6-9-2005

An Empirical Process Limit Theorem for Sparsely Correlated Data

Thomas Lumley
University of Washington, tlumley@u.washington.edu

1 INTRODUCTION

Empirical process limit theorems for processes indexed by classes of functions or sets are a powerful tool, particularly for studying estimation and inference in semiparametric models (eg van der Vaart & Wellner, 1996; Pollard, 1990). Even for models with no explicit infinite-dimensional component they are useful for reducing the degree of smoothness needed to establish asymptotic Normality (eg a Lipschitz condition on estimating functions suffices to establish bracketing entropy bounds).

Initially established for independent data (Pollard, 1982; Koltchinskii, 1981; Ossiander, 1987), they have been extended from independent to dependent sequences (Doukhan et al., 1995; Yu, 1994; Arcones & Yu, 1994; Andrews & Pollard, 1994; Dehling et al., 2002) and from simple means to U-statistics (Nolan & Pollard, 1987; Arcones & Giné, 1993; de la Peña & Giné, 1999).

Here we show how to extend arguments based on bracketing entropy to a case where the correlation is sparse, meaning that most small sets of observations are independent of each other, but there need not be a simple pattern to the correlations.

For each observation X_i, $i = 1, \ldots, n$ we define a set of indices $\mathcal{S}_i$ called the *neighbourhood* with *center i* such that

1. $j \notin \mathcal{S}_i$ and $i \notin \mathcal{S}_j$ implies X_i and X_j are independent.

2. $\{i_1, i_2, \ldots, i_k\} \cap \left(\cup_{\ell=1}^{k'} \mathcal{S}_{j_\ell} \right) = \emptyset$ and $\{j_1, j_2, \ldots, j_{k'}\} \cap \left(\cup_{\ell=1}^{k} \mathcal{S}_{i_\ell} \right) = \emptyset$ implies $\{X_{i_1}, X_{i_2}, \ldots, X_{i_k}\}$ is independent of $\{X_{j_1}, X_{j_2}, \ldots, X_{j_k}\}$.

Write $M = \max_i |\mathcal{S}_i|$ and m for the size of the largest subset $\mathcal{T}$ such that $\mathcal{S}_i \cap \mathcal{S}_j = \emptyset$ for all pairs $i, j \in \mathcal{T}$ and write $m(n)$ and $M(n)$ when it is important to make explicit the dependence on n. We refer to data as *sparsely correlated* (Lumley & Mayer-Hamblett, 2003) if we can choose $\mathcal{S}_j$ so that $Mm = O(n)$. We call M the *neighbourhood size* and m the *neighbourhood number*. We write $\mathcal{S}(\mathcal{J})$ for $\bigcup_{i_j \in \mathcal{J}} \mathcal{S}_{i_j}$.

Special cases of sparsely correlated data include independent data, where we have $m = n$ and $M = 1$; panel data, where m as the number of panels and M is the maximum number of observations per panel; and k-dependent sequences where $m = 2k + 1$ and $M = n/2k$. Another, more interesting, special case of sparsely correlated data comes from U-statistics of order r, where the data X_i are r-tuples of m independent observations and f is an antisymmetric function of its r arguments. In this case we have $n = m^r$ and $M \leq rm^{r-1}$, giving $Mm \leq rn$, so U-statistics of all orders are sparsely correlated.

When specialised to U-statistics, however, our results are substantially inferior to the best known ones. For the central limit theorem and law of large numbers we require stronger moment conditions than are needed for independent data or U-statistics. For the empirical process central limit theorem we also consider only entropy with bracketing, as our approach does not allow us to extend either decoupling by symmetrisation or Hoeffding's inequality,

1

both of which are used in the standard proofs of stochastic equicontinuity under metric entropy conditions.

Practical examples of this sort of data that do not reduce to independent multivariate observations or to U-statistics are discussed by Lumley & Mayer-Hamblett (2003). The principal examples are incomplete U-statistics formed by summing a function f over only some r-tuples of observations, and incomplete crossed experiments. They include examples from the study of HIV genetics and medical diagnostics and the notorious salamander mating data of McCullagh & Nelder (1989).

In section 2 we extend Bernstein's inequality (Bennett, 1962) to sparsely correlated data. We then use this in section 3 to prove a strong law of large numbers, and in section 4 to derive an empirical process central limit theorem under bracketing entropy assumptions using a proof based on that of Ossiander (1987) as reformulated by van der Vaart (1998, Chapter 19), but replacing Bernstein's inequality with our extended version.

