Central limit theorem for functions of weakly dependent variables

Jensen, Jens Ledet
Aarhus University, Department of Mathematical Sciences
Ny Munkegade, 8000 Aarhus C, Denmark
E-mail: jlj@imf.au.dk

Result

We describe a central limit theorem for the sum $S_n = \sum_{i=1}^n X_i$, $X_i \in \mathbf{R}^p$, where the X_i's can be approximated by weakly dependent variables. The proof is for $i \in \mathbf{Z}$, but can be directly generalized to the case of a random field, $i \in \mathbf{Z}^d$. The weak dependency is formulated through a set of σ-algebras, D_j, $j \in \mathbf{Z}$. The strong mixing coefficients of these are $\alpha(k,l,d) = \sup\big|P(A_1 \cap A_2) - P(A_1)P(A_2)\big|$, where the supremum is taken over sets $A_i \in \sigma\big(D_j : j \in I_i\big)$, $i = 1,2$, with $|I_1| \le k$, $|I_2| \le l$, and the distance $d(I_1, I_2)$ between the two sets I_1 and I_2 is at least d. The central limit theorem has applications in the study of nonhomogeneous hidden Markov chains (Jensen 2005).

Theorem 1. *Assume that there exist* $\delta_0, \epsilon_0 > 0$, $\delta_1, \delta_2 \ge 0$, $\theta > \delta_1 + \delta_2 + \max\big\{(2+\delta_0)/\delta_0, 1+\delta_2, 2\big\}$ *and constants* c_0, c_1, c_2 *such that*

(1) $\quad EX_i = 0, \quad E|X_i|^{2+\delta_0} \le c_0, \quad \mathrm{Var}(a \cdot S_n) \ge \epsilon_0 n|a| \quad \forall a \in \mathbf{R}^p$

(2) $\quad \alpha(k,l,d) \le c_1 k^{\delta_1} l^{\delta_2} \max\{1,d\}^{-\theta}$,

(3) $\quad \forall\, m \in \mathbf{N}\ \exists X_j^m \in \sigma\big(D_k : d(k,j) \le m\big): \quad E\big|X_j - X_j^m\big| \le c_2 m^{-\theta}$.

Then we have that $\mathrm{Var}(S_n)^{-1/2} S_n \xrightarrow{\sim} N_p(0,I)$ *for* $n \to \infty$.

(For the case of a random field, X_i, $i \in \mathbf{Z}^\nu$, the lower bound on θ is multiplied by ν.) We divide the proof into a number of subsections. In the first two subsections we use truncation to reduce the problem to that of bounded variables. In the last section the method of Bolthausen (1982) is used for the bounded variables.

Truncation

We use the trunctation function T_M where $T_M(x)$ equals x for $|x| \le M$ and equals $Mx/|x|$ otherwise. Let $Q_M(x) = x - T_M(x)$. Using that $EX_j = 0$ we write the sum S_n as $S_n = S'_n + S''_N$, with

$$(4) \quad S'_n = \sum_{j=1}^n \big[T_M(X_j) - E(T_M(X_j))\big] \quad \text{and} \quad S''_n = \sum_{j=1}^n \big[Q_M(X_j) - E(Q_M(X_j))\big],$$

and the idea is to prove that a CLT for S'_n for any fixed M implies a CLT for S_n.

In the lemma below we consider fixed values of j, k and fixed unit vectors a_j and a_k. We define $U = a_j \cdot X_j$, $U_M = a_j \cdot T_M(X_j)$ and $\hat{U}_M = a_j \cdot Q_M(X_j)$, and define V, V_M and $\hat{V}_M$ similarly with j replaced by k.

Lemma 2. *There exists a constant* c_3, *depending on* c_1, c_2, δ_0 *and* θ *only, such that*

$$\mathrm{Cov}(U,V) \le c_3\big[c_0^{1/(2+\delta_0)} + c_0^{2/(2+\delta_0)}\big] \max\big\{1, d(j,k)\big\}^{-\kappa},$$

where $\kappa = (\theta - \delta_1 - \delta_2)\delta_0/(2+\delta_0) > 1$.

