The Korean Journal of Applied Statistics (2010)
23(1), 73–79

Nonstationary Time Series and Missing Data

Dong Wan Shin[1] · Oesook Lee[2]

[1]Department of Statistics, Ewha Woman's University
[2]Department of Statistics, Ewha Woman's University

(Received December 2009; accepted January 2010)

Abstract

Missing values for unit root processes are imputed by the most recent observations. Treating the imputed observations as if they are complete ones, semiparametric unit root tests are extended to missing value situations. Also, an invariance principle for the partial sum process of the imputed observations is established under some mild conditions, which shows that the extended tests have the same limiting null distributions as those based on complete observations. The proposed tests are illustrated by analyzing an unequally spaced real data set.

Keywords: High frequency data, invariance principle, missing value imputation, semiparametric unit root test.

1. Introduction

Some finance time series are observed at high frequencies such as day or transaction-by-transaction, see Tsay (2005, Chapter 5). For daily data sets, observations are not available on weekends and holidays. For transaction-by-transaction data sets, transactions occur at irregular times. These situations produce missing observations. Also, some macro economic data sets contain missing values due to changes of sampling frequencies, for example, from quarterly sampling to monthly sampling.

We focus our interest on nonstationary time series data sets, which are very common for economic and financial time series. In the literature, nonstationarity for economic time series is extensively discussed in the context of unit root, see Fuller (1996, Chapter 10) and many others. We note that a basic probability model for modern finance analysis is Brownian motion, see Shreve (2004) and many others.

Statistical theories and methods for nonstationary time series data are largely based on the invariance principle that the partial sum processes converge in distribution to Brownian motions as the series length increases to infinity. If missing data occur, we need a proper imputation method in which the imputed process enjoy the invariance principle.

This research is supported by the Korea Research Foundation grant funded by the Korean Government (MOEHRD, Basic Research Promotion Fund)(KRF-2007-312-C00114).
[1]Corresponding author: Professor, Department of Statistics, Ewha Woman's University, Seoul 120-750, Korea. E-mail: shindw@ewha.ac.kr

In this paper, we study a simple imputation method under which missing values are imputed by the most recent observations. The partial sum process of the imputed data has the same limiting Brownian motion as that of the (unavailable) complete data. This implies that the long-run variances of the imputed data set and complete data set are the same.

Therefore, statistical methods based on the invariance principle and developed for complete data sets can also be applied to the imputed data set treating the imputed data sets as if complete data sets. Methods for financial engineering such as volatility estimation for Brownian motion are still valid for the imputed data sets. Unit root tests can extend to missing situations. Related study is by Shin and Sarkar (1996, 1998) who constructed unit root tests adopting parametric ARMA error structures. In this note, we impose no parametric structure for the error process and develop semiparametric unit root tests. The proposed tests are applied to a real data set, a transaction-by-transaction data set for IBM stock prices.

2. Imputed Process

In this section, we discuss imputation, invariance principle, and estimation of the long-run variance. Let x_t be a time series specified by $x_t = x_{t-1} + u_t$, where u_t is a zero-mean processes with finite variance $\sigma_u^2 > 0$. The error process u_t is assumed to be an α-mixing which is more general than stationary ARMA processes. Assume that the long-run variance $\sigma^2 = \lim_{T\to\infty} T^{-1}\mathrm{var}(\sum_{t=1}^{T} u_t) > 0$ exists. According to an invariance principle, under some conditions such as C1 and C2 below, as $T \to \infty$, we have

$$T^{-\frac{1}{2}} x_{[Tr]} \xrightarrow{d} \sigma W(r), \tag{2.1}$$

where $[Tr]$ is the integer part of $[Tr], 0 \le r \le 1$, $\xrightarrow{d}$ denotes convergence in distribution, and $W(r)$ is a standard Brownian motion.

Assume that observations are subject to missing so that x_t are available for time points $1 = t_1 < t_2 < \cdots < t_{n-1} < t_n = T$. Therefore, the set of observations is $\{x_{t_k}, k = 1, 2, \ldots, n\}$. For each k, the missing values x_t for $t \in \{t_k + 1, \ldots, t_{k+1} - 1\}$ are imputed by the most recent observation x_{t_k}, $k = 1, \ldots, n-1$. Therefore, the imputed values are

$$x^*_{t_k+j} = x_{t_k}, \quad j = 0, 1, \ldots, \Delta_k - 1, \ k = 1, \ldots, n-1,$$

where $\Delta_k = t_{k+1} - t_k$, $k = 1, \ldots, n-1$. This imputation is optimal if u_t is a white noise process. We note that this imputation is widely used in practice. For example, many asset prices reported at regular times are in fact imputed prices, by the prices of the most recent transactions. Note that x_t^* is the observation if x_t is observed and x_t^* is the imputed value if x_t is missing.