2   BERNSTEIN'S INEQUALITY FOR SPARSE CORRELATION

Bernstein's inequality (Bennett, 1962) in its original form gives a tail bound for the sum of uniformly bounded independent random variables. The moment condition we give is implied by a bound of $\pm 3K$.

Lemma 1 *Suppose we have X_i, $i = 1, 2, \ldots, n$ mean zero and sparsely correlated.*
Suppose that for each X_i

$$EX_i^r \le K^{r-2}\sigma^2/2 \tag{2·1}$$

Then

$$\Pr\left(\left|\sum_i X_i\right| > t\right) \le 2e^{-\frac{1}{2}\frac{t^2}{Mn\sigma^2 + MKt}}.$$

Proof: The proof largely follows Bennett (1962), but differs in how the individual moment bound is related to the exponential moments of the sum.

The rth moment of the sum is

$$E(\sum_{i=1}^n X_i)^r = \sum_{i_1, i_2, \ldots, i_r}^n EX_{i_1} X_{i_2} \cdots X_{i_r}$$

For independent summands a term $EX_{i_1} X_{i_2} \cdots X_{i_r}$ is non-zero only if every distinct index appears an even number of times. For sparsely correlated summands a term may also be non-zero if a repeated index is replaced by another index in its neighbourhood. For example, in addition to $EX_1 X_1 X_2 X_2$ there are all terms $EX_1 X_i X_2 X_j$ where $i \in \mathcal{S}_1$ and $j \in calS_2$. This expands the number of non-zero terms by a factor not exceeding

2
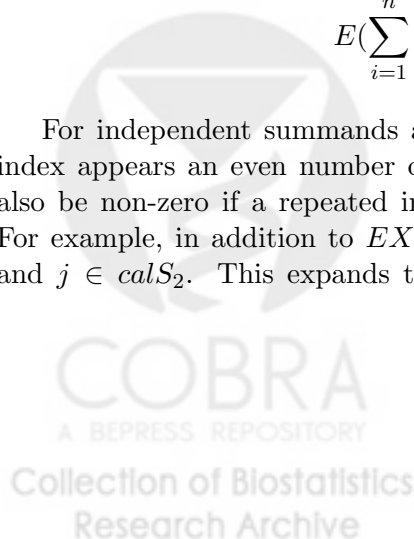
$M^{r-2}(r-1)!$. Each term is bounded in magnitude by $K^{r-2}\sigma^2/2$. In Bennett (1962), the moments of sums with independent summands are also bounded by $K^{r-2}\sigma^2/2$, so we can simply multiply his bounds by $M^{r-2}(r-1)!$.

$$
\begin{aligned}
Ee^{cS_n} &= 1 + \frac{n}{2}\sigma^2 c^2 \sum_{r=2}^{\infty} \frac{c^{r-2}ES_m^r}{nr!\sigma^2/2} \\
&< \exp\left[\frac{n}{2}\sigma^2 c^2 \sum_{r=2}^{\infty} \frac{c^{r-2}ES_m^r}{nr!\sigma^2/2}\right] \\
&\leq \exp\left[\frac{n}{2}\sigma^2 c^2 \sum_{r=2}^{\infty} \frac{c^{r-2}nM^{r-1}(r-1)!K^{n-2}\sigma^2/2}{nr!\sigma^2/2}\right] \\
&< \exp\left[\frac{n}{2}M\sigma^2 c^2 \sum_{r=2}^{\infty}(cMK)^{r-2}\right] \\
&= \exp\left[\frac{nM\sigma^2 c^2}{2(1-cMK)}\right]
\end{aligned}
$$

Write $\tilde{K}$ for MK and $\tilde{\sigma}^2$ for $nM\sigma^2$ to get

$$
Ee^{cS_n} < \exp\left[\frac{\tilde{\sigma}^2 c^2}{2(1-c\tilde{K})}\right]
$$

Now for every positive c

$$
P[S_n \geq t\tilde{\sigma}] \leq \frac{Ee^{cS_n}}{e^{ct\tilde{\sigma}}} < \exp\left[\frac{\tilde{\sigma}^2 c^2}{2(1-c\tilde{K})} - ct\tilde{\sigma}\right],
$$

equation 2a of Bennett (1962), who shows this implies

$$
\begin{aligned}
P[S_n \geq t\tilde{\sigma}] &< \exp\left[-\frac{t^2}{1+\tilde{K}t/\tilde{\sigma}+\sqrt{1+2\tilde{K}t/\sigma}}\right] \\
&< \exp\left[-\frac{t^2}{2+2\tilde{K}t/\tilde{\sigma}}\right]
\end{aligned}
$$