Proof. Following Deo (1973) we expand $\mathrm{Cov}(U,V) = \mathrm{Cov}\big(U_M + \hat{U}_M, V_M + \hat{V}_M\big)$ into four terms and bound each of these. Thus, using the simple bound $\big|Q_M(x)\big| \le \big(|x|/M\big)^{1+\delta_0}|x|$ we find

$$\big|\mathrm{Cov}\big(U_M, \hat{V}_M\big)\big| \le 2ME\big|\hat{V}_M\big| \le 2Mc_0/M^{1+\delta_0} = 2c_0/M^{\delta_0}.$$

Similarly, $\left|\mathrm{Cov}\big(\hat U_M, V_M\big)\right| \le 2c_0/M_0^\delta$, and

$$\left|\mathrm{Cov}\big(\hat U_M, \hat V_M\big)\right| \le \big\{\mathrm{Var}(\hat U_M)\,\mathrm{Var}(\hat V_M)\big\}^{1/2} \le c_0/M^{\delta_0},$$

since $E\hat U_M^2 \le E\big|X_j\big|^{2+\delta_0}/M^{\delta_0}$. To bound $\mathrm{Cov}\big(U_M, V_M\big)$ we define $U_M^m = a_j \cdot T_M\big(X_j^m\big)$ and $V_M^m = a_k \cdot T_M(X_k^m)$. Since $\big|T_M(X_j) - T_M(X_j^m)\big| \le \big|X_j - X_j^m\big|$ we find

$$\left|\mathrm{Cov}\big(U_M - U_M^m, V_M\big)\right| \le 2ME\big|U_M - U_M^m\big| \le 2c_2 Mm^{-\theta},\ \left|\mathrm{Cov}\big(U_M^m, V_M - V_M^m\big)\right| \le 2c_2 Mm^{-\theta},$$

and therefore $\big|\mathrm{Cov}(U_M, V_M)\big| \le \big|\mathrm{Cov}(U_M^m, V_M^m)\big| + 4c_2 Mm^{-\theta}$. Finally, we use the classical bound (Ibragimov and Linnik, 1971 [17.2.1]) $\big|\mathrm{Cov}(U_M^m, V_M^m)\big| \le 4M^2\alpha\big(2m+1, 2m+1, \max\{0, d(j,k) - 2m\}\big)$. Putting all the terms together we obtain

$$\big|\mathrm{Cov}(U, V)\big| \le 5c_0/M^{\delta_0} + 4c_2 Mm^{-\theta} + 4c_1 M^2 (2m+1)^{\delta_1+\delta_2} \max\big\{1, d(j,k) - 2m\big\}^{-\theta}.$$

Choosing $m = \lfloor d(j,k)/3 \rfloor$ and $M = c_0^{1/(2+\delta_0)} d(j,k)^{\kappa/\delta_0}$ we get the result of the lemma. □

Lemma 3. *Let S_n'' be defined in (4). There exists a function $b(M)$ with $b(M) \to 0$ for $M \to \infty$ such that for all unit vectors a and for all n we have* $\mathrm{Var}\big(a \cdot S_n''\big)/n \le b(M)$.

Proof. We use Lemma 2 with the random variable X replaced by $Z = Q_M(X) - E(Q_M(X))$. Let $0 < \xi < \delta_0$ be so large that $\kappa_1 = (\theta - \delta_1 - \delta_2)\xi/(2+\xi) > 1$. Remembering the simple inequality $\big|Q_M(x)\big| \le \big(|x|/M\big)^\alpha |x|$ we have $\big|ET_M(X_i)\big| = \big|EQ_M(X_i)\big| \le c_0/M^{1+\delta_0}$ and $E\big|Q_M(X_i)\big|^{2+\xi} \le c_0/M^{\delta_0-\xi}$. Thus, replacing δ_0 by ξ we use the bound

$$\begin{aligned} E\big|Z_i\big|^{2+\xi} &\le 2^{1+\xi}\big\{E\big|Q_M(X_i)\big|^{2+\xi} + \big|EQ_M(X_i)\big|^{2+\xi}\big\} \\ &\le 2^{1+\xi}\big\{\big(c_0/M^{\delta_0-\xi}\big) + \big(c_0/M^{1+\delta_0}\big)^{2+\xi}\big\} = c_0(M,\xi). \end{aligned} \tag{5}$$