We can write x_t^* as an integrated process

$$x_t^* = x_{t-1}^* + u_t^*$$

if we define u_t^* as follows

$$\begin{aligned}
&u^*_{t_1} = u_{t_1}; \quad u^*_j = 0, \quad j = t_1 + 1, \ldots, t_2 - 1;\\
&u^*_{t_2} = u_{t_1+1} + \cdots + u_{t_2}; \quad u^*_j = 0, \quad j = t_2 + 1, \ldots, t_3 - 1;\\
&\vdots\\
&u^*_{t_n} = u_{t_{n-1}+1} + \cdots + u_{t_n}.
\end{aligned}$$

Therefore, under some mild conditions listed below we have an invariance principle for $T^{-1/2}x^*_{[Tr]}$ as stated in Theorem 2.1 below.

C1. u_t is a zero mean process with a positive long-run variance σ^2 and satisfies $\sup_t E|u_t|^{2+\delta} < \infty$ for some $\delta > 0$,

C2. u_t is an α-mixing with α-mixing coefficient $\alpha(m) = O(m^{-\lambda})$ for some $\lambda > (2+\delta)/\delta$,

C3. letting $\Delta = \max_{1\le k\le n} \Delta_k$, we have $E(\Delta) = o(T^{\delta/2(3+\delta)})$.

C4. we assume MCAR(Missing Completely At Random) for the missing mechanism.

Note that C2 implies $\sum_m \alpha(m)^{1-2/(2+\delta)} = O(\sum_m m^{-\lambda\delta/(2+\delta)}) = O(1)$ because $\lambda\delta/(2+\delta) > 1$. This, together with C1, guarantees the invariance principle of Herrndorf (1984) for $T^{-1/2}x_{[Tr]}$. The conditions C1 and C3 ensure that $\max_{1\le t\le T} |x^*_t - x_t| = o_p(T^{1/2})$ as shown in the proof of Theorem 2.1, leading to the invariance principle for $x^*_T(r)$ below.

We observe that $E(\Delta)$ need not to be bounded. If u_t has uniformly bounded high moments with large δ, then $E(\Delta)$ can be large without violating C3. If, for example, $\delta > 6$, then C3 becomes $E(\Delta) = o(T^{1/3})$. Therefore, in such case, if we collect observations with sampling intervals satisfying $E(\Delta) = o(T^{1/3})$, the imputed process still satisfy the invariance principle. Note that as $\delta \to \infty$, $T^{\delta/2(3+\delta)}$ increase to $T^{1/2}$. Therefore, if $\delta = \infty$ as in normal case, Δ with $E(\Delta) = O(T^{1/2-\epsilon})$ for some $\epsilon > 0$ satisfies C3.

According to C4, the probabilistic structure generating missing values is independent of the process $\{X_t\}$. This would be the usual case of the missing situations of multiple sampling frequency data sets and high frequency data sets discussed in Section 1. See Little and Rubin (2002) for more about MCAR.

Theorem 2.1. *Under C1∼C4, as* $T \to \infty$, $T^{-1/2}x^*_{[Tr]} \xrightarrow{d} \sigma W(r)$.

Shin (2008) showed that the long-run variance σ^2 can be consistently estimated from the imputed data. Let

$$\hat{\sigma}^{*2}_{T\ell} = T^{-1}\sum_{t=1}^{T} u_t^{*2} + 2T^{-1}\sum_{\tau=1}^{\ell}\sum_{t=\tau+1}^{T} u^*_t u^*_{t-\tau},$$

where ℓ is a given nonnegative integer called bandwidth. Theorem 2.2 below establishes consistency of $\hat{\sigma}^{*2}_{T\ell}$, for which we need more conditions than C1∼C4. The conditions are that u_t has bounded $2(2+\delta)$ moment for some $\delta > 0$, an order condition is imposed on Δ, and a rate condition is imposed on ℓ. These conditions are not binding ones for practical use.

