Chapter 7

Cointegration Testing Using Pseudo Likelihood Ratio Tests

Chapters 4 through 6 dealt with the problem of constructing outlier robust, univariate unit root testing procedures. In Chapters 7 and 8, multivariate unit root tests, also known as cointegration tests, are studied.

The present chapter is set up as follows. Section 7.1 introduces the concept of cointegration and discusses the cointegration testing procedure of Johansen (1988, 1991). This section also comments on the possibilities of using outlier robust estimation procedures in order to construct outlier robust cointegration tests. Section 7.2 introduces the pseudo maximum likelihood (PML) estimation principle and discusses the classes of estimators and data generating processes that are used. In Section 7.3, the asymptotic distribution of a PML based cointegration test is derived. The relation between this new test and the one put forward by Johansen (1988) is discussed. Section 7.4 derives an optimality result for the choice of the pseudo likelihood. It turns out that power can be gained in situations with fat-tailed innovations if non-Gaussian PML estimators are used. In Section 7.5, a simple Bartlett-type correction factor is proposed for the PML based test. The corrected test is designed to have approximately the same critical values as the test of Johansen (1988). In Section 7.6, the results of a small simulation experiment are described, illustrating the performance of different PML based cointegration tests. Section 7.7 briefly discusses the problems of introducing deterministic regressors or additional nuisance parameters into the model. Chapter 8 deals in more detail with deterministic regressors in nonstationary, multivariate time series models. Finally, Section 7.8 contains some concluding remarks. The proofs of the theorems in this chapter can be found in Appendix 7.A.

7.1 Testing for Cointegration

Chapters 4 through 6 dealt with the topic of autoregressive unit root testing in a univariate framework. Consider the simple autoregressive model of order one,

$$\Delta y_t = \phi y_{t-1} + \varepsilon_t, \tag{7.1}$$

with Δ the first difference operator, $\Delta y_t = y_t - y_{t-1}$, and $\{\varepsilon_t\}$ and i.i.d. process. The tests developed so far in this thesis were concerned with the hypothesis $H_0: \ \phi = 0$. For $\phi = 0$, (7.1) describes a regression model in first differences, such that y_t can be rewritten as

$$y_t = y_0 + \sum_{s=1}^{t} \varepsilon_s. \tag{7.2}$$

(7.2) clearly demonstrates that if ε_t has a positive variance, the variable y_t is nonstationary. Using the terminology of Section 4.1, y_t in (7.2) is integrated of order one ($I(1)$), because y_t is nonstationary, while Δy_t is stationary.

One of the easiest ways to generalize the univariate unit root tests to the multivariate setting is to replace the univariate series y_t by a vector of k different series, $y_t = (y_{1t}, \ldots, y_{kt})^\top$. One then obtains the model $\Delta y_t = \Pi y_{t-1} + \varepsilon_t$, or

$$\begin{pmatrix} \Delta y_{1t} \\ \vdots \\ \Delta y_{kt} \end{pmatrix} = \Pi \begin{pmatrix} y_{1,t-1} \\ \vdots \\ y_{k,t-1} \end{pmatrix} + \begin{pmatrix} \varepsilon_{1t} \\ \vdots \\ \varepsilon_{kt} \end{pmatrix}, \tag{7.3}$$

with Π a $(k \times k)$ matrix of unknown regression parameters. Assume that $|I_k - (I_k + \Pi)z| = 0$ implies either $|z| > 1$ or $z = 1$, with $z \in \mathbb{C}$ and I_k denoting the unit matrix of order k. So, explosive processes and processes with complex unit roots or a root at minus one are excluded. Complex unit roots are encountered in the analysis of seasonal time series (see, e.g., Hylleberg et al. (1990) and Franses (1991, Chapter 4)). Because seasonal time series analysis is not the focus of this thesis, the intricacies of complex unit roots and roots of minus one are discarded. All methods in Chapters 7 and 8 can, however, be generalized towards the case with complex roots.

Just as in the univariate case, one can test whether the matrix Π in (7.3) is equal to zero. If $\Pi = 0$, then all the components of y_t are driven by different partial sum processes. Apart from the cases $|\Pi| \neq 0$ and $\Pi = 0$, one can have situations in which Π is singular, but nonzero ($|\Pi| = 0$, while $\Pi \neq 0$). This leads to several complications in the multivariate setting that were not encountered in the univariate context. The different components of y_t can now be integrated of order one, without the complete Π matrix being equal to zero.

Example 7.1 Let Π be equal to the $(k \times k)$ matrix

$$\Pi = \begin{pmatrix} -1 & 0 & \cdots & 0 & 1 \\ 0 & -1 & \cdots & 0 & 1 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & -1 & 1 \\ 0 & 0 & \cdots & 0 & 0 \end{pmatrix},$$

then (7.3) can be rewritten as

$$\left\{ \begin{array}{rcl} y_{jt} & = & y_{kt} + \varepsilon_{jt}, \quad \text{for } j \neq k, \\ \Delta y_{kt} & = & \varepsilon_{kt}. \end{array} \right. \tag{7.4}$$

△

From (7.4) one can easily see that all the components of y_t are nonstationary. There are, however, two additional properties of (7.4) that are important. First, there is only one partial sum process causing the nonstationarity in the elements of y_t, namely $\sum_{s=1}^{t} \varepsilon_{ks}$. This contrasts with the case $\Pi = 0$, for which there are k different random walks generating the nonstationarity in y_t. Second, there are $(k-1)$ linear combinations of the elements of y_t that are stationary, namely $y_{jt} - y_{kt}$ for $j \neq k$. This also contrasts with the case $\Pi = 0$, for which there are no linear combinations of the elements of y_t that are stationary. Both of these properties are dealt with in more detail, below.

Section 4.1 discussed the notion of integratedness of order d. A process $\{y_t\}$ is said to be integrated of order d if its dth order differences form a stationary process, while its $(d-1)$th order differences still form a nonstationary process. If one is confronted with a vector process $\{y_t\}$ that is integrated of order d, part of the nonstationarity may be common accross the different components of y_t. This was illustrated in Example 7.1. The fact that different variables may have a common source of nonstationarity led to the introduction of the concept of cointegration (see Engle and Granger (1987)). Assume that all the elements of the vector process $\{y_t\}$ are integrated of order[1] d, $y_{it} \sim I(d)$ for $i = 1, \ldots, k$. The elements of y_t are said to be cointegrated of order (d, b), $y_t \sim CI(d, b)$, with $0 < b \leq d$, if there exists a linear combination of the elements of y_t, $a^\top y_t$ with $a \in \mathbb{R}^k$ and $a \neq 0$, such that $a^\top y_t \sim I(d-b)$. The vector a is called the cointegrating vector. It is possible that there are several linearly independent cointegrating vectors for one process. In Example 7.1, $y_t \sim CI(1, 1)$ and there are $(k-1)$ linearly independent cointegrating vectors.

If variables are cointegrated, then one can construct linear combinations of those variables that are integrated of a smaller order than the original variables. In economics, the case that is mostly studied is that in which the original variables are $I(1)$ and some linear combination(s) are $I(0)$. Recently, also the $I(2)$ case has empirically been dealt with, see, e.g., Juselius (1995) and the

[1] In principle, the assumption that all the elements of y_t are integrated of order d can be replaced by the assumption that the elements of y_t are integrated of at most order d. For example, consider the case with two variables $y_{1t} \sim I(1)$ and $y_{2t} \sim I(0)$. Then the elements of $y_t = (y_{1t}, y_{2t})^\top$ are integrated of at most order one and there exists a linear combination of the elements of y_t, namely $0 \cdot y_{1t} + 1 \cdot y_{2t}$, that is stationary. So the cointegrating vector in this case is just the second column of the unit matrix of order two. It is a bit an abuse of language, however, to say that y_{1t} and y_{2t} are *co*integrated, because the '*co*-part' of *co*integration suggests that there is a *common* source of nonstationarity in *both* variables. Keeping in mind this problem with the terminology, all the test procedures that are discussed in this second part of the thesis can also be used in situations where the elements of the vector process $\{y_t\}$ are integrated of at most order d.

references cited therein. Typical examples of $I(2)$ series are monetary variables like the nominal money stock and prices. Note that if prices are $I(2)$, inflation rates are $I(1)$. Similarly, if the nominal money stock and prices have the same $I(2)$ trend, it might be the case that the real money stock is $I(1)$ instead of $I(2)$. In this thesis I restrict my attention to the simpler case of variables that are at most $I(1)$.

I will now show under which conditions the rank of the matrix Π in (7.3) coincides with the number of linearly independent cointegrating vectors for the process $\{y_t\}$. Assume that there are r linearly independent cointegrating vectors and let B denote a $(k \times r)$ matrix whose columns are equal to the cointegrating vectors. If r is equal to zero, the matrix B is undefined. The structure of the argument that the rank of Π is equal to r, is as follows. First, it is shown that Π can be written as $AB^\top$, with A a $(k \times r)$ matrix. Next, it is shown that A has rank r.[2]

Note that for $r > 0$ the matrix B has full column rank r. Let $B_\perp$ denote a $(k \times (k-r))$ matrix of full column rank, such that $B_\perp^\top B = 0$. Using these definitions, (7.3) can be rewritten as

$$\Delta y_t = \Pi B(B^\top B)^{-1}B^\top y_{t-1} + \Pi B_\perp(B_\perp^\top B_\perp)^{-1}B_\perp^\top y_{t-1} + \varepsilon_t. \tag{7.5}$$

From the assumption that the elements of y_t are at most $I(1)$, it follows that Δy_t is stationary. The stationarity of $B^\top y_{t-1}$ follows from the fact that the columns of B constitute all the linearly independent cointegrating vectors of the system. The stationarity of the ε_t process follows directly from the i.i.d. assumption for the error process. So the only term in (7.5) that exhibits $I(1)$ behavior is $\Pi B_\perp(B_\perp^\top B_\perp)^{-1}B_\perp^\top y_{t-1}$. Because the left-hand side of (7.5) is $I(0)$, it must follow that the right-hand side of (7.5) is also $I(0)$. This can only be the case if $\Pi B_\perp(B_\perp^\top B_\perp)^{-1}B_\perp^\top y_{t-1} \equiv 0$, or equivalently, if $\Pi B_\perp = 0$. This condition implies that Π can be written as $AB^\top$, where A is a $(k \times r)$ matrix. Therefore, (7.5) can be rewritten as

$$\Delta y_t = AB^\top y_{t-1} + \varepsilon_t. \tag{7.6}$$

It now remains to be shown that A has full column rank r. Premultiplying (7.6) by $B^\top$, one obtains

$$B^\top \Delta y_t = (B^\top A)B^\top y_{t-1} + B^\top \varepsilon_t. \tag{7.7}$$

Let the singular value decomposition of the matrix $(B^\top A)$ be given by $U\Sigma V^\top$, with U and V two orthogonal $(r \times r)$ matrices and Σ an $(r \times r)$ diagonal matrix

[2] Note that the proof of the equality of the rank of Π and the number of linearly independent cointegrating relations departs slightly from the argument one usually finds in the literature. Usually, one assumes that the rank of Π is r and then derives that (under certain conditions) there are r independent stationary linear relations (see, e.g., Johansen (1991)). Here, in contrast, it is assumed that the linear relations $B^\top y_t$ are stationary, from which it is derived that the rank of Π must be equal to r (under the same regularity conditions as in the usual approach).

containing the singular values of $B^\top A$. Define $z_t = U^\top B^\top y_t$, then (7.7) can be rewritten as

$$\Delta z_t = \Sigma V^\top U z_{t-1} + U^\top B^\top \varepsilon_t. \tag{7.8}$$

(7.8) implies that the diagonal elements of Σ must be strictly positive, otherwise the stationarity of $B^\top y_t$ is contradicted. Thus, the matrix $B^\top A$ has full rank, which implies two things. First, the matrix A has full column rank r, which establishes that the rank of Π is r. Second, the matrix $B_\perp^\top A_\perp$ has full rank, where $A_\perp$ is a $(k \times (k-r))$ matrix of full column rank, such that $A_\perp^\top A = 0$. The second of these two results provides a condition that is used in Granger's representation theorem (see Johansen (1991)). If the condition fails, one can show using (7.6) and (7.8) that y_t must be integrated of at least order two.

The main conclusion from the above paragraph is that there are two major conditions for the number of cointegrating vectors to be equal to the rank of the matrix Π. These conditions are that: 1. the roots of the equation $|I_k - (I_k + \Pi)z| = 0$ lie outside the unit circle or are equal to one; 2. $|B_\perp^\top A_\perp| \neq 0$, i.e., the elements of y_t are at most $I(1)$.

As a side result of the above derivations, one obtains the vector error[3] correction model (VECM) representation of the y_t process in (7.6). Here the variables $B^\top y_{t-1}$ are interpreted as long-run equilibrium relationships (see, e.g., Banerjee et al. (1993), Lütkepohl (1993, Section 11.1.2), and Hamilton (1994, Section 19.1)). The parameters in the matrix A are called the error correction parameters. The idea is that if the system is in equilibrium, i.e., $B^\top y_{t-1} \approx 0$, then the changes in the elements of y_t will be small. Alternatively, if the system is far out of equilibrium, the ECM induces a change in y_t towards the equilibrium relations $B^\top y_t$, at least if the elements of A have the correct signs.