So

$$
P[S_n \geq t] < \exp\left[-\frac{1}{2}\frac{t^2}{\tilde{\sigma}^2+\tilde{K}t}\right]
$$

is a one-sided bound, and the conclusion of the theorem follows by adding the corresponding lower bound. ∎

3

3   STRONG LAW OF LARGE NUMBERS

This sub-exponential bound on the tails of the mean immediately gives a strong law of large numbers for bounded sparsely correlated sequences using the Borel–Cantelli lemma.

A stronger theorem can be obtained by truncating X_i at $m(i)$, provided that $\sum_i \Pr(|X_i| > m(i))$ can be controlled, where $m(i)$ is the size of the largest subset $\mathcal{T}_i$ of $1, 2, \ldots, i$ with $\mathcal{S}_j \cap \mathcal{S}_k = \emptyset$ for $i, j \in \mathcal{T}_i$.

Theorem 2 *Let X_i be a mean zero sparsely correlated sequence with $M = O(m^d)$ and with $P(|X_i| > t) < Ct^{-\alpha}$ for some $\alpha > 0$ and C not depending on i. If $\alpha > d + 1$ then $\bar{X}_i \overset{a.s.}{\to} 0$.*

Proof: We truncate X_i at $\pm m(i)$ to Y_i and use Bernstein's inequality from Lemma 1 to obtain

$$\Pr(|\bar{Y}_i| < t) \leq 2\exp\left(-\frac{1}{2}\frac{t^2}{M(i)L^2/i + M(i)m(i)t/i}\right) < 2e^{-Dt} \tag{3.1}$$

for some $D > 0$. By the Borel–Cantelli lemma $\bar{Y}_i \overset{a.s.}{\to} 0$. We now need to handle the tails $Z_i = X_i - Y_i$. Now

$$
\begin{aligned}
\sum_{i=1}^{\infty} P(Z_i \neq 0) \;&<\; \sum_{i=1}^{\infty} Cm(i)^{-\alpha)} \\
&=\; \sum_{m=1}^{\infty} \sum_{i:m(i)=m} Cm(i)^{-\alpha} \\
&\leq\; \sum_{m=1}^{\infty} CMm^{-\alpha} \\
&\leq\; \sum_{m=1}^{\infty} Cm^{d-\alpha}
\end{aligned}
$$

This converges if $d - \alpha < -1$, so by the Borel–Cantelli lemma, $P(Z_i \neq 0 \text{ i.o.}) = 0$. So $\bar{X}_i \overset{a.s.}{\to} 0$. ∎

A weak law of large numbers for sparsely correlated data with $Mm = O(n)$ and $2 + \delta$ finite moments follows the central limit theorem proved by Lumley & Mayer-Hamblett (2003). In contrast, our proof of the strong law requires $M = O(m)$ for this moment condition. Presumably this gap could be narrowed by a more sophisticated proof.

These laws of large numbers give Glivenko–Cantelli theorems for classes of functions with finite bracketing numbers by exactly the same finite approximation arguments as for independent data (eg van der Vaart & Wellner, 1996, Theorem 2.4.1).

4

The moment condition (2·1) is satisfied by a random variable bounded by K with ν_i being its variance. This leads to an important maximal inequality for finite classes of functions via the following lemma, which we quote from van der Vaart & Wellner (1996, Lemma 2.2.10).