Furthermore, we can approximate Z by $Q_M(X^m) - E(Q_M(X))$ with a mean error

$$\begin{aligned} E\big|Q_M(X) - Q_M(X^m)\big| &\le E\big|X - X^m\big| + E\big|T_M(X) - T_M(X^m)\big| \\ &\le 2E\big|X - X^m\big| \le 2c_2 m^{-\theta}. \end{aligned}$$

We can now use Lemma 2 with δ_0 replaced by ξ, c_0 replaced by $c_0(M,\xi)$ and c_2 replaced by $2c_2$. For some constant $\tilde c_3$ and any unit vector a we then have

$$\mathrm{Cov}\big(a \cdot Q_M(X_j), a \cdot Q_M(X_k)\big) \le \tilde c_3 \tilde b(M) \max\big\{1, d(j,k)\big\}^{-\kappa_1},$$

where $\tilde b(M) = c_0(M,\xi)^{1/(2+\xi)} + c_0(M,\xi)^{2/(2+\xi)}$. Writing $\mathrm{Var}(S_N'')$ as a double sum of covariances we find the result of the lemma with $b(M) = \tilde c_3 \tilde b(M)[3 + 2/(\kappa_1 - 1)]$. From (5) we see that $c_0(M,\xi)$ tends to zero, and therefore $b(M)$ tends to zero as M tends to infinity. □

Variance

Writing $\mathrm{Var}(a \cdot S_n)$ as a double sum of covariances we get directly from Lemma 2 the bound $\mathrm{Var}(a \cdot S_n) \le c_4 n$ with $c_4 = c_3\big[c_0^{1/(2+\delta_0)} + c_0^{2/(2+\delta_0)}\big][3 + 2/(\kappa - 1)]$.

From Lemma 3 we find

$$\begin{aligned} \big|\mathrm{Var}\big(a \cdot S_n/\sqrt{n}\big) - \mathrm{Var}\big(a \cdot S_n'/\sqrt{n}\big)\big| &\le 2\big|\mathrm{Cov}\big(a \cdot S_n/\sqrt{n}, a \cdot S_n''/\sqrt{n}\big)\big| + \mathrm{Var}\big(a \cdot S_n''/\sqrt{n}\big) \\ &\le 2\sqrt{c_4 b(M)} + b(M) \to 0 \ \text{ for } M \to \infty. \end{aligned}$$

The assumption of a lower bound on the variance of $S_n/\sqrt{n}$ in Theorem 1 therefore gives that

(6) $\quad \mathrm{Var}\big(a \cdot S_n/\sqrt{n}\big)/\mathrm{Var}\big(a \cdot S_n'/\sqrt{n}\big) \to 1$ for $M \to \infty$.

As in Ibragimov and Linnik (1971, page 346) we have from Lemma 3 and (6) that it suffices to prove a CLT for S_n' for fixed M to obtain a CLT for S_n.

Proof of central limit theorem for S_n'

Let a be a fixed unit vector and define $Y_i = a \cdot \big(T_M(X_i) - ET_M(X_i)\big)$ and $\tilde{S}_n = \sum_{i=1}^n Y_i = a \cdot S_n'$.

We saw in the proof of Lemma 3 that $\big|ET_M(X_i)\big| \leq c_0/M^{1+\delta_0}$ so that for M sufficiently large we have $|Y_i| \leq M+1$. Furthermore, if we set $Y_i^m = T_M(X_i^m) - ET_M(X_i)$ we have $E\big|Y_i - Y_i^m\big| \leq E\big|X_i - X_i^m\big| \leq c_2 m^{-\theta}$ and $\big|Y_i^m\big| \leq M+1$.

We will use the method of proof from Bolthausen (1982). For this we need the following estimates.

Lemma 4. *There exists a constant c_4 such that*

$$\mathrm{Cov}\big(Y_j, Y_k\big) \leq c_4(M+1)^2 \max\{1, d(j,k)\}^{-\gamma},\ \mathrm{Cov}\big(Y_jY_k, Y_rY_s\big) \leq c_4(M+1)^4 \max\big\{1, d(\{j,k\},\{r,s\})\big\}^{-\gamma},$$

and $\mathrm{Cov}\big(Y_j, Y_kY_rY_s\big) \leq c_4(M+1)^4 \max\big\{1, d(j,\{k,r,s\})\big\}^{-\gamma}$, *where* $\gamma = \theta - \delta_1 - \delta_2$.