Another way of looking at $CI(1,1)$ processes is by decomposing the y_t process into stationary and nonstationary components. Using Theorem 8.1 from Section 8.3, one obtains that under the two assumptions mentioned earlier,

$$y_t = y_0 + B_\perp (A_\perp^\top B_\perp)^{-1} A_\perp^\top \sum_{s=1}^{t} \varepsilon_s + S(L)(\varepsilon_t - \varepsilon_0), \tag{7.9}$$

with $S(L)$ a matrix polynomial in the lag operator L and $S(L)(\varepsilon_t - \varepsilon_0)$ a stationary process. The decomposition in (7.9) reveals that the elements of y_t are driven by the stationary process $S(L)(\varepsilon_t - \varepsilon_0)$ and the $(k-r)$-dimensional nonstationary partial sum processes $A_\perp^\top \sum_{s=1}^t \varepsilon_s$, which are called the common trends of the system. In Example 7.1 the common trend is given by $\sum_{s=1}^t \varepsilon_{ks}$. Cointegration is concerned with finding the linear combinations of the elements of y_t that eliminate the common trends. It is easily seen from (7.9) that the linear combinations $B^\top y_t$ are exactly the ones that satisfy this objective.

[3] Hendry recently advocates the use of the term *equilibrium* instead of *error* correction mechanism. The two terms have identical acronyms. I leave it to the reader to decide upon which of the two terminologies is best.

So far, only the relationship between the cointegrating vectors, the common trends, and the rank of the matrix Π have been discussed. Nothing has been said about the estimation of the unknown parameters in (7.3) nor about the determination of the rank of Π. There is a huge literature on the construction of statistics for determining the number of cointegrating relationships and the exact form of these relationships, see, e.g., Engle and Granger (1987), Phillips and Durlauf (1986), Johansen (1988, 1989, 1991), Park and Phillips (1988), Phillips (1988, 1991a), Boswijk (1992), Park (1992), Kleibergen and van Dijk (1994), and Stock (1994). In all of these procedures the ordinary least-squares (OLS) estimator plays an important role. Skimming the empirical literature for applications of cointegration testing procedures, one finds that the likelihood based testing procedure of Johansen (1988, 1991) is mostly used. The procedure of Johansen also forms the basis of the present chapter and is, therefore, explained in more detail below. The next chapter, instead, starts from the procedure put forward by Kleibergen and van Dijk (1994).

Johansen begins with the model

$$\Delta y_t = \Pi y_{t-1} + \Phi_1 \Delta y_{t-1} + \ldots + \Phi_p \Delta y_{t-p} + \varepsilon_t. \tag{7.10}$$

The multivariate process y_t is observed for $t = -p, \ldots, T$. In addition to the regressors present in (7.10), deterministic trend functions and seasonal dummies can be added. This is postponed until Chapter 8. Model (7.10) differs from (7.3) in that additional dynamics are incorporated. This turns out to be irrelevant for the asymptotic analysis (compare Remark 6.1). Johansen now proceeds by assuming that $\{\varepsilon_t\}$ is a Gaussian i.i.d. process with mean zero and covariance matrix Ω_{11}. Conditioning on $y_0, \ldots, y_{-p}$, one obtains the following (conditional) likelihood for (7.10):

$$\prod_{t=1}^{T} |2\pi\Omega_{11}|^{-1/2} \cdot \exp\left(-\frac{1}{2}(\Delta y_t - \Pi y_{t-1} - \Gamma Z_t)^\top \Omega_{11}^{-1}(\Delta y_t - \Pi y_{t-1} - \Gamma Z_t)\right), \tag{7.11}$$

where $\Gamma = (\Phi_1, \ldots, \Phi_p)$, and $Z_t^\top = (\Delta y_{t-1}^\top, \ldots, \Delta y_{t-p}^\top)$. Under the hypothesis that there are r cointegrating relationships, the matrix Π can be written as $\Pi = AB^\top$, with A and B denoting two $(k \times r)$ matrices of full column rank. Conditional on A and B, the maximum likelihood (ML) estimator for Γ from (7.11) is the OLS estimator

$$\hat{\Gamma} = \left(\sum_{t=1}^{T}(\Delta y_t - AB^\top y_{t-1})Z_t^\top\right)\left(\sum_{t=1}^{T} Z_t Z_t^\top\right)^{-1}.$$

Now define the vectors $R_{1t} = y_{t-1} - M_{1k}M_{kk}^{-1}Z_t$ and $R_{0t} = \Delta y_t - M_{1k}M_{kk}^{-1}Z_t$, with $M_{1k} = \sum_{t=1}^{T} y_{t-1}Z_t^\top$, $M_{0k} = \sum_{t=1}^{T} \Delta y_t Z_t^\top$, and $M_{kk} = \sum_{t=1}^{T} Z_t Z_t^\top$. One obtains that $\hat{\Gamma} = (M_{0k} - \Pi M_{1k})M_{kk}^{-1}$ and that R_{1t} and R_{0t} are the residuals from the regressions of y_{t-1} on Z_t and of Δy_t on Z_t, respectively. Next, conditional on B, the ML estimators for A and Ω_{11} from (7.11) are the OLS estimators

$$\hat{A} = \left(\sum_{t=1}^{T} R_{0t} R_{1t}^{\top} B \right) \left(\sum_{t=1}^{T} B^{\top} R_{1t} R_{1t}^{\top} B \right)^{-1},$$

and

$$\hat{\Omega}_{11} = T^{-1} \sum_{t=1}^{T} (R_{0t} - \hat{A} B^{\top} R_{1t})(R_{0t} - \hat{A} B^{\top} R_{1t})^{\top}.$$

Let $S_{00} = \sum_{t=1}^{T} R_{0t} R_{0t}^{\top}$, $S_{01} = \sum_{t=1}^{T} R_{0t} R_{1t}^{\top}$, $S_{10} = S_{01}^{\top}$, and $S_{11} = \sum_{t=1}^{T} R_{1t} R_{1t}^{\top}$, then the ML estimator for B has to minimize

$$|\hat{\Omega}_{11}| = |S_{00} - S_{01} B (B^{\top} S_{11} B)^{-1} B^{\top} S_{10}|,$$

which is equivalent to minimizing

$$|S_{00}| \cdot |B^{\top}(S_{11} - S_{10} S_{00}^{-1} S_{01}) B| / |B^{\top} S_{11} B|. \tag{7.12}$$

(7.12) is minimized by setting the columns of B equal to the eigenvectors corresponding to the r largest eigenvalues of the matrix $S_{11}^{-1/2} S_{10} S_{00}^{-1} S_{01} S_{11}^{-1/2}$. Note that B cannot be determined uniquely, as different normalizations can be chosen for the eigenvectors. For example, B and $B\Phi$ produce the same value of the objective funtion (7.12) if Φ is a nonsingular matrix. Stated differently, the matrix B is not identified. This is immediately clear from (7.11). Because the pair (A, B) produces the same matrix Π as the pair $(A\Phi^{-1}, B\Phi)$, these two parameter configurations also lead to identical values of the likelihood function. From this it follows that A and B are not identified and that only the column spaces of these variables can be determined from the data. One has to impose restrictions on the elements of either A or B (or both) in order to obtain estimates of the parameters. These restrictions can be arbitrary normalization restrictions or restrictions that are derived from economic theory. For the remainder of this chapter, it is assumed that no restrictions are available from economic theory, and that B is normalized such that $B^{\top} S_{11} B = I_r$.

Given the maximum likelihood estimators of all the unknown parameters, the maximum value of the log-likelihood under the hypothesis that the rank of Π is equal to r, is

$$-\frac{1}{2} \ln(|S_{00}|) - \frac{1}{2} \sum_{i=1}^{r} \ln(1 - \lambda_i), \tag{7.13}$$

with $\lambda_1 > \ldots > \lambda_k$ denoting the eigenvalues of the matrix $S_{11}^{-1/2} S_{10} S_{00}^{-1} S_{01} S_{11}^{-1/2}$. One can now construct the likelihood ratio test for the null hypothesis H_r : rank$(\Pi) = r$ against the alternative H_k : rank$(\Pi) = k$. Using (7.13), one obtains the test statistic

$$LR_r = -2T \sum_{i=r+1}^{k} \ln(1 - \lambda_i).$$

This statistic is known as the trace-test. Alternatively, one can test the null hypothesis H_r : rank$(\Pi) = r$ against the alternative H_{r+1} : rank$(\Pi) = r + 1$. This results in the test statistic $LR_r^{\max} = -2T \ln(1 - \lambda_{r+1})$, which is known as the maximum eigenvalue test. Johansen derived the asymptotic distributions of both tests and showed that they can be expressed in terms of functionals of Brownian motions (see also Section 7.3).

Both the theoretical foundation in the likelihood principle and the computational ease of the Johansen method have stimulated its widespread use in applied econometric work. As appears from the above discussion, the procedure of Johansen heavily relies on the Gaussian (pseudo) maximum likelihood ((P)ML) estimator. In Chapter 2, several disadvantages of the Gaussian PML estimator were mentioned. For example, the variance of the Gaussian PML estimator quickly increases as the disturbances become more heavy-tailed. Moreover, outliers and influential observations can have a large impact on the estimator. For both issues, see Huber (1981) and Hampel et al. (1986). The sensitivity of the Gaussian PML estimator has its effect on inference procedures that are based on it. In order to reduce the sensitivity of inference procedures, one can base them on other estimators, e.g., maximum likelihood type (M) estimators (see Huber (1981)) and pseudo maximum likelihood estimators (see White (1982) and Gouriéroux et al. (1984)). Such alternative estimators can be chosen such that they are less sensitive than Gaussian PML and, at the same time, have a reasonable efficiency if the errors are normally distributed. Moreover, some of these estimators outperform the Gaussian PML estimator in terms of efficiency if the errors in the model are nonnormal.

Also in the present context of cointegration testing, one can expect that tests based on the Gaussian PML estimator are more sensitive to outliers and fat-tailed innovations than tests based on non-Gaussian PML estimators. In particular, one can expect that the nonnormality of the error process can be exploited in order to improve the power properties of the cointegration testing procedure. For example, it is well-known that in a stationary context the maximum likelihood estimator is, in general, more efficient than a Gaussian PML estimator. Therefore, it is intuitively clear that maximum likelihood type estimators are more efficient if the hypothesis of integration (no cointegration) is slightly violated. This is formalized in the present chapter within the framework of nearly non-cointegrated time series (see Phillips (1988)).

The main objective of the present chapter is to develop a cointegration testing procedure based on pseudo maximum likelihood estimators (see Gouriéroux et al. (1984)) and to study the properties of this procedure by means of an asymptotic analysis and simulations. The considered test is a generalization of the trace test of Johansen (1988, 1991) presented below (7.13). It uses the ratio of two possibly non-Gaussian pseudo likelihoods. As mentioned above, the motivation for this approach is twofold. First, in many economic applications, e.g., in finance (see de Vries (1994)), the normality assumption for the error term is untenable. This leaves some room for improving the power properties of the cointegration test of Johansen. Second, dealing with outliers and influential observations in the data is a common feature of empirical econometric model building. A test that automatically corrects for some of these atypical observations seems a useful tool for the applied researcher.

This chapter only considers vector autoregressive (VAR) time series models. This means that attention is restricted to the parametric cointegration testing approach as adopted by Johansen (1988, 1991). This contrasts with the semiparametric approach of, e.g., Phillips (1987, 1988, 1991), Phillips and Durlauf (1986), and Park and Phillips (1988). An advantage of using the parametric approach is that one can easily construct tests that, like those of Johansen, are based on the (pseudo) likelihood ratio principle.

This chapter extends the literature in several ways. First, pseudo maximum likelihood estimators are used for testing the cointegration hypothesis. This leads to the construction of new test statistics. The relation of these statistics to the likelihood ratio test of Johansen is established. Second, the optimal choice of the pseudo likelihood and the score function is discussed in a multivariate, nearly nonstationary framework. In this way, the findings of Cox and Llatas (1991) are generalized. Third, simulation evidence is provided, illustrating that the new cointegration tests outperform the likelihood ratio test of Johansen if the innovations are fat-tailed. The notation used in this chapter was explained in Subsection 1.4.4.

7.2 Preliminaries

The central model in this chapter is the vector autoregressive (VAR) model of order $p+1$, given in (7.10). The error terms ε_t are assumed to satisfy the following conditions.

Assumption 7.1 *(i)* $\{\varepsilon_t\}_{t=0}^{\infty}$ *is an i.i.d. process with density function* $f(\varepsilon_t)$*; (ii)* $E(\varepsilon_0)=0$*; (iii)* $\Omega_{11}=E(\varepsilon_0\varepsilon_0^\top)$ *is finite and positive definite.*

Assumption 7.1 is stronger than the assumptions made in Phillips (1988). As a result, the asymptotic distribution of the cointegration test statistic has a simpler form. The requirement that the second moment of ε_t exists, can be dispensed with. The limiting distribution of the cointegration tests for innovations with infinite variance is probably distributed as a χ^2 random variable with r degrees of freedom or as a weighted sum of r independent $\chi^2(1)$ random variables, at least if the function ψ of Assumption 7.3 in Section 7.3 below is bounded (compare Knight (1989, 1990)). This statement, however, is not proved formally in this thesis.[4] Finally, the introduction of deterministic

[4] Although no formal proof is given, one can make the result intuitively clear. Assume that the first moment of ε_t exists, but that the second moment does not exist. Further assume that ψ is bounded. The (canonical) correlations between ε_t and $\psi(V^{1/2}\varepsilon_t)$ are then equal to zero. This follows from the observation that $E(\psi(\Omega_{11}^{-1/2}\varepsilon_t)\varepsilon_t^\top)$ and $E(\psi(\Omega_{11}^{-1/2}\varepsilon_t)\psi(\Omega_{11}^{-1/2}\varepsilon_t)^\top)$ are bounded, while $E(\varepsilon_t\varepsilon_t^\top)$ is infinite. As a result, the Brownian motion $\hat{W}_2$ and the stochastic process $\hat{U}$ in Theorem 7.1 are uncorrelated, giving rise to a χ^2 limiting result.

regressors in the model is not of central interest in this chapter and, therefore, delayed until Section 7.7 and Chapter 8.