Lemma 3 *Let $X_1, X_2, \ldots, X_k$ be random variables that satisfy the tail bound*

$$\Pr(|X_i| > x) \le 2e^{-\frac{1}{2}\frac{x^2}{b+ax}}$$

for all x and fixed a, $b > 0$. Then

$$\left\| \max_{1 \le i \le k} X_i \right\|_{\psi_1} \le K\left(a\log(1+k) + \sqrt{b}\sqrt{\log(1+k)} \right),$$

for a universal constant K. Here $\|\cdot\|_{\psi_1}$ is the Orlicz norm corresponding to the function $e^x - 1$, defined as

$$\|X\|_{\psi_1} = \inf\left\{ C > 0 : e^{|X|/C} - 1 \le 1 \right\}$$

Under the condition $Mm = O(n)$, Lumley & Mayer-Hamblett (2003) proved a central limit theorem with normalisation $\sqrt{m}/n$ rather than $1/\sqrt{n}$, suggesting

$$\mathbb{G}_n f \equiv \frac{\sqrt{m}}{n} \sum_{i=1}^{n} f(X_i) - Ef(X_i)$$

as the appropriate definition for the empirical process.

Combining this definition with Theorem 2 and Lemma 3 we have for a finite set $\mathcal{F}$ of functions with cardinality $|\mathcal{F}| > 2$

$$E \left\| \mathbb{G}_n \right\|_{\mathcal{F}} \lesssim \max_f \frac{\|f\|_\infty}{\sqrt{m}} \log|\mathcal{F}| + \max_f \|f\|_{P,2} \sqrt{\frac{mM}{n}} \sqrt{\log|\mathcal{F}|},$$

where we write $a \lesssim b$ if $a \le Kb$ for some constant K that depends only on $\sup Mm/n$.

As $Mm = O(n)$ this simplifies to

$$E \left\| \mathbb{G}_n \right\|_{\mathcal{F}} \lesssim \max_f \frac{\|f\|_\infty}{\sqrt{m}} \log|\mathcal{F}| + \max_f \|f\|_{P,2} \sqrt{\log|\mathcal{F}|}, \tag{4·1}$$

extending van der Vaart (1998, lemma 19.33). Note that in independent data $m = n$ and $M = 1$ and this inequality reduces to the usual one. This is the maximal inequality needed to prove a central limit theorem with L_2 bracketing.

This gives the following bound for the empirical process

5

Lemma 4 *Let X_i be sparsely correlated with identical marginal distribution P and $mM = O(n)$, and let $\mathcal{F}$ be a class of measurable functions with $Pf^2 < \delta^2$. Then with*

$$a(\delta) = \frac{\delta}{\sqrt{\log 1 + N_{[]}(\delta, \mathcal{F}, L_2(P))}}$$

we have

$$E^* \|\mathbb{G}_n\|_{\mathcal{F}} \lesssim \int_0^{\delta} \sqrt{\log 1 + N_{[]}(\epsilon, \mathcal{F}, L_2(P))}\, d\epsilon + \sqrt{m} E^* F\{F > \sqrt{m} a(\delta)\}$$

where E^* denotes outer expectation.

Proof: The proof is almost identical to that of Lemma 19.34 of van der Vaart (1998) with n replaced by m everywhere and his Lemma 19.33 replaced by inequality 4·1 above.

Define $\text{Log}(x) = \log(1 + x)$ for convenience. Let

$$a(\delta) = \delta / \sqrt{(\text{Log}N_{[]}(\epsilon, \mathcal{F}, L_2(P)))}$$

and truncate $f \in \mathcal{F}$ at $a(\delta)$. The tail from the truncation has

$$E^* \|\mathbb{G}_n\|_{\mathcal{F}} \leq 2m^{-1/2} F\{F > a(\delta)\},$$

which goes to zero with increasing m for any fixed δ by the square-integrability of F.

Now assume that each $F \in \mathcal{F}$ is bounded by $\sqrt{m} a(\delta)$. Fix q_0 such that $4\delta < 2^{q_0} \leq 8\delta$. Partition $\mathcal{F}$ into a nested sequence $\cup_{i=1}^{N_q} \mathcal{F}_{qi}$ of subsets, such that

$$\sum_{q \geq q_0} 2^{-q} N_q \lesssim \int_0^{\delta} \sqrt{\text{Log}N_{[]}(\epsilon, \mathcal{F}, L_2(P))}\, d\epsilon,$$

and $\sup_{f,g \in \mathcal{F}_{qi}} |f - g| \leq \Delta_{qi} < 2F$ and $E\Delta_{qi}^2 < 2^{-2q}$. This is possible by the finiteness of the bracketing integral.