Theorem 2.2. *Under some regularity conditions, as* $T \to \infty$, $\hat{\sigma}^{*2}_{T\ell} \to \sigma^2$ *in probability.*

The estimator $\hat{\sigma}^{*2}_{T\ell}$ is not always nonnegative. Bartlett modification

$$\bar{\sigma}^{*2}_{T\ell} = T^{-1}\sum_{t=1}^{T} u_t^{*2} + 2T^{-1}\sum_{\tau=1}^{\ell}\left(1-\frac{\tau}{\ell}\right)\sum_{t=\tau+1}^{T} u^*_t u^*_{t-\tau}$$

is always nonnegative. Note that the estimators $\hat{\sigma}^{*2}_{T\ell}$ and $\bar{\sigma}^{*2}_{T\ell}$ are constructed from the imputed data set $\{x^*_t, t = 1,\dots,T\}$ in the same way as the usual complete data estimators $\hat{\sigma}^2_{T\ell} = T^{-1}\sum_{t=1}^{T} u_t^2 + 2T^{-1}\sum_{\tau=1}^{\ell}\sum_{t=\tau+1}^{T} u_t u_{t-\tau}$ and $\bar{\sigma}^2_{T\ell} = T^{-1}\sum_{t=1}^{T} u_t^2 + 2T^{-1}\sum_{\tau=1}^{\ell}(1-\tau/\ell)\sum_{t=\tau+1}^{T} u_t u_{t-\tau}$ are

constructed from the complete data set $\{x_t, t = 1,\ldots,T\}$. Therefore, together with Theorem 2.1, statistical methods for unit root processes such as unit root test, cointegration test, and regression for integrated time series can extend to missing data situations. In the following section, we choose unit root test for more detailed investigation.

3. Applications to Unit Root Test

We apply the results of Section 2 to construct semiparametric unit root tests. Two cases are considered. The first case tests the null hypothesis of unit root and the second case tests the alternative hypothesis of unit root.

Consider a mean model, $x_t = \mu + \rho x_{t-1} + u_t$, where u_t is an error process satisfying C1~C3. We are interested in testing the null hypothesis of nonstationarity $H_0 : \rho = 1$. The semiparametric tests of Phillips (1987) and Phillips and Perron (1988) allow simple extensions to missing data cases. An estimator of ρ is $\hat{\rho} = (\sum_{t=2}^{T} \tilde{x}_{t-1}^{*2})^{-1}(\sum_{t=2}^{T} \tilde{x}_{t-1}^{*}\tilde{x}_t^{*})$, where $\tilde{x}_t^{*} = x_t^{*} - \bar{x}^{*}$ is the demeaned process and $\bar{x}^{*} = T^{-1}\sum_{t=1}^{T} x_t^{*}$. According to Theorem 2.1, as in Phillips (1987) for complete data case, under H_0, as $T \to \infty$,

$$T(\hat{\rho} - 1) \xrightarrow{d} \left\{ \int_0^1 \tilde{W}(r)dW(r) + 0.5\frac{\sigma^2 - \sigma_u^{*2}}{\sigma^2} \right\} \Big/ \int_0^1 \tilde{W}^2(r)dr$$

and

$$t_{\hat{\rho}} = \frac{\hat{\rho} - 1}{\mathrm{se}(\hat{\rho})} \xrightarrow{d} \left\{ \int_0^1 \tilde{W}dW + 0.5\frac{\sigma^2 - \sigma_u^{*2}}{\sigma^2} \right\} \Big/ \left\{ \int_0^1 \tilde{W}^2(r)dr \right\}^{\frac{1}{2}},$$

where

$$\tilde{W}(r) = W(r) - \int_0^1 W(s)ds, \qquad \mathrm{se}(\hat{\rho}) = \left(\hat{\sigma}_u^{*2} \Big/ \sum_{t=2}^{T} \tilde{x}_{t-1}^{*} \right)^{\frac{1}{2}},$$

$$\sigma_u^{*2} = p\lim_{T\to\infty} T^{-1}\sum_{t=2}^{T} E(x_t^{*} - x_{t-1}^{*})^2, \qquad \hat{\sigma}_u^{*2} = T^{-1}\sum_{t=2}^{T}(x_t^{*} - x_{t-1}^{*})^2.$$