Next, the two central restrictions mentioned in Section 7.1 are imposed on the model.

Assumption 7.2 *(i) The elements of y_t are integrated of at most order one; (ii) the equation*

$$|I_k - z(I_k + \Pi) - z(1-z)\Phi_1 - \ldots - z^p(1-z)\Phi_p| = 0$$

with $z \in \mathbb{C}$ has roots that satisfy either $|z| > 1$ or $z = 1$.

In order to determine the rank of the matrix Π, estimates are needed of the parameters in (7.10). Motivated by the arguments raised in Section 7.1, this chapter considers the class of Pseudo Maximum Likelihood (PML) estimators as an alternative to the Gaussian maximum likelihood estimator of Johansen (1988). Assume that the pseudo likelihood has the form

$$\mathcal{L}_T(\theta) \propto \prod_{t=1}^{T} |\Omega_{11}|^{-1/2} \cdot \exp\left(-\rho(\Omega_{11}^{-1/2} e_t)\right), \tag{7.14}$$

where $\rho(\cdot)$ is a function satisfying the regularity conditions of Assumption 7.3 in Section 7.3 below. This function can be interpreted as the negative of the (pseudo) log likelihood. The derivative of ρ with respect to the unknown parameters can, therefore, be interpreted as the (pseudo) score. The matrix Ω_{11} in (7.14) is a scaling matrix, $e_t = \Delta y_t - \Pi y_{t-1} - \Phi_1 \Delta y_{t-1} - \ldots - \Phi_p \Delta y_{t-p}$, and θ is the vector of unknown parameters. If the matrix Ω_{11} is not known, it can be estimated along with the parameters from (7.10). The pseudo likelihood may be improper in the sense that $\int \exp(-\rho(\Omega_{11}^{-1/2}\varepsilon_t))d\varepsilon_t$ need not exist. In this way, pseudo likelihoods with a redescending score function are also covered by the results in this paper. The PML estimator is given by the vector $\hat{\theta}_T$ that maximizes $\ell_T(\theta) = \ln(\mathcal{L}_T(\theta))$. Note that (7.14) comprises most likelihood functions that are used in the literature. The Gaussian maximum likelihood estimator of Johansen (1991), for example, is obtained by setting $\rho(e) = e^\top e/2$. Also the Student t maximum likelihood estimator, as discussed by Prucha and Kelejian (1984), and the class of maximum likelihood type (M) estimators (see, e.g., Huber (1981) and Hampel et al. (1986)) are contained as special cases of (7.14).

In Section 7.1, the two likelihood ratio based testing procedures of Johansen were presented, namely the trace test and the maximum eigenvalue test. This chapter only discusses the trace test. The results for the maximum eigenvalue test can be obtained using similar techniques as the ones employed here.

The hypotheses of interest concern the rank of the matrix Π. As this rank can range from zero to k, there are k hypotheses of interest. The rth hypothesis postulates that there are at most r cointegrating relationships, H_r : $\text{rank}(\Pi) \leq r$, with $r = 0, \ldots, k-1$. The alternative hypothesis in each case is H_k : $\text{rank}(\Pi) = k$.

Using the pseudo log likelihood $\ell(\theta)$, three testing principles can be employed, namely the likelihood ratio (LR), the Lagrange multiplier (LM), and the Wald principle. The Wald and LM testing principles are considered in the next chapter. Here, the focus is on the likelihood ratio principle for constructing a test statistic. If $\tilde{\theta}_{T,r}$ denotes the PML estimator under the null hypothesis, H_r, and $\hat{\theta}_T$ denotes the PML estimator under the alternative hypothesis, then the Pseudo Likelihood Ratio (PLR) test is given by

$$PLR_r = 2(\ell(\hat{\theta}_T) - \ell(\tilde{\theta}_{T,r})), \tag{7.15}$$

(compare White (1982)). A subscript r is added to the test statistic in order to indicate the null hypothesis that is tested. If no confusion is caused, this subscript is omitted. The limiting distribution of PLR_r is derived in the next section.

It was show in the previous section that under the null hypothesis H_r, Π can be written as $\Pi = AB^\top$, with A and B two $(k \times r)$ matrices of full column rank. One of the interesting questions in this chapter concerns the possibility of gaining power by exploiting the nonnormality of the error process in the estimation stage. This question cannot be addressed if one only considers the asymptotic distribution theory of the test statistic under the null hypothesis of no cointegration. Instead, (local) alternatives to the null hypothesis have to be considered. An adequate way of analysing the asymptotic distribution theory of the PLR test under local alternatives is given in Johansen (1989) and Rahbek (1994), who use of theory for nearly-nonstationary processes as presented in Phillips (1988). Following Johansen (1989), the local alternatives considered here are of the form

$$\Pi = AB^\top + T^{-1}A_1B_1^\top, \tag{7.16}$$

with A_1 and B_1 two $(k \times r_1)$ matrices of full column rank and $0 \leq r_1 \leq k - r$. The decomposition in (7.16) implies that there are r_1 additional cointegration vectors B_1, which enter model (7.10) with loadings that tend to zero as the sample size increases. The order of T^{-1} is necessary for obtaining a nondegenerate power function. It is assumed that the matrix B_1 satisfies $B^\top B_1 = 0$, such that the cointegrating vectors in B_1 are orthogonal to the cointegrating vectors in B. If $B_1^\top B \neq 0$, the matrix A in all subsequent derivations should be replaced by a matrix $A_T = A + O(T^{-1})$. This has no effect on the final results. Therefore, the assumption $B_1^\top B = 0$ is maintained throughout in order not to burden the notation more than is necessary.

7.3 Asymptotic Distribution Theory

In this section, the asymptotic distribution of the PLR test statistic (7.15) is discussed under the sequence of local alternatives (7.16). Apart from the conditions on the behavior of ε_t stated in Assumption 7.1, also some regularity conditions are needed for the function $\rho(\cdot)$ used in the definition of the pseudo likelihood (7.14). I assume the following conditions are satisfied.

Assumption 7.3 *(i) $\rho(\cdot)$ is twice continuously differentiable; the first and second order derivatives are denoted by $\psi(\Omega_{11}^{-1/2}\varepsilon_t) = \partial\rho(\Omega_{11}^{-1/2}\varepsilon_t)/\partial\varepsilon_t$ and $\psi'(\Omega_{11}^{-1/2}\varepsilon_t) = \partial\psi(\Omega_{11}^{-1/2}\varepsilon_t)/\partial\varepsilon_t^\top$, respectively; (ii) $\psi'(\Omega_{11}^{-1/2}\varepsilon_t)$ is first order Lipschitz; (iii) $E(\psi(\Omega_{11}^{-1/2}\varepsilon_0)) = 0$; (iv) $E(\psi'(\Omega_{11}^{-1/2}\varepsilon_0)) = C_1$, with C_1 positive definite; (v) the random vector $\psi(\Omega_{11}^{-1/2}\varepsilon_0) \otimes \varepsilon_0$ has finite second order moments; (vi) $E(\psi'(\Omega_{11}^{-1/2}\varepsilon_0) \otimes \varepsilon_0) = 0$.*

Parts (i) and (ii) of Assumption 7.3 impose some smoothness conditions on the pseudo likelihood. The conditions are somewhat stricter than necessary. Discontinuities in the function ψ can be handled if the density of ε_t is sufficiently smooth. This can be seen by comparing the results of Herce (1993) for the Least Absolute Deviations estimator with those of Chapter 6 for smooth M estimators. If allowance is made for discontinuities in, e.g., ψ, the methods of proof have to be changed considerably. Therefore, the attention here is restricted to smooth versions of ρ. Part (iii) of Assumption 7.3 is another centering condition in order to guarantee the consistency of the PML estimator. It is important to realize that this condition is nontrivial. It implies that more is known about the distribution of the ε_t process than just its mean and the finiteness of its variance. For the simple case that the distribution of ε_t is spherically symmetric, however, it suffices that the function ρ is spherically symmetric in order to meet part (iii) of the assumption. Part (iv) implies that the PML estimator can be approximated using a first order Taylor series expansion of the first order condition that defines the estimator. Part (v) is a moment condition. For the Gaussian PML estimator, it states that 4th order moments of the errors exist. The condition is somewhat too strict and is mainly used to facilitate the proofs of the theorems. It can, for example, be replaced by the conditions that $\psi(\Omega_{11}^{-1/2}\varepsilon_0)$ has finite second order moments and that $\partial\ell_T(\theta)/\partial\Omega_{11}$ has finite absolute moments of order $1+\eta$ for some $\eta > 0$. Note that (v) implies that the second order moment of $\psi(\Omega_{11}^{-1/2}\varepsilon_0)$ exists and is finite. Finally, part (vi) implies that we can abstract from the fact that Ω_{11} is estimated rather than known. If this part of Assumption 7.3 is not met, the limiting distribution of the *PLR* test changes. Note that (vi) is satisfied if both $\rho(\cdot)$ and $f(\cdot)$ are even, i.e., $f(\varepsilon_t) = f(-\varepsilon_t)$, and if the appropriate moments exist.

The limiting behavior of $(B_\perp^\top B_\perp)^{-1}B_\perp^\top y_{t-1}$ is presented in the next lemma, which follows directly from Phillips (1988), Johansen (1989), and Phillips and Durlauf (1986).

Lemma 7.1 *Given Assumptions 7.1 through 7.3, then*

$$T^{-1/2}\sum_{t=1}^{\lfloor sT\rfloor}(\varepsilon_t^\top, \psi(\Omega_{11}^{-1/2}\varepsilon_t)^\top) \Rightarrow (W_1(s)^\top, W_2(s)^\top),$$

with $(W_1(s)^\top, W_2(s)^\top)^\top$ a multivariate Brownian motion with covariance matrix

$$\Omega = \begin{pmatrix} \Omega_{11} & \Omega_{12} \\ \Omega_{21} & \Omega_{22} \end{pmatrix},$$

and $s \in [0,1]$*. Moreover, define* $C_2 = A_\perp^\top A_1 B_1^\top B_\perp$*,* $\Psi = \sum_{i=1}^{p+1} i\Pi_i$*,* $\Pi_1 = I_k + \Pi + \Phi_1$*,* $\Pi_{p+1} = -\Phi_p$*, and* $\Pi_i = \Phi_i - \Phi_{i-1}$ *for* $i = 2, \ldots, p$*, then*

$$T^{-1/2}(B_\perp^\top B_\perp)^{-1} B_\perp^\top y_{\lfloor sT \rfloor} \Rightarrow U(s),$$

where $U(s)$ *is the Ornstein-Uhlenbeck process that satisfies the stochastic differential equation*

$$(A_\perp^\top \Psi B_\perp) dU(s) = C_2 U(s) ds + dA_\perp^\top W_1(s).$$

In order to establish the limiting behavior of the PLR statistic, it turns out to be useful to define the matrix

$$S_0 = (A_\perp^\top \Omega_{11} A_\perp)^{-1/2} A_\perp^\top \Omega_{12} C_1^{-1} A_\perp (A_\perp^\top C_1^{-1} \Omega_{22} C_1^{-1} A_\perp)^{-1/2}, \tag{7.17}$$

which is the correlation matrix between $A_\perp^\top \varepsilon_t$ and $A_\perp^\top C_1^{-1} \psi(\Omega_{11}^{-1/2} \varepsilon_t)$. Let $S_1 R S_2^\top$ denote the singular value decomposition of S_0, with S_1 and S_2 two orthogonal matrices and R a diagonal matrix containing the absolute values of the canonical correlations between $A_\perp^\top \varepsilon_t$ and $A_\perp^\top C_1^{-1} \psi(\Omega_{11}^{-1/2} \varepsilon_t)$. Finally, let $\hat\varepsilon_t$ denote the residuals calculated at some estimate $\hat\theta$. Using these definitions, the limiting behavior of the PLR statistic can be established.