Proof. The proof is based on successively replacing Y_i by Y_i^m in the mean of a product of Y's. Thus, using the second inequality as an example, $\big|E(Y_jY_kY_rY_s) - E(Y_j^mY_k^mY_r^mY_s^m)\big| \leq 4(M+1)^3 c_2 m^{-\theta}$. From inequalities of this form we obtain

$$\big|\mathrm{Cov}(Y_jY_k, Y_rY_s) - \mathrm{Cov}(Y_j^mY_k^m, Y_r^mY_s^m)\big| \leq 2 \cdot 4(M+1)^3 c_2 m^{-\theta}.$$

Next, the strong mixing implies (Ibragimov and Linnik, 1971 [17.2.1])

$$\big|\mathrm{Cov}(Y_j^mY_k^m, Y_r^mY_s^m)\big| \leq 4(M+1)^4 c_1[2(2m+1)]^{\delta_1+\delta_2} \max\big\{1, d(\{j,k\},\{r,s\}) - 2m\big\}^{-\theta}.$$

Combining the two inequalities and taking $m = \lfloor d(\{j,k\},\{r,s\})/4 \rfloor$ we obtain the second inequality of the lemma. □

For a number r we introduce the notation

$$S_{i,n} = \sum_{j=1, d(i,j)\leq r}^{n} Y_j \quad \text{and} \quad \alpha_n = \sum_{i=1}^{n} E(Y_i S_{i,n}),$$

where eventually r will be tending to infinity with n. From Lemma 4 we find that

$$\mathrm{Var}\big(\tilde{S}_n/\sqrt{n}\big) - \frac{\alpha_n}{n} = \frac{1}{n}\sum_{i=1}^{n} \sum_{j=1, d(i,j)>r}^{n} \mathrm{Cov}(Y_i, Y_j) \leq \frac{4c_4(M+1)^2}{(\gamma-1)r^{\gamma-1}},$$

where $\gamma - 1 = \theta - \delta_1 - \delta_2 - 1 > 0$. Now, consider the case $r = n^{\omega}$, with $\omega > 0$. Then the right hand side above tends to zero. Since also we have from (6) that $\mathrm{Var}(\tilde{S}_n)$ is of the same order as $\mathrm{Var}(S_n)$, and we have assumed the lower bound $\epsilon_0 n$ for the latter, we find that α_n has a similar lower bound and $\mathrm{Var}(\tilde{S}_n)/\alpha_n \to 1$. We must therefore show that $\bar{S}_n = \tilde{S}_n/\sqrt{\alpha_n}$ converges to a standard normal distribution.

Proof of CLT for $\bar{S}_n$ We follow the proof of Bolthausen (1982), where the definitions of A_1, A_2 and A_3 below can be found. These terms depend on the argument t of the characteristic function for

$\bar{S}_n$. One needs to show that $E\big(A_1 - A_2 - A_3\big) \to 0$. Using the expression in Bolthausen (1982) and Lemma 4 we have with $J_r = \{1 \le i, i', j, j' \le n : d(i,j) \le r, d(i',j') \le r\}$,

$$\begin{aligned}
E\big|A_1\big|^2 &= \frac{t^2}{\alpha_n^2}\sum_{J_r}\mathrm{Cov}(Y_iY_j, Y_{i'}Y_{j'}) \\
&= \frac{t^2}{\alpha_n^2}\Big\{\sum_{J_r, d(i,i')\ge 3r}\mathrm{Cov}(Y_iY_j, Y_{i'}Y_{j'}) + \sum_{J_r, d(i,i')<3r}\mathrm{Cov}(Y_iY_j, Y_{i'}Y_{j'})\Big\} \\
&\le \frac{c_4t^2(M+1)^4}{\alpha_n^2}\Big\{n(2r+1)^2 2\sum_{k=3r}^{\infty}(k-2r)^{-\gamma} + n(2r+1)\cdot 8\cdot 6r\big(1+\sum_{j=1}^{3r} j^{-\gamma}\big)\Big\} \\
\text{(7)} \qquad &= O\big(t^2(M+1)^4r^2/n\big) = O\big(t^2(M+1)^4/n^{1-2\omega}\big),
\end{aligned}$$