Applying the procedure of Phillips and Perron (1988) to the imputed data set, we construct the following Z-tests

$$Z(\hat{\rho}) = T(\hat{\rho} - 1) - \hat{\theta} \Big/ \left(T^{-1}\sum_{t=1}^{T} \tilde{x}_t^{*2} \right), \qquad Z(t_{\hat{\rho}}) = \left(\frac{\hat{\sigma}_u^{*}}{\hat{\sigma}_{T\ell}} \right) t_{\hat{\rho}} - \hat{\theta} \Big/ \left(\hat{\sigma}_{T\ell}^2 T^{-1} \sum_{t=1}^{T} \tilde{x}_t^{*2} \right)^{\frac{1}{2}},$$

where $\hat{\theta} = 0.5(\hat{\sigma}_{T\ell}^2 - \hat{\sigma}_u^{*2})$, where

$$\hat{\sigma}_{T\ell}^{*2} = T^{-1}\sum_{t=1}^{T}(x_t^{*} - x_{t-1}^{*})^2 + 2T^{-1}\sum_{\tau=1}^{\ell}\sum_{t=\tau+1}^{T}(x_t^{*} - x_{t-1}^{*})(x_{t-\tau}^{*} - x_{t-1-\tau}^{*}).$$

Thanks to Theorems 2.1, 2.2, the proposed tests have the standard limiting null distributions

$$Z(\hat{\rho}) \xrightarrow{d} \int_0^1 \tilde{W}dW \Big/ \int_0^1 \tilde{W}^2(r)dr, \qquad Z(t_{\hat{\rho}}) \xrightarrow{d} \int_0^1 \tilde{W}dW \Big/ \left\{ \int_0^1 W^2(r)dr \right\}^{\frac{1}{2}}.$$

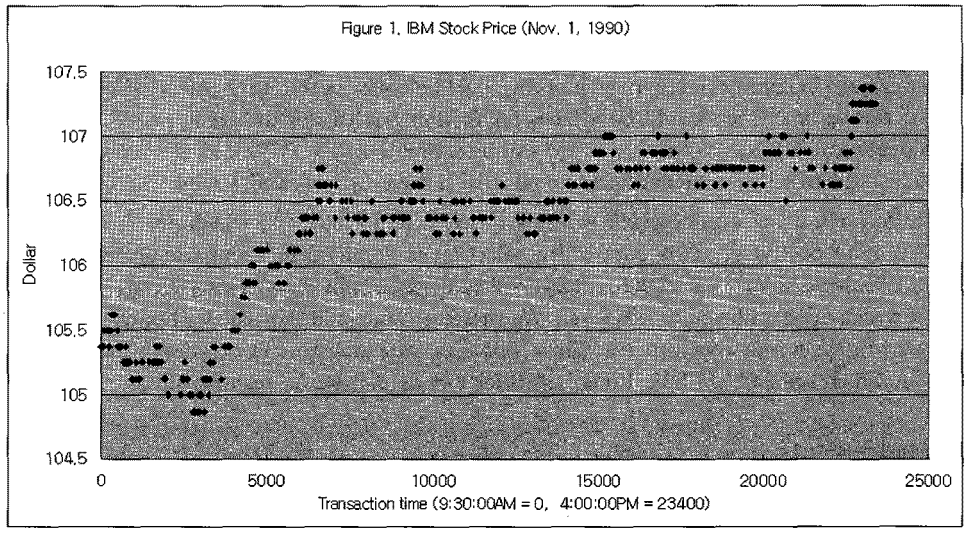

Figure 4.1. IBM stock price

It is obvious that the previous discussion in this section extends to the trend model $x_t = \mu + \beta t + \rho x_{t-1} + u_t$. The Z-statistics for the trend model constructed from the imputed data set have the same standard limiting null distribution as those constructed from the complete data set.