Theorem 7.1 *Let* $\hat\varepsilon_t$ *and* $\tilde\varepsilon_t$ *denote the residuals calculated at* $\hat\theta_T$ *and* $\tilde\theta_{r,T}$*, respectively. Given Assumptions 7.1 through 7.3 and* $\hat\varepsilon_t - \varepsilon_t = o_p(1)$ *and* $\tilde\varepsilon_t - \varepsilon_t = o_p(1)$ *uniformly in* t*, then* $PLR \Rightarrow \overline{PLR}$*, with*

$$\begin{aligned}
\overline{PLR} \;=\; & \operatorname{tr}\left(\tilde K_0 (\int \hat U d\hat W_2^\top)^\top (\int \hat U \hat U^\top)^{-1} (\int \hat U d\hat W_2^\top)\right) + \\
& 2 \cdot \operatorname{tr}\left(\tilde K_0 S_2^\top \bar K_0^{-1/2} (A_\perp^\top \Omega_{11} A_\perp)^{1/2} S_1 \tilde C_2 (\int \hat U d\hat W_2^\top)\right) + \\
& \operatorname{tr}\left(\tilde C_2^\top S_1^\top (A_\perp^\top \Omega_{11} A_\perp)^{1/2} K_0 (A_\perp^\top \Omega_{11} A_\perp)^{1/2} S_1 \tilde C_2 (\int \hat U \hat U^\top)\right),
\end{aligned}$$

where $\operatorname{tr}(\cdot)$ *is the trace operator,* $\hat U(s)$ *is an Ornstein-Uhlenbeck process satisfying the stochastic differential equation*

$$d\hat U(s) = \tilde C_2 \hat U(s) ds + d\hat W_1(s),$$

$\hat W_1(s)$ *and* $\hat W_2(s)$ *are two standard Brownian motions with diagonal correlation matrix* R *(defined below (7.17)), and*

$$\begin{aligned}
\tilde C_2 &= S_1^\top (A_\perp^\top \Omega_{11} A_\perp)^{-1/2} A_\perp^\top A_1 B_1^\top B_\perp (A_\perp^\top \Psi B_\perp)^{-1} (A_\perp^\top \Omega_{11} A_\perp)^{1/2} S_1, \\
K_0 &= (A_\perp^\top C_1^{-1} A_\perp)^{-1}, \\
\bar K_0 &= (A_\perp^\top C_1^{-1} \Omega_{22} C_1^{-1} A_\perp), \\
\tilde K_0 &= S_2^\top \bar K_0^{1/2} K_0 \bar K_0^{1/2} S_2.
\end{aligned}$$

As mentioned in Section 7.1, the proofs of all theorems and lemmas can be found in Appendix 7.A.

Remark 7.1 The additional two conditions in Theorem 7.1, $\hat{\varepsilon}_t - \varepsilon_t = o_p(1)$ and $\tilde{\varepsilon}_t - \varepsilon_t = o_p(1)$ uniformly in t, ensure that the correct optimum is chosen from the (possibly large) set of local optima. In fact, the conditions imply that $A, \Phi_1, \ldots, \Phi_p$, and Ω_{11} be consistently estimated, while the actual unit root parameters and cointegrating vectors (α_{22} and β in Appendix 7.A) are consistently estimated at a rate higher than $T^{1/2}$. Low-level conditions for consistency in a (possibly) nonlinear context can be found in, e.g., Gallant (1987).

It is illustrative to consider the two main differences between the result of Theorem 7.1 and the result of Johansen (1989). Johansen assumes Gaussian error terms, thus imposing $\psi(\Omega_{11}^{-1/2}\varepsilon_t) = \Omega_{11}^{-1}\varepsilon_t$. It then follows that the Brownian motion $W_2(s)$ is a linear transformation of the Brownian motion $W_1(s)$, namely $W_2(s) = \Omega_{11}^{-1}W_1(s)$. This implies that $\hat{W}_2(s)$ and $\hat{W}_1(s)$ are perfectly correlated in the sense that $R = I_{k-r}$. If a different specification is chosen for ψ, $\hat{W}_1$ and $\hat{W}_2$ are imperfectly correlated, which complicates the asymptotic distribution of the PLR test.

In order to explicate the difference between the general PML estimator and the Gaussian one, define the Brownian motion $\hat{W}_3(s) = \hat{W}_2(s) - R\hat{W}_1(s)$. All stochastic integrals of the form $\int \hat{U} d\hat{W}_2^\top$ can now be split into two parts. The first part equals $\int \hat{U} d\hat{W}_1^\top R$, which is a Gaussian functional (see Phillips (1991a)). The second part is $\int \hat{U} d\hat{W}_3(s)$, which is mixed normally distributed. One obtains the following theorem.

Theorem 7.2 *Define the matrices* $K_1 = \int \hat{U} d\hat{W}_1^\top$ *and* $K_2 = K_1^\top K_3^{-1} K_1$, *with* $K_3 = \int \hat{U}\hat{U}^\top$. *Then under the assumptions of Theorem 7.1,*

$$
\begin{aligned}
PLR \;\Rightarrow\; & \operatorname{tr}\left(\tilde{C}_2^\top S_1^\top (A_\perp^\top \Omega_{11} A_\perp)^{1/2} K_0 (A_\perp^\top \Omega_{11} A_\perp)^{1/2} S_1 \tilde{C}_2 K_3\right) + \\
& 2 \cdot \operatorname{tr}\left(\tilde{K}_0 S_2^\top \bar{K}_0^{-1/2} (A_\perp^\top \Omega_{11} A_\perp)^{1/2} S_1 \tilde{C}_2 \left(\int \hat{U} dW_3^\top + K_1 R\right)\right) + \\
& \operatorname{tr}\left(\tilde{K}_0 R K_2 R + 2\tilde{K}_0 \left(\int \hat{U} d\hat{W}_3^\top\right) K_3^{-1} K_1 R\right) + \\
& \operatorname{tr}\left(\tilde{K}_0 \left(\int \hat{U} d\hat{W}_3^\top\right)^\top K_3^{-1} \left(\int \hat{U} d\hat{W}_3^\top\right)\right). \qquad (7.18)
\end{aligned}
$$

By imposing the Gaussian specification of ψ as in Johansen (1989), Theorem 7.2 has the following obvious corollary.

Corollary 7.1 *Let the assumptions of Theorem 7.1 be satisfied. If* $\psi(\Omega_{11}^{-1/2}\varepsilon_t) = \Omega_{11}^{-1}\varepsilon_t$, *then* S_1 *can be chosen equal to* I_{k-r} *and*

$$
PLR \Rightarrow \operatorname{tr}\left(K_2 + 2\tilde{C}_2 K_1 + \tilde{C}_2 K_3 \tilde{C}_2^\top\right).
$$

Theorem 7.2 reveals that the PLR statistic depends in several ways on nuisance parameters. Consider the asymptotic distribution of the PLR statistic under the null hypothesis, so $A_1B_1^\top = 0$. Then from Theorem 7.2, one can distinguish two sources of dependence. First, the asymptotic distribution is influenced by the premultiplication with the matrix $\tilde{K}_0$ in each of the terms of (7.18) that do not involve $\tilde{C}_2$. This source of dependence is due to the discrepancy between the pseudo and the true likelihood. As was already noted in White (1982), misspecification of the likelihood causes a breakdown of the information matrix equality. In the present setting this means that if the pseudo likelihood does not coincide with the true likelihood, $C_1 \neq \Omega_{22}$. Therefore, the matrix $\tilde{K}_0$ can be eliminated either by correctly specifying the likelihood or by using the Gaussian PML estimator. The second source of dependence enters through the presence of the matrix R in (7.18). This dependence is due to the use of a non-Gaussian PML estimator. It is interesting to note that both types of dependence disappear if one sets $\psi(\Omega_{11}^{-1/2}\varepsilon_t) = \Omega_{11}^{-1}\varepsilon_t$, see Corollary 7.1.

As a side result of Theorem 7.2 one obtains that nuisance parameters remain present in the limiting distribution of PLR, even if the pseudo likelihood coincides with the actual likelihood. This result is presented in the following corollary.

Corollary 7.2 *If ε_t has density $f(\varepsilon_t) = c|\Omega_{11}|^{-1/2}\exp(-\rho(\Omega_{11}^{-1/2}\varepsilon_t))$, where c is such that $\int f(\varepsilon_t)d\varepsilon_t = 1$, and if $A_1B_1^\top = 0$, then*

$$PLR \Rightarrow \operatorname{tr}\left((\int \hat{W}_1 d\hat{W}_2^\top)^\top (\int \hat{W}_1\hat{W}_1^\top)^{-1}(\int \hat{W}_1 d\hat{W}_2^\top)\right),$$

with $E(\hat{W}_1(s)\hat{W}_2(s)^\top) = sR$.

Corollary 7.2 states that if the pseudo likelihood is correctly specified, then the only nuisance parameters that enter the asymptotic distribution of the PLR test are the canonical correlations between $A_\perp^\top\varepsilon_t$ and $A_\perp^\top C_1^{-1}\psi(\Omega_{11}^{-1/2}\varepsilon_t)$. Under the conditions of Corollary 7.1, this correlation is perfect. In most other circumstances, however, the correlation is imperfect, which results in a more complicated asymptotic distribution of the test statistic.

Corollary 7.2 can be used to simulate critical values of LR cointegration tests for correctly specified non-Gaussian likelihoods. A procedure for obtaining consistent estimates of these values is fairly straightforward. For given parameter estimates, the matrix S_0 in (7.17) can be consistently estimated: replace Ω_{11} by $T^{-1}\sum_{t=1}^T \hat{\varepsilon}_t\hat{\varepsilon}_t^\top$, C_1 by $T^{-1}\sum_{t=1}^T \psi'(\hat{\Omega}_{11}^{-1/2}\hat{\varepsilon}_t)$, etc., with $\hat{\varepsilon}_t$ denoting the tth regression residual. The estimate of S_0 can then be used to estimate the canonical correlations R by means of a singular value decomposition. Let $\hat{R}$ denote a diagonal matrix containing the estimated singular values of S_0. Then the critical values of $\overline{PLR}$ can be simulated in the usual way by generating random walks $\hat{W}_{1,T}$ and $\hat{W}_{2,T}$ of length T with correlation matrix $\hat{R}$, and replacing the integrals in Corollary 7.2 by sums. Note that this methodology can be extended to simulate the critical values of the PLR test for misspecified pseudo likelihoods. In that case, also a consistent estimate of $\tilde{K}_0$ is needed. Such an estimate can be constructed in the same way as described above using the residuals $\hat{\varepsilon}_t$.

The above procedure for computing critical values has two major drawbacks. First, critical values have to be simulated for every estimate of R. This might prove too time consuming for useful practical purposes. Second and more important, the critical values of $\overline{PLR}$ provide poor approximations to the critical values of the PLR test in finite samples, see Chapter 8. Therefore, Section 7.6 simulates the values of the PLR test directly in order to obtain critical values. Moreover, Section 7.5 proposes some simple strategies to correct the PLR test, such that simulations are not needed altogether, and standard available tables can be used.

7.4 The Choice of the Pseudo Likelihood

In order to use a PML estimator, one has to specify the function ρ in (7.14). Different objectives lead to different choices of the pseudo likelihood. In this section, the optimal choice of ρ is investigated if the criterion is to minimize $E(\overline{PLR})$, where $\overline{PLR}$ is defined in Theorem 7.1. This criterion is intimately linked to the minimum (asymptotic) mean squared error (MSE) criterion of Cox and Llatas (1991). The MSE considered is that of the estimator for the restricted elements of the matrix Π in (7.10), i.e., the actual unit root parameters α_{22} in Appendix 7.A. This is seen by looking at the proof of Theorem 7.1. For example, when testing $H_0: \ r = 0$, versus $H_k: \ r = k$, the relevant MSE is that of the estimator for the complete matrix Π. In the stationary context, minimization of $E(\overline{PLR})$ produces the true maximum likelihood estimator. Therefore, the criterion also seems natural for guiding the choice of the pseudo likelihood in a setting with nonstationary variables. Moreover, as in the stationary context, one can expect the minimum MSE of the estimators for the unit root parameters to be translated in a better power behavior of the cointegration tests based on these estimators.

This section produces two major conclusions. First, given the objective function stated above, the optimal pseudo score function ψ^* is a linear combination of the score function of the Gaussian PML estimator and the score function of the true ML estimator. This contrasts with the results one obtains in the stationary setting. Second, only part of the pseudo likelihood is identified if one uses the above criterion.

Before stating the main theorem of this section, the following additional assumption is introduced.

Assumption 7.4 *The density function $f(\varepsilon_t)$ is twice continuously differentiable with respect to its argument and vanishes on the edge of its support.*

Assumption 7.4 imposes a smoothness condition on the distribution of the error terms in (7.10). This condition can again be relaxed at the expense of additional complications in the derivations and the resulting formulae.

In constructing an optimal PML estimator, one could directly try to minimize $E(\overline{PLR})$ with respect to $\rho(\cdot)$ subject to the restriction $E(\psi(\Omega_{11}^{-1/2}\varepsilon_t)) = 0$. This results in a parabolic partial differential equation in k variables, which is rather hard to solve. Therefore, the problem is tackled from a slightly different angle. First, the class of PML estimators is enlarged to the class of PMLF estimators. The latter type of estimators solve a first order condition rather than a maximization problem. Differentiating the logarithm of (7.14) with respect to θ and equating to zero, one obtains a system of equations in the vector of unknown parameters θ. The value $\hat{\theta}$ that solves this system of equations is labeled the PMLF estimator of θ, where the F stands for the First order conditions. For every PML estimator there is a corresponding PMLF estimator. The converse, however, is not true if $k \geq 2$. The criterion $E(\overline{PLR})$ is now minimized with respect to $\psi(\cdot)$ subject to the restriction $E(\psi(\Omega_{11}^{-1/2}\varepsilon_t)) = 0$. This produces the optimal PMLF estimator. It turns out that this estimator corresponds to a PML estimator only in certain special cases. For $k = 1$, one again obtains the results of Cox and Llatas (1991).

A problem with the approach sketched above is that the function $\psi(\cdot)$ does not uniquely define a PMLF estimator. If instead of $\psi(\cdot)$ one uses $\tilde{\psi}(\cdot) = C_3\psi(\cdot)$, with $|C_3| \neq 0$, then the same PMLF estimator is obtained. This indicates that further restrictions are needed in order to uniquely define $\psi(\cdot)$. The set of restrictions chosen here is $E(\psi'(\Omega_{11}^{-1/2}\varepsilon_t)) = C_1 = I_k$. Other normalizations are, of course, also possible. This particular set of restrictions, however, results in considerable simplifications in the derivations below. The following lemma presents the expectation of $\overline{PLR}$ subject to the two relevant restrictions.

