\sup_{|\theta-\theta_n|>\sqrt{\alpha_n}} Pf_n(\cdot,\theta) \;\leq\; \sup_{|\theta|>\sqrt{\alpha_n/2}} Pf_n(\cdot,\theta).$$

Since the last quantity is maximized at $(\sqrt{\alpha_n}/2, \sqrt{\alpha_n}/2)$,

$$Pf_n(\cdot,\theta_n) - \sup_{|\theta|>\frac{\sqrt{\alpha_n}}{\sqrt{2}}} Pf_n(\cdot,\theta) > \frac{c_2}{4}\pi\alpha_n^2 \quad \text{eventually.}$$

Put $\epsilon_n = c_2/8\pi\alpha_n$. Then (2.2) of Theorem 2.2 is satisfied when

$$\alpha_n \gg n^{-1/3}(\log n)^{1/3}. \tag{4.2}$$

Therefore with probability one it is eventually true that

$$|\hat{\theta}_n - \theta_n| \leq \sqrt{\alpha_n}.$$

Define $g_n(\cdot,\theta) = f_n(\cdot,\theta) - f_n(\cdot,\theta_n)$. Then for each n, $\mathcal{G}_n = \{g_n(\cdot,\theta) : \theta \in \Theta\}$ is also a VC subgraph class for an envelope $G_n = \sup_{\mathcal{G}_n} |g_n|$. For each

R, define subclasses $\mathcal{G}_n(R) = \left\{g_n(\cdot,\theta) : |\theta - \theta_n| \le R\right\}$, and their envelopes, $G_n(R) = \sup_{\mathcal{G}_n(R)} |g_n|$. For $R \le M\sqrt{\alpha_n}$, $G_n(R)$ is an indicator function of a set with volume $O(R\sqrt{\alpha_n})$. So $PG_n(R)^2 \le CR\sqrt{\alpha_n}$ for a finite constant C.

For rate of convergence of $(\hat{\theta}_n - \theta_n)$, we need to calculate the second derivative matrix, $-V_n$, of $Pg_n(\cdot,\theta)$ at $\theta = \theta_n$. A straightforward calculation shows that V_n is a diagonal matrix, $(2c_2\pi\alpha_n + o(\alpha_n))I$ where I is a two by two identity matrix. So $\alpha_n^{-1}V_n$ goes to a negative definite matrix, $V = 2c_2\pi I$. By Theorem 2.5, $|\hat{\theta}_n - \theta_n|$ has a $n^{-1/3}\alpha_n^{-1/2}$, which is guaranteed to be much smaller than $\sqrt{\alpha_n}$ by (4.2).

Reparametrize θ by setting $\theta = \theta_n + \lambda\gamma_n$, where $\gamma_n = n^{-1/3}\alpha_n^{-1/2}$. Then $\hat{\lambda}_n = n^{1/3}\alpha_n^{1/2}(\hat{\theta}_n - \theta_n)$ maximizes the rescaled process

$$X_n(\lambda) \;=\; \alpha_n^{-1}\gamma_n^{-2}Pg_n(\cdot,\theta_n + \lambda\gamma_n) + \alpha_n^{-1}\gamma_n^{-2}(P_n - P)g_n(\cdot,\theta_n + \lambda\gamma_n)$$

Theorem 3.4 establishes the limit distribution of X_n. Since we have shown (i) of Lemma 3.2, (ii) of Lemma 3.3 and the condition on the second deritive matrix of $Pg_n(\cdot,\theta)$, we only need to check (i) of Lemma 3.3 and (ii) of Lemma 3.2. We start with (i) of Lemma 3.3, because (ii) of Lemma 3.2 follows easily afterwards.