Consider, next, $y_t = v_t + z_t$, $v_t = v_{t-1} + w_t$, where z_t and w_t are independent zero-mean stationary processes with positive long-run variances with $\text{var}(w_t) = \sigma_w^2$. We are interested in testing the null hypothesis $H_0 : \sigma_w^2 = 0$ of stationarity of y_t against the alternative hypothesis $H_1 : \sigma_w^2 > 0$ of nonstationarity. Kwiatkowski *et al.* (1992) proposed a locally best invariant test given by $S = T^{-2} \sum_{t=1}^{T} x_t^2 / (T^{-1} \sum y_t^2)$, where $x_t = \sum_{s=1}^{t} y_s$. We extend the test to a situation in which some of y_t are missing. If y_t is missing, then it's imputed value is $y_t^* = 0$. Otherwise, $y_t^* = y_t$. Let $x_t^* = \sum_{s=1}^{t} y_s^*$. The test of Kwiatkowski *et al.* (1992), applied to the imputed observations, is

$$S = T^{-2} \sum_{t=1}^{T} \frac{x_t^{*2}}{\hat{\sigma}_{T\ell}^{*2}}$$

and has the standard limiting null distribution $\int_0^1 W^2(r)dr$, where

$$\hat{\sigma}_{T\ell}^{*2} = T^{-1} \sum_{t=1}^{T} y_t^{*2} + 2T^{-1} \sum_{\tau=1}^{\ell} \sum_{t=\tau+1}^{T} y_t^* y_{t-\tau}^*.$$

4. Example

We analyze an IBM stock price data set discussed in Tsay (2005, Chapter 5). The data set contains transaction times and transaction prices for all transactions of IBM stock for the period between November 1, 1990 and January 31, 1991. We select the data for the first day, November 1, 1990, which is depicted in Figure 4.1. Transaction began at 9:30AM and ended at 4:00 PM. The time unit is second. Time index varies from 0 to (3600sec/hour) $*$ (6.5hours/day) = (23400seconds/day).

Total number of transaction is 757. As seen in Figure 4.1, transactions occurred at unequally spaced time intervals. For example, the first four transactions occurred at $t = 1, 9, 10, 15$.

We are interested in testing whether the first day local behavior of the stock price, x_t say, is governed by a unit root or not. This would be a basis for further analysis of the stock price. For example, one-day local volatility $\sigma_d = \sqrt{\text{var}(x_T - x_0)}$ can be estimated by $\sqrt{T}\bar{\sigma}^*_{T\ell}$ if x_t is a unit root process. This issue is addressed by considering a model $x_t = \mu + \rho x_{t-1} + u_t$ and testing $H_0 : \rho = 1$.

We computed $\bar{\sigma}^{*2}_{T\ell}$ using 4 bandwidth values $\ell = L_k = [k(T/100)^{0.25}]$, which are 7, 15, 31, 46 for $k = 1, 2, 3, 4$, respectively. This bandwidth selection is common in semiparametric econometric analysis. We have $Z(t_{\hat{\rho}}) = -2.026, -2.251, -2.573, -2.767$ for $\ell = L_1, L_2, L_3, L_4$, respectively, neither of which are significant at 5% level. We therefore conclude that the within-day dynamic for IBM stock price is a unit root process. From this fact, we can estimate one-day local volatility σ_d by $\sqrt{T}\bar{\sigma}^*_{T\ell}$, which are 2.025, 1.823, 1.594, 1.483 for $\ell = L_1, L_2, L_3, L_4$ respectively.

Acknowledgements

The authors appreciate the comments of two referees.

Appendix: Proofs

Proof of Theorem 2.1. For each t, we can find $k \in \{1, \ldots, n\}$ and $h \in \{0, 1, \ldots, \Delta_k - 1\}$ such that $t = t_k + h$. Let $r_t = \sum_{j=1}^{h} u_{t_k+j}$. Then $x_t - x^*_t = r_t$. Because of (2.1), if we show $T^{-1/2} \max_{1\le t\le T} |r_t| \to 0$ in probability, then Theorem 4.1 of Billinsley (1968) with the metric $d(x_T, x^*_T) = \sup_{0\le r\le 1} |x_{[Tr]} - x^*_{[Tr]}| = \max_{1\le t\le T} |r_t|$ is applicable to yield the desired result. Noting that $\Delta_k \le \Delta$ for all k, we have

$$\max_{1\le t\le T} |r_t| = \max_{1\le k\le n} \max_{0\le h\le \Delta_k - 1} \left| \sum_{j=1}^{h} u_{t_k+j} \right| \le \max_{1\le k\le n} \max_{0\le h\le \Delta - 1} \left| \sum_{j=1}^{h} u_{t_k+j} \right| \le \max_{1\le t\le T} \max_{0\le h\le \Delta - 1} \left| \sum_{j=1}^{h} u_{t+j} \right|.$$