Lemma 7.2 *If the conditions of Theorem 7.1 are satisfied and $C_1 = I_k$, then*

$$\begin{aligned} E(\overline{PLR}) &= \operatorname{tr}(K_0\bar{C}_2\bar{K}_3\bar{C}_2^\top) + (k-r)\operatorname{tr}(K_0P) + \\ &\quad \operatorname{tr}(K_0(A_\perp^\top\Omega_{21}A_\perp)K_5^{-1/2}\bar{K}_2K_5^{-1/2}(A_\perp^\top\Omega_{12}A_\perp)), \end{aligned}$$

with $\bar{K}_i = E(S_1K_iS_1^\top)$ for $i = 2, 3$,

$$\begin{aligned} K_5 &= A_\perp^\top\Omega_{11}A_\perp, \\ \bar{C}_2 &= C_2(A_\perp^\top\Psi B_\perp)^{-1}(A_\perp^\top\Omega_{11}A_\perp)^{1/2}, \\ P &= \bar{K}_0 - (A_\perp^\top\Omega_{21}A_\perp)K_5^{-1}(A_\perp^\top\Omega_{12}A_\perp), \end{aligned}$$

S_1 defined below (7.17), K_i, $i = 1, 2, 3$ defined in Theorem 7.2, C_2 and Ψ defined in Lemma 7.1, and K_0 and $\bar{K}_0$ defined in Theorem 7.1.

Note that the distribution of $S_1K_iS_1^\top$ does not depend upon S_1 for $i = 2, 3$, because $S_1\hat{W}_1$ is a standard Brownian motion due to the orthogonality of S_1. Moreover, $(S_1\tilde{C}_2S_1^\top)$ is independent of the value of S_1. Therefore, if $\tilde{W}_1(s)$ denotes $S_1\hat{W}_1(s)$, then $\tilde{U}(s) = S_1\hat{U}(s)$ follows the Ornstein-Uhlenbeck process

$$d\tilde{U}(s) = (S_1\tilde{C}_2S_1^\top)\tilde{U}(s)ds + d\tilde{W}_1(s).$$

The Lagrangean L of the constrained optimization problem can now be written as

$$L = E(\overline{PLR}) - \Lambda_1^\top E(\psi(\Omega_{11}^{-1/2}\varepsilon_t)) + \mathrm{vec}(\Lambda_2)^\top \mathrm{vec}(I_k - E(\psi'(\Omega_{11}^{-1/2}\varepsilon_t))),$$

where Λ_1 and Λ_2 are the Lagrange multipliers. Setting the first order variation of the Lagrangean with respect to ψ equal to zero, one can solve for the optimal choice of ψ. The result is presented in the following theorem.

Theorem 7.3 *Let Assumption 7.4 and the conditions of Theorem 7.1 be satisfied. Then the function ψ^* that minimizes $E(\overline{PLR})$ subject to the restrictions $E(\psi(\Omega_{11}^{-1/2}\varepsilon_t)) = 0$ and $E(\psi'(\Omega_{11}^{-1/2}\varepsilon_t)) = I_k$, has to satisfy*

$$(k-r)A_\perp^\top \psi^*(\Omega_{11}^{-1/2}\varepsilon_t) = K_4 A_\perp^\top \varepsilon_t + (K_4 - (k-r)I_{k-r})A_\perp^\top \mathcal{I}^{-1}\frac{d\ln f(\varepsilon_t)}{d\varepsilon_t},$$

with

$$K_4 = -A_\perp^\top \Omega_{21}^* A_\perp (K_5^{-1/2}\bar{K}_2 K_5^{-1/2} - (k-r)K_5^{-1}),$$

$f(\varepsilon_t)$ the density function of ε_t, $\mathcal{I} = -E((d\ln f(\varepsilon_t))/(d\varepsilon_t^\top d\varepsilon_t))$ the information matrix, and Ω_{21}^* such that

$$(k-r)A_\perp^\top \Omega_{21}^* A_\perp = K_4 A_\perp^\top \Omega_{11} A_\perp + ((k-r)I_{k-r} - K_4)A_\perp^\top \mathcal{I}^{-1} A_\perp.$$

Theorem 7.3 reveals that only $A_\perp^\top \psi$ can be identified if one uses the objective function and the restrictions above. So only that part of ψ that is orthogonal to the space spanned by the error correction vectors A, matters asymptotically for the optimal choice of ψ. Furthermore, Theorem 7.3 shows that the optimal choice of the score function ψ is, in general, not proportional to the likelihood score. In fact, omitting for the moment the premultiplication by the matrix $A_\perp^\top$, the optimal score function is a linear combination of the Gaussian pseudo score and the true likelihood score. Similar results were found in the univariate context by Cox and Llatas (1991). The present results are surprising, because in the stationary setting it is known that the (asymptotically) optimal estimator from an MSE perspective is the ML estimator. This no longer holds in the nearly-nonstationary setting. Finally, if $f(\cdot)$ is the Gaussian density, the optimal ψ has to satisfy $A_\perp^\top \psi^*(\Omega_{11}^{-1/2}\varepsilon_t) = A_\perp^\top \varepsilon_t$. As a consequence,

$$\psi^*(\Omega_{11}^{-1/2}\varepsilon_t) = A_\perp(A_\perp^\top A_\perp)^{-1}A_\perp^\top \varepsilon_t + A(A^\top A)^{-1}A^\top g(\varepsilon_t), \tag{7.19}$$

where the function $g(\cdot)$ is such that ψ^* still satisfies Assumption 7.3 and the restriction $E(\partial g(\varepsilon_t)/\partial \varepsilon_t^\top) = I$. For example, if $f(\cdot)$ is the standard bivariate Gaussian density, both $g(\varepsilon_t) = (\varepsilon_{1,t}, \varepsilon_{2,t})^\top$ and $g(\varepsilon_t) = (\varepsilon_{1,t}^3/3, \varepsilon_{2,t}^3/3)^\top$ are suitable specifications for $g(\cdot)$, which again illustrates the partial identification of ψ^*.

Given the optimal PMLF estimator of Theorem 7.3, one can try to find the corresponding PML estimator. This, however, raises a problem. In the univariate context, one can easily write ρ^* as a linear combination of the Gaussian log likelihood and the true log likelihood (see Cox and Llatas (1991)). In the multivariate setting this cannot be achieved in general. A sufficient condition for the existence of a PML estimator corresponding to the optimal PMLF estimator is that Ω_{11} is proportional to the inverse of the information matrix. This condition can even be slightly relaxed, as is shown in the following corollary.

Corollary 7.3 *If the conditions of Theorem 7.3 are satisfied and if* $A_\perp^\top \Omega_{11} A_\perp \propto A_\perp^\top \mathcal{I}^{-1} A_\perp$*, then the constants* a *and* b *can be chosen such that*

$$\rho^*(\Omega_{11}^{-1/2}\varepsilon_t) = a\varepsilon_t^\top \mathcal{I}\varepsilon_t/2 + b\ln(f(\varepsilon_t))$$

defines a PML estimator that is optimal in the sense of Theorem 7.3.

In order to illustrate the construction of the optimal PMLF estimator, consider the following simple example.

Example 7.2 Let the density of ε_t be the k-variate Student t density with $\nu > 2$ degrees of freedom,

$$f(\varepsilon_t) = \frac{\Gamma((\nu+k)/2)}{\Gamma(\nu/2)|\pi(\nu-2)\omega_{11}|^{k/2}}\left(1+\varepsilon_t^\top\varepsilon_t/((\nu-2)\omega_{11})\right)^{-(\nu+k)/2}, \qquad (7.20)$$

with $\omega_{11} > 0$. It follows that $\Omega_{11} = E(\varepsilon_t\varepsilon_t^\top) = \omega_{11}I_k$. It is straightforward to verify that the negative score function of the Student t distribution equals

$$\psi(\Omega_{11}^{-1/2}\varepsilon_t) = (\nu+k)(\varepsilon_t)/((\nu-2)\omega_{11}+\varepsilon_t^\top\varepsilon_t),$$

and that the Fisher information matrix equals

$$\mathcal{I} = E(\partial\psi(\Omega_{11}^{-1/2}\varepsilon_t)/\partial\varepsilon_t) = \frac{\nu(\nu+k)}{(\nu-2)(\nu+k+2)\omega_{11}}I_k.$$

Using these ingredients, one can compute the weighting factor K_4 of Theorem 7.3. Consider the test of H_{k-1} : rank$(\Pi) \leq k-1$, versus H_k : rank$(\Pi) = k$. Furthermore, assume that A_1 and B_1 in (7.16) have rank one, implying that there is one additional cointegrating vector with loadings that decrease to zero as the sample size increases. Without loss of generality, it can be assumed that $A_\perp$ and $B_\perp$ are such that $A_\perp^\top\Omega_{11}A_\perp = 1$ and $A_\perp^\top\Psi B_\perp = 1$, respectively. One obtains

$$K_4 = -\omega_{21}(\bar{K}_2 - 1), \qquad (7.21)$$

with $\bar{K}_2 \in \mathbb{R}$ and $\omega_{21} = A_\perp^\top\Omega_{21}^*A_\perp \in \mathbb{R}$. Following Theorem 7.3, ω_{21} must satisfy

$$\omega_{21} = K_4 + (1-K_4)\iota, \qquad (7.22)$$

with

$$\iota = A_\perp^\top\mathcal{I}^{-1}A_\perp = \frac{(\nu-2)(\nu+k+2)}{\nu(\nu+k)}.$$

Rewriting (7.22) using (7.21), one obtains $\omega_{21} = \iota/(1+(1-\iota)(\bar{K}_2-1))$. Substituting this expression for ω_{21} back into (7.21), one gets the weighting factor

$$K_4 = -\iota(\bar{K}_2-1)/(1+(1-\iota)(\bar{K}_2-1)).$$

To evaluate the weighting factor K_4, the techniques described in Cox and Llatas (1991) can be employed. Using the results of Bobkoski (1983), the joint moment generating function of K_1 and K_3 (see the definitions in Theorem 7.2) is given by

$$\begin{aligned}\Lambda(s_0,s) &= E(\exp(-s_0K_3-sK_1)) \\ &= \exp((C_2+s)/2)(\cosh(z)+(C_2+s)\sinh(z)/z)^{-1/2},\end{aligned}$$

with $z=(C_2^2+2C_2s+2s_0)^{1/2}$. From this one obtains

$$\bar{K}_2 = \int_0^\infty \partial^2\Lambda(s_0,0)/\partial s^2 ds_0.$$

The above integral can be approximated using numerical integration. Some plots of the weighting factor K_4 for several values of C_2 are presented in Figure 7.1. Notice that for increasing values of $1/\nu$, the weight of the Gaussian part in the optimal pseudo score function decreases. Moreover, the further one is away from the null hypothesis, i.e., the larger C_2, the more weight is attached to the true maximum likelihood score. This corresponds with the fact that for stationary time series, i.e., for $C_2\to\infty$, the maximum likelihood estimator is optimal. △

Example 7.2 suggests that power can be gained by exploiting the distributional properties of the innovations that drive the time series. The gain is higher for data that exhibit a higher degree of leptokurtosis and can be realized by using estimators that are, in a sense, "between" the Gaussian and the true maximum likelihood estimator.

In order to apply the optimal PMLF estimator, one needs an estimate of C_2. As this matrix contains the parameters that determine the *local* alternative, it cannot be estimated consistently (see Cox and Llatas (1991)). The information on C_2 does not grow sufficiently fast with the sample size. An operational two step procedure can be devised along the lines of Cox and Llatas (1991, p. 1116). Alternatively, one can use the true ML estimator without the Gaussian part in the score function. As the weights of the true ML part in the optimal specification of ψ^* are larger than 0.8 in the example above, one can still expect a power gain by using such an alternative to Gaussian PML estimation.[5] As a third possibility, one can use a PML estimator that

[5] Note that one can also consider a more ambitious approach. For example, one can try to estimate the density of the innovations ε_t using preliminary estimates of the parameters and a kind of kernel estimator. The estimated density can then be used to compute (nonparametric) maximum likelihood estimators, which can be used to construct a PLR test. This approach should have very appealing properties from an asymptotic point of view, e.g., power and few nuisance parameters. The implications in finite samples, however, remain to be investigated (also compare the npml estimator used in Section 3.3).

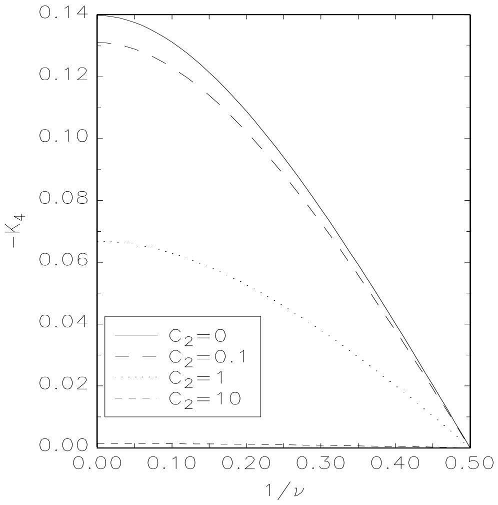

Figure 7.1.— Weight of the Gaussian score function in the optimal pseudo score function $\psi^*(\cdot)$.

offers some protection against leptokurtosis and outliers if one suspects that such phenomena are important problems in the data set that is dealt with. Otherwise, the Gaussian PML estimator can be used. This third procedure is often employed in robust statistics and is especially useful if one is uncertain about the exact specification of the true likelihood function. The main problem with the latter two approaches is the presence of nuisance parameters in the limiting distribution of the PLR test statistic. Some simple corrections for solving this problem in practical circumstances are given in Section 7.5. An additional concern for the third approach is that one has to verify that part (iii) of Assumption 7.3 is satisfied, or stated differently, that one is estimating the correct quantity (compare the simulations with recentered χ^2 distribution in Section 3.3).