For each λ and λ', $\left|g_n(\cdot,\theta_n + \lambda\gamma_n) - g_n(\cdot,\theta_n + \lambda'\gamma_n)\right| = \left|f_n(\cdot,\theta_n + \lambda\gamma_n) - f_n(\cdot,\theta_n + \lambda'\gamma_n)\right|$ is an indicator function of a symmetric difference of two discs with radius $\sqrt{\alpha_n}$. Using surface integration (see Example 6.2 of Kim and Pollard(1990)), $P|f_n(\cdot,\theta_n + \lambda\gamma_n) - f_n(\cdot,\theta_n + \lambda'\gamma_n)|$ can be expressed as

$$4c_0\sqrt{\alpha_n}\gamma_n|\lambda - \lambda'| + o(\sqrt{\alpha_n}\gamma_n),$$

which satisfies the condition (i) of Lemma 3.3. Since $\left|g_n(\cdot,\theta_n+\lambda\gamma_n) - g_n(\cdot,\theta_n+\lambda'\gamma_n)\right| = \left[g_n(\cdot,\theta_n + \lambda\gamma_n) - g_n(\cdot,\theta_n + \lambda'\gamma_n)\right]^2$ and for real a, b, it is true that $ab = \frac{1}{2}[a^2 + b^2 - (a-b)^2]$ and $\alpha_n^{-2}\gamma_n^{-4}n^{-1}\sqrt{\alpha_n}\gamma_n = 1$, the covariance kernel is

$$H(\lambda,\lambda') \;=\; 2c_0\Big(|\lambda| + |\lambda'| - |\lambda - \lambda'|\Big),$$

which is a constant multiple of the covariance kernel of two-sided Brownian Motion. So the process X_n converges in distribution to a process

$$X(\lambda) \;=\; -c_2\pi|\lambda|^2 \;+\; 2\sqrt{c_0}B,$$

where B is a two-sided Brownian Motion. Since conditions in Theorem 3.5 are satisfied,

$$\hat{\lambda}_n \;=\; n^{1/3}\alpha_n^{1/2}(\hat{\theta}_n - \theta_n) \;\rightsquigarrow\; \arg\max X.$$

The optimal rate of α_n is determined as to minimize the mean square error, $\mathbb{E}\hat{\theta}_n^2$, which is divided into two parts: variance, $\mathbb{E}(\hat{\theta}_n - \theta_n)^2$; squared bias term, θ_n^2. Since minimum is achieved when two terms have same order, an optimal rate of α_n should satisfy $n^{-2/3}\alpha_n^{-1} \sim \alpha_n^2$, which gives the optimal rate of α_n is $n^{-\frac{2}{9}}$. This also satisfies the condition on α_n in (4.2). So $\hat{\theta}_n$ converges to origin, which is the true mode, at a rate of convergence, $n^{-2/9}$. More precisely,

$$n^{\frac{2}{9}}\hat{\theta}_n \;\rightsquigarrow\; \arg\max X + \frac{3c_3}{8c_2}(1,1).$$

REFERENCES

(1) Chernoff, H. (1964). Estimation of the mode. *Annals of Institute of Statistical Mathematics*, **16**, **31–41**.

(2) Dudley, R.M. (1985). An extended Wichura theorem, definitions of Donsker class, and weighted empirical distributions. *Springer Lecture Notes in Mathematics*, **1153**, **141–178**.

(3) Dudley, R.M. (1987). Universal Donsker classes and metric entropy. *Annals of Probability*, **15**, **1306–1326**.

(4) Hoffmann-Jørgensen, J. (1984). *Stochastic Processes on Polish Spcaces.* Unpublished manuscript.

(5) Kim, J. (1988). An asymptotic theory for optimization estimators with non-standard rates of convergence. Ph. D. thesis, Yale University.

(6) Kim, J. and Pollard, D.B. (1990). Cube root asymptotics. *Annals of Statistics*, **18**, **191–219**.

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

(8) Pollard, D. (1989). Asymptotics via empirical processes. *Statistical Science*, **4**, **341–366**.

(9) Vapnik, V.N. and Cervonenkis, A.YA. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Application*, **16**, **264–280**.