where the last expression follows upon taking $r = n^{\omega}$. For the A_2 term we have from Bolthausen (1982) and Lemma 4

$$\begin{aligned}
E\big|A_2\big| &\le \frac{nt^2(M+1)}{\alpha_n^{3/2}}\max_i \sum_{j,k=1, d(i,j)\le r, d(i,k)\le r}^{n}\mathrm{Cov}(Y_j, Y_k) \\
\text{(8)} \qquad &\le \frac{c_4nt^2(M+1)^3}{\alpha_n^{3/2}}(2r+1)\Big\{1 + 2\sum_{j=1}^{2r+1} j^{-\gamma}\Big\} = O\big(t^2(M+1)^3/n^{\frac{1}{2}-\omega}\big).
\end{aligned}$$

Finally, we need to consider $A_3 = \alpha_n^{-1/2}\sum_{i=1}^n Y_i \exp\big\{it(\bar{S}_n - \bar{S}_{i,n})\big\}$, where $\bar{S}_{i,n} = S_{i,n}/\sqrt{\alpha_n}$. Considering the exponential part, and replacing all the variables by the approximating variables Y_j^m, we see that

$$E\big|\exp\big\{it(\bar{S}_n - \bar{S}_{i,n})\big\} - \exp\big\{it(\bar{S}_n^m - \bar{S}_{i,n}^m)\big\}\big| \le \frac{|t|}{\sqrt{\alpha_n}}E\big|(\tilde{S}_n - S_{i,n}) - (\tilde{S}_n^m - S_{i,n}^m)\big| \le \frac{|t|}{\sqrt{\alpha_n}}nc_2m^{-\theta}.$$

Using this we obtain

$$\big|\mathrm{Cov}\big(Y_i, \exp\{it(\bar{S}_n - \bar{S}_{i,n})\}\big) - \mathrm{Cov}\big(Y_i^m, \exp\{it(\bar{S}_n^m - \bar{S}_{i,n}^m)\}\big)\big| \le c_2m^{-\theta} + (M+1)\frac{|t|}{\sqrt{\alpha_n}}nc_2m^{-\theta}.$$

From the mixing we have (Ibragimov and Linnik, 1971 [17.2.1])

$$\big|\mathrm{Cov}\big(Y_i^m, \exp\{it(\bar{S}_n^m - \bar{S}_{i,n}^m)\}\big)\big| \le 4(M+1)c_1(2m+1)^{\delta_1}n^{\delta_2}(r-2m)^{-\theta}.$$

Taking $m = r/3$ and combining the two bounds we find for some constant c_5

$$\big|\mathrm{Cov}\big(Y_i, \exp\{it(\bar{S}_n - \bar{S}_{i,n})\}\big)\big| \le c_5\big\{(M+1)n^{\delta_2}r^{-\theta+\delta_1} + r^{-\theta}[1 + (M+1)|t|n/\sqrt{\alpha_n}]\big\}.$$

Since $EY_i = 0$ the terms in A_3 are covariances, and taking $r = n^{\omega}$ we therefore have the bound

$$\text{(9)} \qquad \big|EA_3\big| = O\big((M+1)[n^{-\omega\theta+1} + n^{-\omega(\theta-\delta_1)+\delta_2+\frac{1}{2}}]\big).$$

Thus, if we choose ω such that $\omega < \frac{1}{2}$, $\omega\theta - 1 > 0$ and $\omega(\theta - \delta_1) - \delta_2 - \frac{1}{2} > 0$, which is possible from the assumption on θ in Theorem 1, we see that all of (7), (8), and (9) tend to zero.

REFERENCES

Bolthausen, E. (1982): On the central limit theorem for stationary mixing random variables. *Ann. Probab.*, **10**, 1047–1050.

Deo, C. M. (1973): A note on empirical processes of strong-mixing sequences. *Ann. Probab.*, **1**, 870–875.

Ibragimov, I. A. and Linnik, Yu. V. (1971): *Independent and Stationary Sequences of Random Variables.* Wolters-Noordhoff Publishing, Groningen.

Jensen, J.L. (2005): Context dependent DNA evolutionary models. Research Report, No. 458, Department of Mathematical Sciences, University of Aarhus.