7.5 Obtaining Critical Values

The critical values of most unit root tests are usually obtained by means of simulation.[6] Simulating the distribution of the PLR test for nonlinear PML

[6] There are techniques that allow one to compute the critical values by means of numerical integration (see Evans and Savin (1981, 1984) and Abadir (1992)), but these techniques are still quite difficult to implement in a multivariate context (see Abadir and Larsson (1994)). Therefore, they are not applied here. Moreover, Chapter 8 reveals that the asymptotic distribution can provide a very crude approximation to the finite sample distribution of cointegration tests based on robust estimators. Therefore, instead of simulating the limiting distribution by approximating the (stochastic) integrals in Theorem 7.1, the critical values

estimators in a multivariate setting is very time consuming. In this section, a simple Bartlett type correction for the PLR statistic is considered. The resulting corrected statistic, PLR^*, aims to have approximately the same critical values as the Gaussian PLR statistic of Johansen. The simple correction does not provide an asymptotically correct inference procedure. It merely serves as a tool for the applied researcher.

The idea is as follows. Consider the distribution of PLR under the null hypothesis, so $C_2 = 0$. Theorem 7.1 reveals that the behavior of the PLR test is similar for different specifications of the pseudo likelihood. In particular, Theorem 7.2 reveals that the PLR test can asymptotically be represented as the squared norm of the sum of two independent stochastic vectors, one of which is a Gaussian functional, while the other one is normally distributed. Consequently, the tail behavior of the PLR statistic is similar for a wide variety of PML estimators. This fact can be exploited in deriving an approximation to the critical values of the test. Let PLR^G denote the PLR statistic based on the Gaussian PML estimator, so $\psi(\Omega_{11}^{-1/2}\varepsilon_t) = \Omega_{11}^{-1}\varepsilon_t$. Similarly, let $\overline{PLR}^G$ denote the weak limit of PLR^G. Moreover, define

$$PLR^* = PLR \cdot E(\overline{PLR}^G) \cdot (E(\overline{PLR}))^{-1}, \tag{7.23}$$

then PLR^* has asymptotically the same mean and the same tail behavior as PLR^G. Therefore, the critical values of the PLR^G statistic intuitively provide a reasonable approximation to the critical values of the PLR^* statistic.

From Lemma 7.2 one can obtain an expression for the correction factor in (7.23). This correction factor depends on the alternative hypothesis through the value of $A_1B_1^\top$ in the matrix C_2. As the main objective here is to provide a correction such that PLR^* has approximately the same limiting distribution under the null hypothesis as PLR^G, the correction factor is evaluated at $A_1B_1^\top = 0$. The expression for this factor is given by

$$\left.\frac{E(\overline{PLR}^G)}{E(\overline{PLR})}\right|_{A_1B_1^\top=0} = \frac{\operatorname{tr}(\bar{K}_2)}{\operatorname{tr}(K_0\tilde{R}\bar{K}_2\tilde{R}^\top) + (k-r)\operatorname{tr}(K_0P)}, \tag{7.24}$$

with $\tilde{R} = (A_\perp^\top C_1^{-1}\Omega_{21}A_\perp)K_5^{-1/2}$, K_5 and P as defined in Lemma 7.2, and K_0 and $\bar{K}_2$ defined in Theorem 7.1 and 7.2, respectively. There are three points to note about the corrected PLR statistic. First, the inference based on PLR^* in combination with the critical values of PLR^G is asymptotically biased. For one thing, the actual and nominal size of the test do not coincide when the sample size tends to infinity. Moreover, the correction factor (7.24) is only appropriate under the null hypothesis. Under the alternative, a different adjustment might be called for. In Section 7.6, it is investigated by means of simulation whether the PLR^* statistic combined with the PLR^G critical values provides a useful testing procedure. It is also checked whether the resulting bias in the inference procedure is negligible in practical situations.

are approximated by simulating the PLR test directly.

A second point to note about PLR^*, is that the PLR^* procedure as it stands is not feasible, because the correction factor (7.24) depends upon unknown parameters. The unknown quantities can, however, be consistently estimated. Given a consistent estimate $\hat{A}$ of A, one can choose $(k-r)$ linearly independent vectors orthogonal to $\hat{A}$. These produce an estimate $\hat{A}_\perp$ of $A_\perp$. The feasible correction factor will be independent of the method for constructing $\hat{A}_\perp$. In particular, if $\hat{A}_\perp$ is replaced by $\hat{A}_\perp C_5$ for some nonsingular matrix C_5, then the correction factor remains the same. Furthermore, from the estimated parameters and the model one can compute residuals $\hat{\varepsilon}_t$, which can be used to estimate Ω_{ij} as

$$\begin{pmatrix} \hat{\Omega}_{11} & \hat{\Omega}_{12} \\ \hat{\Omega}_{21} & \hat{\Omega}_{22} \end{pmatrix} = T^{-1} \sum_{t=1}^{T} \left(\hat{\varepsilon}_t^\top, \psi(\hat{\Omega}_{11}^{-1/2} \hat{\varepsilon}_t)^\top) \right)^\top \left(\hat{\varepsilon}_t^\top, \psi(\hat{\Omega}_{11}^{-1/2} \hat{\varepsilon}_t)^\top) \right).$$

Similarly, C_1 can be estimated by $\hat{C}_1 = T^{-1} \sum_{t=1}^{T} \psi'(\hat{\Omega}_{11}^{-1/2} \hat{\varepsilon}_t)$. Substituting these estimates into the appropriate formulae, one obtains estimates of K_0, K_5, $\tilde{R}$, and P. The computation of $\bar{K}_2$ is somewhat more problematic. In effect, one needs the joint characteristic function of $\int \hat{W}_1 d\hat{W}_1^\top$ and $\int \hat{W}_1 \hat{W}_1^\top$. This function can be found in Abadir and Larsson (1994). Another, perhaps simpler route is to obtain an estimate of $\bar{K}_2$ by means of simulation. Constructing random walks y_t of length $T+1$ using Gaussian innovations with mean zero and covariance matrix I_{k-r}, one can approximate $\int \hat{W}_1 d\hat{W}_1^\top$ and $\int \hat{W}_1 \hat{W}_1^\top$ by $\hat{K}_1 = T^{-1} \sum_{t=1}^{T} y_t (y_{t+1} - y_t)^\top$ and $\hat{K}_3 = T^{-2} \sum_{t=1}^{T} y_t y_t^\top$, respectively. Similarly, one can approximate K_2 by $\hat{K}_2 = \hat{K}_1^\top \hat{K}_3^{-1} \hat{K}_1$. The Monte-Carlo mean of this estimator over 10,000 replications serves as an estimator of $\bar{K}_2$. Both from the simulations and from the theoretical results of Abadir and Larsson (1994) and Abadir et al. (1994) it appears that $\bar{K}_2$ is proportional to the unit matrix. Therefore, (7.24) can be rewritten as

$$\frac{\bar{k}_2 (k-r)}{\bar{k}_2 \mathrm{tr}(K_0 \tilde{R} \tilde{R}^\top) + (k-r) \mathrm{tr}(K_0 P)}, \tag{7.25}$$

with $\bar{k}_2$ a positive constant, $\tilde{R}$ defined below (7.24), and P defined in Lemma 7.2. Some values of $\bar{k}_2$ are presented in Table 7.1. Using the proposed estimators for the unknown parameters in (7.25), one can construct a feasible PLR^* procedure.

A third point to note about the PLR^* test is that the Bartlett type correction factor in (7.24) only corrects the mean of the test statistic. Correcting for the mean, however, is insufficient for correcting all the cumulants of the limiting distribution simultaneously. Further modifications can be thought of that, for example, also correct the variance or other higher order moments of PLR^* to that of PLR^G. The resulting correction factors must again be consistently estimated for the procedure to be feasible. This becomes considerably more complicated if additional moments are taken into account. Therefore, I stick to the simple correction proposed in (7.24).

TABLE 7.1

Monte-Carlo Estimates of $\bar{k}_2$

$k-r$	1	2	3	4	5
$\bar{k}_2$	1.1428	3.0830	4.9801	7.0453	9.0597
	(0.0150)	(0.0260)	(0.0331)	(0.0395)	(0.0442)

The matrix $\bar{K}_2$, defined in Theorem 7.2, is proportional to $\bar{k}_2 I_{k-r}$. The table presents estimates of $\bar{k}_2$ based on 10,000 Monte-Carlo simulations of samples of size 1,000. Monte-Carlo standard errors are between parenthesis.

7.6 Simulation Results

This section presents the results of a small simulation experiment. The experiment serves as an illustration of the properties of some simple PLR tests relative to the Johansen test.

First, the simulations under the null hypothesis are described. These simulations provide the critical values of the PLR test. The simulation experiment was set up as follows. For several values of $k-r$, a $(k-r)$-variate random walk y_t, $t=0,\ldots,T$, was generated with standard Gaussian innovations. Using the generated time series y_t, the test statistics PLR_0 and PLR_0^* were computed, which test the hypothesis of zero cointegrating relations versus $k-r$ cointegrating relations. This was done over N Monte-Carlo simulations. The number of observations and replications used, are $T=100$ and $N=1,000$, respectively.

In order to illustrate the properties of the PLR test, only a very simple pseudo likelihood was considered, namely the multivariate Student t (compare (7.20)),

$$\rho(\Omega_{11}^{-1/2}\varepsilon_t) = \frac{1}{2}(\nu+k)\ln(\nu+\varepsilon_t^{\top}\Omega_{11}^{-1}\varepsilon_t).$$

The cases considered were $\nu = 1,3,5,7,10,\infty$. Note that setting $\nu=\infty$ yields the Gaussian PML estimator of Johansen. The results of the experiment are summarized in Table 7.2 and in the left panels of Figure 7.2.

One feature that appears from both the figures and the table is that the distribution shifts to the right if either the degrees of freedom parameter in the pseudo likelihood, ν, decreases, or the dimension of the Brownian motion, $k-r$, increases. The effect of a decrease in ν is larger in higher dimensions.

The results for the corrected tests, PLR^*, are presented in Table 7.3 and in the right panels of Figure 7.2. For the simulations described above, $\hat{A}_\perp = I_k$. This follows from the fact that under the null hypothesis of no cointegrating relations rank$(\Pi)=0$, which implies rank(A) = rank$(B)=0$. By comparing the plots of the empirical cumulative distribution functions (c.d.f.'s) in the right panels of Figure 7.2 with those in the left panels, one sees that the distribution of the Gaussian PLR test gives a reasonable approximation to the c.d.f. of the feasible PLR^* test.

It is interesting to know whether there is an ordering in the different PLR^* statistics with respect to the degrees of freedom parameter ν for a given data

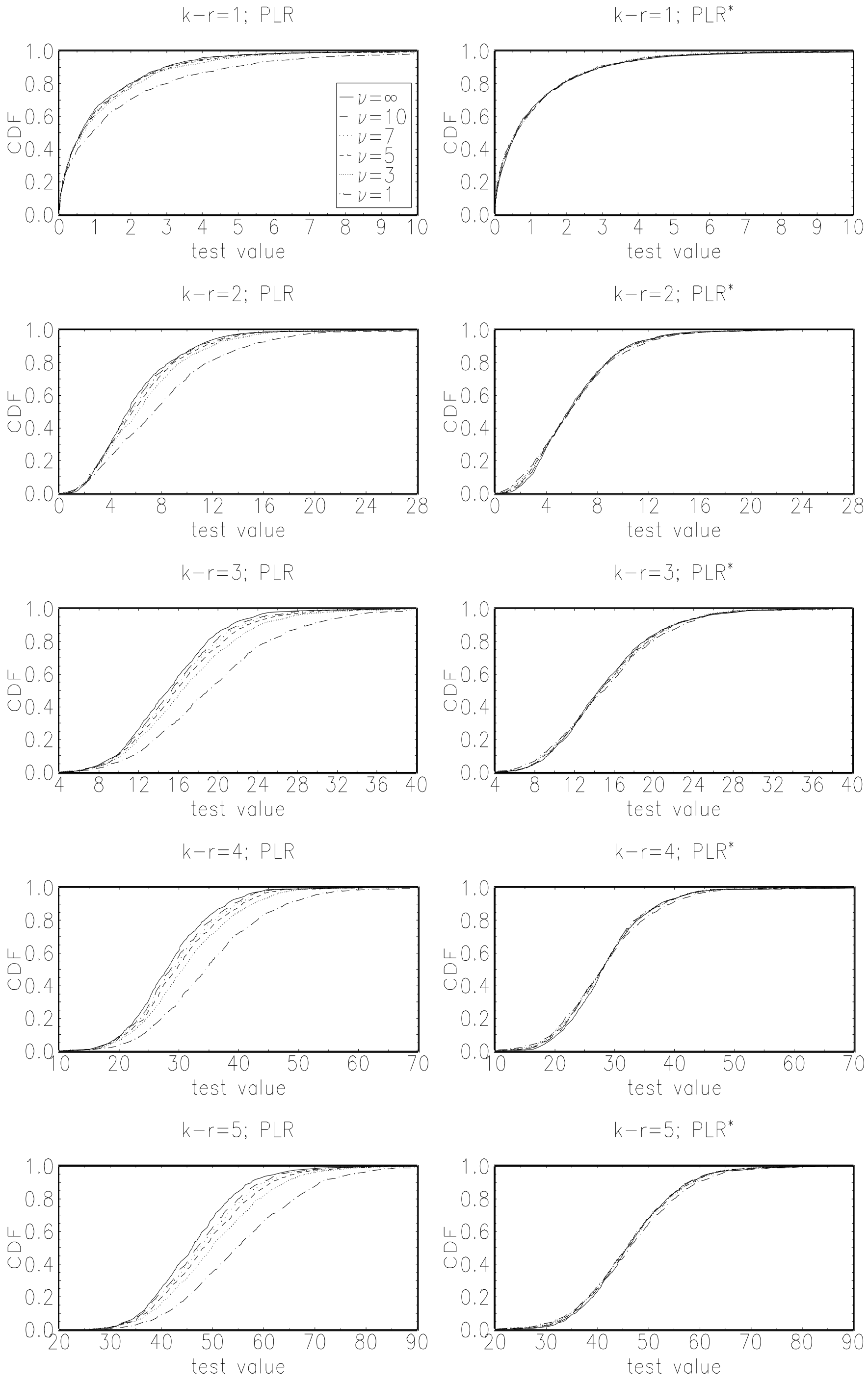

Figure 7.2.— Empirical distribution functions of PLR and PLR^*.