Choose for each $q \geq q_0$ a fixed element $f_{qi} \in \mathcal{F}_{qi}$ and define $\pi_q f = f_{qi}$ and $\Delta_q f = \Delta_{qi}$ if $f \in \mathcal{F}_{qi}$. Define for each fixed m and all $q \geq q_0$

$$
\begin{aligned}
a_q &= 2^{-q} / \sqrt{\text{Log}N_{q+1}} \\
A_{q-1} f &= I\{\Delta_{q_0} f \leq \sqrt{m} a_{q_0}, \ldots, \Delta_{q-1} f \leq \sqrt{m} a_{q-1}\} \\
B_q f &= A_{q-1} f\{\Delta_q f > \sqrt{m} a_q\}.
\end{aligned}
$$

Now decompose pointwise in x

$$f - \pi_{q_0} f = \sum_{q_0+1}^{\infty} (f - \pi_q f) B_q f + \sum_{q_0+1}^{\infty} (\pi_q f - \pi_{q-1} f) A_{q-1} f$$

6

noting that either all $B_q f = 1$ and $A_q f = 1$ or there is a unique $q_1 > q_0$ with $B_q f = 1$, in which case $A_q f = 1$ for $q < q_1$ and $A_q f = 0$ for $q > q_1$.

Now apply the empirical process $\mathbb{G}_n$ separately to each series, and use the triangle inequality and the finite-class maximal inequality(4·1)

$$
E^* \left\| \sum_{q_0+1}^{\infty} \mathbb{G}_n (f - \pi_q f) B_q f \right\|_{\mathcal{F}} \leq \sum_{q_0+1}^{\infty} E^* \| \mathbb{G}_n \Delta_q f B_q f \| \mathcal{F}
$$

$$
\lesssim \sum_{q_0+1}^{\infty} \left[a_{q-1} \mathrm{Log} N_q + 2^{-q} \sqrt{\mathrm{Log} N_q} + \frac{4}{a_q} 2^{-2q} \right]
$$

We can bound the right-hand side by a multiple of $\sum_{q_0+1}^{\infty} 2^{-q} \sqrt{\mathrm{Log} N_q}$.

Next

$$
E^* \left\| \sum_{q_0+1}^{\infty} \mathbb{G}_n (\pi_q f - \pi_{q-1} f) A_{q-1} f \right\|_{\mathcal{F}} \lesssim \sum_{q_0+1}^{\infty} a_{q-1} \mathrm{Log} N_q + 2^{-q} \sqrt{\mathrm{Log} N_q}.
$$

Again this is bounded above by a multiple of $\sum_{q_0+1}^{\infty} 2^{-q} \sqrt{\mathrm{Log} N_q}$. Finally, consider $\pi_q f$. By assumption $|\pi_{q_0} f| < F < a(\delta) \sqrt{m}$ and $\|\pi_{q_0} f\|_{P,2} \leq \delta$, so

$$
E^* \| \mathbb{G}_n \pi_{q_0} f \|_{\mathcal{F}} \lesssim a_{q_0} \mathrm{Log} N_{q_0} + \delta \sqrt{\mathrm{Log} N_{q_0}}.
$$

By the choice of q_0 this is also bounded above by (the initial terms of) $\sum_{q_0+1}^{\infty} 2^{-q} \sqrt{\mathrm{Log} N_q}$.
∎

Now we can prove the empirical process central limit theorem

Theorem 5 *Let X_i be sparsely correlated with common marginal distribution P and $mM = O(n)$, and let $\mathcal{F}$ be a class of measurable functions with finite $2 + \delta$ moments and envelope function F. If $E[F^2] < \infty$ and*

$$
\int_0^1 \sqrt{\log 1 + N_{[]}(\epsilon, \mathcal{F}, L_2(P))} \, d\epsilon < \infty
$$

then $\mathcal{F}$ is Donsker, i.e., $\mathbb{G}_n$ converges weakly in the sense of Hoffman-Jørgensen to a continuous Gaussian process indexed by $\mathcal{F}$

Proof: The convergence of finite-dimensional distributions is proved by Lumley & Mayer-Hamblett (2003, Theorem 4) using the method of Stein (1972). It is therefore sufficient to prove stochastic equicontinuity: to show that for every $\eta > 0$ there exists a finite partition of $\mathcal{F}$ into sets $\mathcal{F}_i$ such that

$$
\lim_{n \to \infty} E^* \left[\sup_i \sup_{f,g \in \mathcal{F}_i} |\mathbb{G}_n (f - g)| \right] \leq \eta.
$$

7

Let $\mathcal{G} = \{f - g | f, g \in \mathcal{F}\}$ be the collection of differences of functions in $\mathcal{F}$. The bracketing entropies of $\mathcal{G}$ and $\mathcal{F}$ are proportional, so the entropy integral

$$\int_0^1 \sqrt{1 + \log N_{[]}(\epsilon, \mathcal{G}, L_2(P))}\, d\epsilon$$

is still finite.