Let B the $\sigma-$algebra generated by $\{t_k, k = 1, 2, \ldots\}$. Let $\epsilon > 0$ be given. We have

$$\begin{aligned}
P\left[T^{-\frac{1}{2}} \max_{1\le t\le T} |r_t| > \epsilon \,\middle|\, B \right] &\le P\left[\max_{1\le t\le T} \max_{0\le h\le \Delta - 1} \left| \sum_{j=1}^{h} u_{t+j} \right| > T^{\frac{1}{2}}\epsilon \,\middle|\, B \right] \\
&\le \sum_{t=1}^{T} \sum_{h=0}^{\Delta-1} P\left[\left| \sum_{j=1}^{h} u_{t+j} \right| > T^{\frac{1}{2}}\epsilon \,\middle|\, B \right] \\
&= \sum_{t=1}^{T} \sum_{h=0}^{\Delta-1} P\left[\left| \sum_{j=1}^{h} u_{t+j} \right| > T^{\frac{1}{2}}\epsilon \right] \quad \text{by C4 of MCAR} \\
&\le \sum_{t=1}^{T} \sum_{h=0}^{\Delta-1} T^{-1-\frac{\delta}{2}} \epsilon^{-(2+\delta)} E \left| \sum_{j=1}^{h} u_{t+j} \right|^{2+\delta} \quad \text{by the Chebychev's inequality} \\
&\le T^{-1-\frac{\delta}{2}} \epsilon^{-(2+\delta)} \sum_{t=1}^{T} \sum_{h=0}^{\Delta-1} h^{1+\delta} \sum_{j=1}^{h} E|u_{t+j}|^{2+\delta} \quad \text{by the Holder inequality,} \\
&\le T^{-\frac{\delta}{2}} \epsilon^{-(2+\delta)} \sum_{h=0}^{\Delta-1} h^{2+\delta} \sup_t E|u_t|^{2+\delta}.
\end{aligned}$$

Taking expectation with respect to $\{t_k, k = 1, 2, \ldots\}$,

$$P\left[T^{-\frac{1}{2}} \max_{1\le t\le T} |r_t| > \epsilon\right] \le T^{-\frac{\delta}{2}} \epsilon^{-(2+\delta)} o\left(E\left(\Delta^{3+\delta}\right)\right) \sup_t E|u_t|^{2+\delta} = o(1), \quad \text{by C1 and C3,}$$

we get $T^{-1/2} \max_{1\le t\le T} |r_t| \to 0$ in probability.

Proof of Theorem 2.2. A proof is given in Shin (2008).

References

Billingsley, P. (1968). *Convergence of Probability Measures*, John Wiley & Sons, New York.

Fuller, W. A. (1996). *Introduction to Statistical Time Series*, 2nd ed. John Wiley & Sons, New York.

Herrndorf, N. (1984). A functional central limit theorem for weakly dependent sequences of random variables, *Annals of Probability*, **12**, 141–153.

Kwiatkowski, D., Phillips, P. C. B., Schmidt, P. and Shin, Y. (1992). Testing the null hypothesis of stationarity against the alternative of a unit root: How sure are we that economic time series have a unit root?, *Journal of Econometrics*, **54**, 159–178.

Little, R. J. and Rubin, D. B. (2002). *Statistical Analysis with Missing Data*, John Wiley & Sons, Hoboken, New Jersey, USA.

Phillips, P. C. B. (1987). Time series regression with a unit root, *Econometrica*, **55**, 277–301.

Phillips, P. C. B. and Perron, P. (1988). Testing for a unit root in time series regression, *Biometrika*, **75**, 335–346.

Shin, D. W. (2008). Estimation of the long-run covariance matrix for unequally spaced nonstationary time series, Unpublished manuscript, Ewha University, Korea.

Shin, D. W. and Sarkar, S. (1996). Testing for a unit root in an AR(1) time series using irregularly observed data, *Journal of Time Series Analysis*, **17**, 309–321.

Shin, D. W. and Sarkar, S. (1998). Testing for a unit root in autoregressive moving-average models with missing Data, *Journal of Time Series Analysis*, **19**, 601–608.

Shreve, S. E. (2004). *Stochastic Calculus for Finance II: Continuous-Time Models*, Springer.

Tsay, R. S. (2005). *Analysis of Financial Time Series*, 2nd ed., John Wiley & Sons, Hoboken, New Jersey.