The figure contains the c.d.f.'s of the PLR test and the corrected PLR test (PLR^*), based on 1,000 Monte Carlo simulations and sample size 100. For a generated data set, the PLR test is directly calculated using the Student t pseudo likelihood with ν degrees of freedom. $k-r$ denotes the dimension of the random walk process for which the test is computed.

TABLE 7.2
Critical Values of the *PLR* Test for the Student t Pseudo Likelihood

ν	$k-r$	quantile							
		0.500	0.600	0.700	0.800	0.900	0.950	0.975	0.990
∞	1	0.593	0.862	1.279	1.980	2.926	3.866	4.999	6.551
	2	5.376	6.067	7.126	8.611	10.747	12.343	14.069	16.599
	3	14.985	16.131	17.588	18.918	21.303	23.510	24.624	28.680
	4	27.982	29.713	31.487	34.300	38.013	40.978	43.477	45.461
	5	46.009	47.970	50.348	53.223	57.451	61.892	65.355	71.507
10	1	0.624	0.902	1.387	2.014	2.925	3.961	4.939	6.380
	2	5.520	6.298	7.269	8.856	10.749	12.729	14.250	16.488
	3	15.280	16.575	17.982	19.593	22.029	24.234	26.964	29.083
	4	28.573	30.250	32.804	35.478	39.018	41.596	43.221	45.648
	5	46.847	49.277	51.945	55.039	59.560	62.844	67.923	72.987
7	1	0.609	0.921	1.395	2.032	3.015	4.052	5.098	6.603
	2	5.608	6.505	7.386	8.793	10.976	12.982	14.392	16.738
	3	15.349	16.796	18.264	19.993	22.512	24.669	27.758	29.968
	4	28.995	30.742	33.056	36.024	39.730	42.386	44.627	46.715
	5	47.517	49.936	52.797	56.121	60.327	63.852	69.874	73.856
5	1	0.642	0.948	1.420	2.067	3.114	4.185	5.457	6.555
	2	5.735	6.729	7.669	9.007	11.354	13.384	14.744	16.753
	3	15.704	17.100	18.610	20.555	22.963	25.430	28.871	30.468
	4	29.576	31.378	33.660	36.711	40.595	43.483	45.659	48.489
	5	48.092	50.644	53.795	57.067	61.505	65.432	71.293	74.638
3	1	0.666	1.005	1.520	2.193	3.313	4.446	5.924	7.222
	2	6.144	7.014	8.135	9.504	11.938	14.096	15.648	17.493
	3	16.390	17.884	19.421	21.554	24.225	27.343	30.178	33.553
	4	30.708	32.607	35.106	38.212	42.311	45.642	48.346	50.738
	5	49.698	52.567	55.820	59.317	64.291	68.185	73.429	77.554
1	1	0.905	1.318	1.960	2.978	4.730	6.597	8.471	11.838
	2	7.421	8.592	9.906	11.649	14.594	17.091	18.890	21.610
	3	18.985	20.862	22.678	25.116	28.991	32.571	35.252	39.386
	4	34.563	36.919	39.233	43.086	47.953	52.507	55.409	61.065
	5	54.824	58.057	61.284	66.168	70.775	76.600	81.671	86.197

ν is the degrees of freedom parameter of the Student t pseudo likelihood, k is the dimension of the time series, and r is the cointegrating rank. The critical values were obtained using 1,000 Monte-Carlo simulations with multivariate random walks of length 100 with standard Gaussian innovations.

TABLE 7.3
Critical Values of the feasible PLR^* Test for the
Student t Pseudo Likelihood

ν	$k-r$	quantile							
		0.500	0.600	0.700	0.800	0.900	0.950	0.975	0.990
∞	1	0.630	0.966	1.427	2.009	3.102	4.138	5.202	7.534
	2	5.448	6.107	7.116	8.273	10.136	12.023	13.648	16.457
	3	14.679	16.130	17.517	19.201	22.439	25.215	27.180	29.212
	4	28.057	29.896	32.200	35.253	39.089	42.152	45.845	49.148
	5	47.108	49.368	52.044	54.705	59.708	63.608	66.166	69.774
10	1	0.606	0.980	1.436	1.956	2.979	4.018	4.913	7.111
	2	5.405	6.152	7.115	8.233	10.204	12.108	13.563	16.325
	3	14.648	15.969	17.342	19.586	22.260	24.316	27.572	29.804
	4	28.003	29.807	31.665	34.665	38.712	42.003	44.932	49.242
	5	47.326	49.260	51.591	54.429	58.612	62.658	65.309	68.525
7	1	0.616	0.988	1.423	1.966	2.947	4.081	4.871	7.015
	2	5.427	6.183	7.115	8.283	10.393	12.215	13.759	16.056
	3	14.629	16.057	17.408	19.719	22.205	24.575	27.360	30.279
	4	28.026	29.838	31.803	34.703	38.963	42.031	45.459	49.717
	5	47.295	49.341	51.697	54.670	58.357	63.215	65.664	66.854
5	1	0.630	0.992	1.404	2.002	2.912	4.085	5.232	6.919
	2	5.466	6.191	7.087	8.295	10.314	12.236	14.153	15.695
	3	14.590	16.094	17.488	19.698	22.150	24.664	27.308	31.070
	4	27.985	29.814	31.895	34.535	39.019	42.310	45.451	50.674
	5	47.310	49.054	51.982	54.965	59.301	63.494	65.328	66.748
3	1	0.636	0.976	1.392	2.058	2.921	3.983	5.300	6.556
	2	5.475	6.209	7.061	8.379	10.615	12.285	14.698	16.470
	3	14.794	16.125	17.695	19.853	22.401	25.126	27.513	31.116
	4	28.047	29.993	32.260	35.105	39.520	42.476	45.891	52.061
	5	47.292	48.948	52.492	55.261	60.517	63.292	66.712	68.268
1	1	0.584	0.843	1.282	1.906	3.189	4.262	5.188	6.635
	2	5.334	6.212	7.162	8.499	11.109	13.001	15.467	17.912
	3	14.566	16.329	18.181	20.280	23.024	26.645	29.534	33.019
	4	28.394	30.467	33.180	36.219	40.016	44.683	48.106	54.271
	5	47.670	50.454	53.390	56.880	62.311	66.579	70.099	72.670

ν is the degrees of freedom parameter of the Student t pseudo likelihood, k is the dimension of the time series, and r is the cointegrating rank. The critical values were obtained using 1,000 Monte-Carlo simulations with multivariate random walks of length 100 with standard Gaussian innovations. PLR^* was computed by applying the correction factor in (7.24) to every realization of PLR.

set. This turns out not to be the case, as is illustrated in Figure 7.3. Consider the setting with $k - r = 1$ For every simulated time series that was used to construct Table 7.3, the difference is computed between the Gaussian test and the Student t based PLR^* statistic with $\nu = 10$ (left panel of Figure 7.3) and $\nu = 1$ (right panel). This difference regularly alternates sign, indicating that the Student t based PLR^* statistics can be above as well as below the Gaussian test statistic for a given data set. The magnitude of the difference seems to decrease with the degrees of freedom parameter ν. For $\nu = 1$, the maximum absolute difference is approximately 7, whereas the maximum absolute difference for $\nu = 10$ is about 2.3. Similar results hold if one considers the absolute difference between two Student t based PLR^* tests.

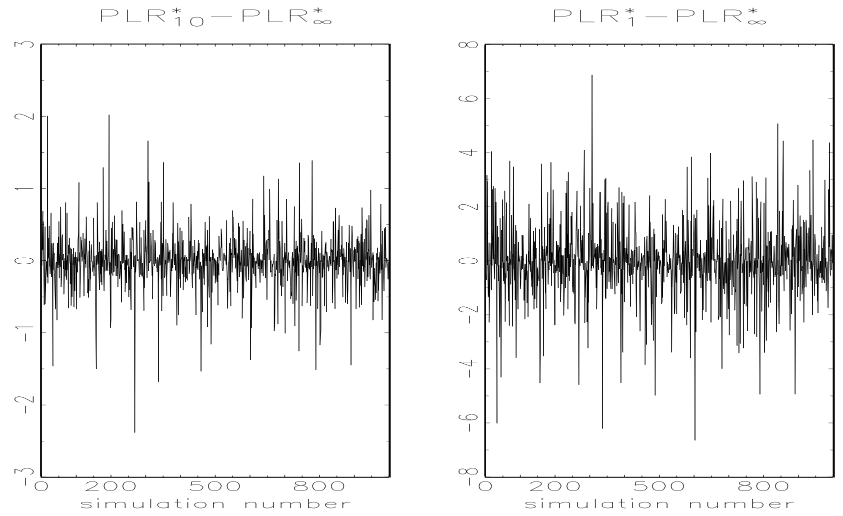

Figure 7.3.— Difference per simulation between the Gaussian feasible PLR^* statistic and the Student t based feasible PLR^* statistic for $k - r = 1$ and $\nu = 10$ and $\nu = 1$

Both findings can be explained by looking at Theorem 7.2. First, the difference between two PLR^* tests based on different pseudo likelihoods is a random variable with zero mean. This follows from the fact that all PLR^* statistics are designed to have the same mean as the Gaussian PLR statistic. Second, the variance of the independent Brownian motion $\hat{W}_3$ depends upon the canonical correlations R between $A_{\perp}^{\top}\varepsilon_t$ and $A_{\perp}^{\top}C_1^{-1}\psi(\Omega_{11}^{-1/2}\varepsilon_t)$ as $E(\hat{W}_3(s)\hat{W}_3(s)^{\top}) = s(I_{k-r} - R^2)$. As the canonical correlations are increasing functions of the degrees of freedom parameter ν for Gaussian ε_t, one can expect the difference between the two tests in Figure 7.3 to be larger for lower values of ν.

Next, the power of the PLR test is considered. As was demonstrated in Section 7.3, the power of the PLR test depends upon the parameter matrix $C_2 = A_\perp^\top A_1 B_1^\top B_\perp$, which contains $(k-r)^2$ parameters. Johansen (1989) demonstrates how to reduce the number of free parameters by exploiting the invariance property of the Ornstein-Uhlenbeck process under rotation. The interesting question in this chapter, however, is not so much the absolute power of the PLR tests, but rather the power of the tests relative to that of the Gaussian PLR test. Therefore, in this chapter a simple model is considered in order to compare the power differences between Gaussian and non-Gaussian based tests in a setting with normal and leptokurtic innovations, respectively.

In the simulation experiment for the power of the PLR test, the following data generating process is used:

$$\begin{pmatrix} \Delta y_{1t} \\ \Delta y_{2t} \end{pmatrix} = \begin{pmatrix} -\tilde{c}_2/T \\ 0 \end{pmatrix} (y_{1,t-1} - y_{2,t-1}) + \begin{pmatrix} \varepsilon_{1t} \\ \varepsilon_{2t} \end{pmatrix}, \tag{7.26}$$

where $\tilde{c}_2$ is a positive constant, and where the ε_t either have a standard normal distribution or a truncated Cauchy distribution (see further below). The two roots of the VAR polynomial in (7.26) are 1 and $(1-\tilde{c}_2/T)^{-1}$, respectively. In order to satisfy Assumption 7.2, it must hold that $0 \le \tilde{c}_2 < 2T$, such that both roots are on or outside the unit circle. Note that for $\tilde{c}_2 = 0$ the system in (7.26) has two unit roots and there exist no cointegrating relationships. If $0 < \tilde{c}_2 < 2T$, then there is one cointegrating relationship and the cointegrating vector lies in the space spanned by $(1,-1)$, while the error correction parameter is $-\tilde{c}_2/T$.