For any small δ we can choose $N_{[]}(\delta, \mathcal{G}, L_2(P))$ brackets and use them to partition $\mathcal{F}$ into $N_{[]}(\delta, \mathcal{G}, L_2(P))$ disjoint sets $\mathcal{F}_i$ with diameters less than δ. The set $\mathcal{G}_i$ of differences of functions in $\mathcal{F}_i$ consists of functions with $L_2(P)$ norm less than δ and has $2F$ as an envelope, so by lemma 4

$$E^* \left[\sup_i \sup_{f,g \in \mathcal{F}_i} |\mathbb{G}_n(f - g)| \right] \lesssim \int_0^\delta \sqrt{1 + \log N_{[]}(\epsilon, \mathcal{F}, L_2(P))}\, d\epsilon + \sqrt{m} E F \{F > a(\delta)\sqrt{m}\}.$$

The first term on the right hand side does not depend on n and goes to zero as $\delta \to 0$. The second term is bounded by $a(\delta)^{-1} E[F^2 \{F > a(\delta)\sqrt{m}\}]$, which goes to zero as $n \to \infty$ for any δ since F has a finite second moment. ∎

References

ANDREWS, D. W. K. & POLLARD, D. (1994). An introduction to functional central limit theorems for dependent stochastic processes. *International Statistical Review* **62**, 119–132.

ARCONES, M. & GINÉ, E. (1993). Limit theorems for U-processes. *Annals of Probability* **21**, 1494–1542.

ARCONES, M. A. & YU, B. (1994). Central limit theorems for empirical and U-processes of stationary mixing sequences. *Journal of Theoretical Probability* **7**, 47–71.

BENNETT, G. (1962). Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association* **57**, 33–45.

DE LA PEÑA, V. & GINÉ, E. (1999). *Decoupling: from dependence to independence.* Springer, New York.

Dehling, H., Mikosch, T., & Sörensen, M., editors (2002). *Empirical Process Techniques for Dependent Data.* Birkhäuser.

DOUKHAN, P., MASSART, P., & RIO, E. (1995). Invariance principles for absolutely regular empirical processes. *Annals de L'Institute Henri Poincaré: Probabilités et Statistiques.* **31**, 393–427.

KOLTCHINSKII, V. I. (1981). On the central limit theorem for empirical measures. *Theory of Probability and Mathematical Statistics* **24**, 71–82.

LUMLEY, T. & MAYER-HAMBLETT, N. (2003). Asymptotics for marginal generalized linear models with sparse correlations. Technical Report 207, Department of Biostatistics, University of Washington.

8

McCullagh, P. & Nelder, J. (1989). *Generalised Linear Models.* Monographs on Statistics and Applied Probability. Chapman and Hall, London, 2nd edition.

Nolan, D. & Pollard, D. (1987). *U*-processes: Rates of convergence. *Annals of Statistics* **15**, 780–799.

Ossiander, M. (1987). A central limit theorem under metric entropy with L_2 bracketing. *Annals of Probability* **15**, 897–919.

Pollard, D. (1982). A central limit theorem for empirical processes. *Journal of the Australian Mathematical Society A* **33**, 235–248.

Pollard, D. (1990). *Empirical Processes: Theory and Applications*, volume 2 of *NSF-CBMS Regional Conference Series in Probability and Statistics*. Institute of Mathematical Statistics.

Stein, C. (1972). A bound for the error in normal approximation of a sum of dependent random variables. In *Proceedings of the 6th Berkeley Symposium in Mathematical Statisics and Probability*, pages 583–603.

van der Vaart, A. W. (1998). *Asymptotic Statistics.* Cambridge University Press, Cambridge, UK.

van der Vaart, A. W. & Wellner, J. A. (1996). *Weak Convergence and Empirical Processes with Applications to Statistics.* Springer Series in Statistics. Springer, New York.

Yu, B. (1994). Rates of convergence of empirical processes for stationary mixing sequences. *Annals of Probability* **22**, 94–116.