Five different test statistics are considered for the power simulations, namely the Gaussian PLR test, the Student t based PLR test with 5 degrees and 1 degree of freedom, and the two corrected Student t based PLR^* tests. The rejection frequencies of these tests are simulated in the usual way. After generating a time series according to (7.26), the value of each of the above test statistics is computed and compared with its 5% and 10% critical value, respectively. For the first three tests, the critical values from Table 7.2 are used. For the last two tests, the critical values of the Gaussian PLR test are used. The simulations use time series of length $T = 1,000$ and $1,250$ Monte Carlo replications. Simulations with $T = 100$ resulted in identical conclusions. The standard errors of the rejection frequencies are smaller than or equal to $0.5N^{-1/2} \approx 0.014$.

Using the data generating process in (7.26), two experiments were performed. In the first experiment, the ε_t were drawn from a bivariate normal distribution with mean zero and covariance matrix I_2. The restriction of the covariance matrix to be the unit matrix is unimportant in the present setup, because of the presence of the scaling matrix Ω_{11} in the pseudo likelihood. For Gaussian ε_t, Theorem 7.3 reveals that the optimal pseudo score function from a minimum MSE perspective is the Gaussian score function, $\psi(\Omega_{11}^{-1/2}\varepsilon_t) = \Omega_{11}^{-1}\varepsilon_t$. Therefore, one can expect the Johansen or Gaussian PLR test to have the largest power in this case.

In the second experiment, the Gaussian distribution for ε_t was replaced by the truncated bivariate Cauchy distribution. The Cauchy distribution was

TABLE 7.4
Rejection Frequencies of the PLR Tests
for Gaussian Innovations

$\tilde{c}_2$	PLR_G	PLR_5	PLR_1	PLR_5^*	PLR_1^*
			10% level		
0	0.101	0.116	0.115	0.113	0.120
1	0.086	0.122	0.137	0.118	0.140
5	0.330	0.304	0.238	0.301	0.244
10	0.702	0.652	0.524	0.648	0.534
20	0.997	0.985	0.870	0.984	0.877
			5% level		
0	0.057	0.056	0.071	0.059	0.078
1	0.050	0.061	0.077	0.067	0.085
5	0.207	0.180	0.152	0.194	0.165
10	0.559	0.514	0.399	0.531	0.421
20	0.981	0.946	0.794	0.953	0.814

The table contains the rejection frequencies of the Gaussian based PLR test (PLR_G), the Students t based PLR test with 5 (PLR_5) and 1 (PLR_1) degrees of freedom, and the corrected PLR tests for the Student t pseudo likelihood with 5 (PLR_5^*) and 1 (PLR_1^*) degrees of freedom. The hypothesis of no cointegrating relationships (H_0) is tested against the alternative of stationarity (H_2). The data generating process is (7.26) with Gaussian innovations ε_t.

truncated to the set $\{\varepsilon_{1t}^2+\varepsilon_{2t}^2 \leq F_{0.95}(2,1)\}$, with $F_{0.95}(2,1)$ the 95th percentile of the F distribution with two degrees and one degree of freedom, respectively. The truncation was introduced in order to guarantee the existence of a sufficient number of moments (compare Assumption 7.1). Although the truncated Cauchy distribution does not satisfy Assumption 7.4 from Section 7.4, one can still expect from Theorem 7.3 that a power gain can be realized by exploiting the non-Gaussian form of the distribution.

The results of the first experiment are presented in Table 7.4. For $\tilde{c}_2 = 0$, the rejection frequency should be equal to the size of the test. This appears to be approximately true for most test statistics. There are, however, some slight size distortions for PLR_1 and PLR_1^* at the 5% level. Under small deviations from the null hypothesis, the number of rejections generally increases for all test statistics. As expected on the basis of Theorem 7.3, the rejection frequencies of the Gaussian PLR test are higher than those of the other tests. Furthermore, the size and power of the corrected PLR statistics are approximately equal to that of their uncorrected counterparts. This argues in favor of the Bartlett type correction of the PLR test, because it avoids the need for computing new critical values for every separate choice of the pseudo likelihood.

The results of the second experiment are given in Table 7.5. The first thing to notice is that the non-Gaussian PLR tests have an actual size below

TABLE 7.5
Rejection Frequencies of the PLR Tests
for Truncated Cauchy Innovations

$\tilde{c}_2$	PLR_G	PLR_5	PLR_1	PLR_5^*	PLR_1^*
			10% level		
0	0.084	0.012	0.017	0.130	0.144
1	0.096	0.035	0.109	0.289	0.395
5	0.307	0.737	0.878	0.949	0.979
10	0.722	0.994	0.998	0.999	1.000
20	0.995	1.000	1.000	1.000	1.000
			5% level		
0	0.049	0.006	0.007	0.090	0.098
1	0.051	0.019	0.062	0.201	0.314
5	0.205	0.606	0.818	0.922	0.966
10	0.576	0.981	0.998	0.999	1.000
20	0.978	1.000	1.000	1.000	1.000

The data generating process is (7.26) with Gaussian innovations ε_t. For further exlanation, see the note to Table 7.4.

the nominal size. The power of these non-Gaussian PLR tests, however, very rapidly exceeds the power of the Johansen test if one considers (local) departures from the null hypothesis. For $\tilde{c}_2 = 5$, the rejection frequencies of PLR_5 and PLR_1 at the 5% level are already three and four times as high as that of PLR_G. This demonstrates that it is worthwhile to exploit the nonnormality of the innovations in order to increase the power of cointegration tests. The power could be further increased if the discrepancy between the actual and nominal sizes of the tests could be corrected. This, however, is a separate subject and it is not dealt with in this thesis.

A second point that can be noticed in Table 7.5 is that the actual sizes of the corrected test statistics PLR_5^* and PLR_1^* are above their nominal values if the innovations are truncated Cauchy. This makes inference based on the corrected test statistics difficult to interpret in this situation.

7.7 Model Extensions

This section briefly discusses two possible model extensions and their effects on the asymptotic distribution of the PLR statistic. First, the consequences of including deterministic functions of time as additional regressors in (7.10) are discussed. Second, the effect of incorporating additional unknown nuisance parameters in the pseudo likelihood (7.14) are dealt with.

It is well known that the incorporation of deterministic time trends in (7.10) complicates the asymptotic analysis. For example, if the data generating process is (7.10) and if one uses a regression model that contains a constant term in addition to the regressors in (7.10), then the Ornstein-Uhlenbeck process $\hat{U}(s)$ in Theorem 7.1 has to be replaced by the demeaned stochastic process $\hat{U}(s)-\int_0^1 \hat{U}(s)ds$. Similarly, the presence of a linear time trend as an additional regressor results in detrended stochastic processes in the limiting distribution of the PLR statistic. The results get even more complicated if one allows for deterministic components to be present in the data generating process (7.10) instead of only in the fitted regression model. A well known example of such a process is the random walk with nonzero drift. For such processes, the interpretation of the deterministic components and their effect on the asymptotic distributions are rather delicate (see Johansen (1994) and Chapter 8).

All of the above points have been addressed in the literature for multivariate time series in the context of Gaussian (pseudo) maximum likelihood estimators. The results, however, carry over in a straightforward manner to the present context of non-Gaussian PML estimators. This is illustrated by the results in the next chapter. Consequently, also the results of Rahbek (1994) for the power of the Gaussian PLR test in the presence of nonzero drift terms in (7.10) go through. This leaves one with the dilemma of choosing the appropriate additional deterministic regressors. If one chooses too few of them, inference is, in general, asymptotically biased. If one chooses the correct regressors, the test statistics are not asymptotically similar (see Johansen (1991, Theorems 2.1 and 2.2)). Finally, if one incorporates too many deterministic functions of time as additional regressors, the power of the PLR test diminishes (see Rahbek (1994)). It is also important to note that the incorporation of additional regressors and of nonzero drift terms in the data generating process complicates the form of the simple Bartlett type corrections discussed in Section 7.5.

A second type of model extension concerns the presence of additional nuisance parameters in the pseudo likelihood. So far, only the presence of a scaling matrix Ω_{11} has been dealt with. If this matrix was unknown, it could be estimated along with the other parameters under suitable regularity conditions (see Assumption 7.1 and Appendix 7.A). From the proof of Theorem 7.1 in Appendix 7.A, one can see that the appropriately normalized Hessian of the pseudo likelihood is asymptotically block diagonal between Ω_{11} and the parameters that are of interest for constructing the PLR test. Consequently, one could also use a consistent preliminary estimate of Ω_{11} in the construction of the PLR test without altering the asymptotic distribution of the test. This finding can easily be generalized towards cases where additional nuisance parameters are present in the pseudo likelihood. A simple example is given by the Student t pseudo likelihood, where the degrees of freedom parameter ν is unknown and estimated.

One can think of three strategies for dealing with unknown nuisance parameters. First, one can set the nuisance parameters equal to some user defined values. This strategy may prove useful if one only uses the PLR test for protection against outliers and leptokurtosis. The nuisance parameters can then be regarded as a type of tuning constants. This way of tackling the problem is often encountered in robust statistics. Second, one can use preliminary consistent estimates of the parameters in order to eliminate them. Third, one can estimate the nuisance parameters along with the other parameters of (7.10) by formulating the relevant pseudo score equations. Such estimators are consistent under suitable regularity conditions (compare Assumption 7.1 and Appendix 7.A).

7.8 Conclusions

In this chapter, the properties of likelihood ratio type tests for testing the cointegration hypothesis were studied. Instead of using the Gaussian likelihood, inference was based on a certain class of pseudo likelihoods. This class contained several well known densities, like the Gaussian and the Student t density. The asymptotic distribution of the pseudo likelihood ratio (PLR) test was derived for a sequence of local alternatives to the null hypothesis of no cointegration. This asymptotic distribution was shown to depend on three types of nuisance parameters, arising from: the distance from the null hypothesis, the possible misspecification of the pseudo likelihood, and the use of a non-Gaussian pseudo likelihood. Even if the likelihood was correctly specified, nuisance parameters remained present if a non-Gaussian pseudo likelihood was used.

Also the optimal choice of the pseudo score vector was discussed. The optimal pseudo score turned out to be only partially identified and equal to a linear combination of the Gaussian PML score and the true ML score. A simple Bartlett type correction for the PLR test was proposed, which had approximately the same critical values as the Gaussian PML test of Johansen (1988, 1991), thus avoiding the need for calculating new critical values for every choice of the pseudo likelihood. Using a simulation experiment, the properties of all the tests were investigated. It was found that the choice of the pseudo likelihood can have a great influence on both the distribution of the PLR test under the null hypothesis, i.e., on the level, and under the alternative, i.e., on the power. It also turned out that the distribution of the Bartlett corrected PLR test was reasonably well approximated by that of the Gaussian PLR test in case of Gaussian random walks.

The power simulations demonstrated the validity of the asymptotic arguments. If the innovations are Gaussian, the Johansen trace test is optimal from a power point of view. If the innovations are, in contrast, truncated Cauchy, the Student t based PLR tests perform better in terms of power. It also appeared that the actual sizes of the non-Gaussian based PLR tests were below the nominal size. Moreover, for irregularly behaved innovations like the truncated Cauchy ones, the corrected PLR tests turned out to be less useful due to the fact that their actual sizes exceeded the nominal ones.

Several interesting topics for future research in this area remain. First, it is worthwhile to devise corrections to the PLR statistic that approximate the Gaussian PLR test better in the tail of the distribution than the simple

Bartlett type corrections used here. Second, more simulation evidence must be gathered in order to demonstrate the advantages and disadvantages of the non-Gaussian PLR test over the Gaussian one in situations that are of practical interest. This especially concerns the inclusion of deterministic functions of time in the regression model as well as nonzero drift terms in the data generating process. Third, the effects of dynamic misspecification of the regression model on the asymptotic distribution of the PLR test must be studied. Unreported preliminary results indicate that the PLR test is very sensitive to dynamic model misspecification. Methods for correcting the effects of misspecification have to be designed. Some interesting possibilities for this approach can be found in Phillips (1991a), who uses the Whittle likelihood for the Gaussian PML estimator, and in Bierens (1994), who constructs a nonparametric cointegration test. Fourth, the outlier robust cointegration tests discussed in this chapter can be generalized in order to deal with periodic and seasonal cointegration. Finally, it remains to be shown how well non-Gaussian PLR tests perform on empirical data. One of the chief difficulties is to construct fast iteration schemes in order to maximize the pseudo likelihood. As this likelihood is, in general, highly nonlinear in the parameters, this might prove a nontrivial task. A simple empirical example using the Student t pseudo likelihood can be found in Franses and Lucas (1995).

7.A Proofs

This Appendix provides the proofs of the statements in Sections 7.3 and 7.4.

In order to prove Theorem 7.1, some further notation is needed. First, normalize the matrix B of cointegration vectors such that $B^\top = (I_r, \beta^\top)$, with β a $((k-r)\times r)$ matrix. Note that under the null hypothesis H_r, such a normalization is always possible, because $\operatorname{rank}(B) = r$. The choice of the leading submatrix in $B^\top$ to be the unit matrix may, however, require a reordering of the elements of y_t. As the (pseudo) likelihood ratio test is invariant under such reparameterizations, no generality is lost by imposing this condition. Next, let $A^\top = (\alpha_{11}^\top, \alpha_{21}^\top)$. Also introduce the $(k \times (k-r))$ matrix K_6, which has the property that $\tilde{A} = (A, K_6)$ has full rank. Under the hypothesis H_k the matrix Π can then be decomposed as

$$\Pi = AB^\top + K_6\alpha_{22}(0, I_{k-r}) = A(I_r, 0) + \tilde{A}(\beta, \alpha_{22}^\top)^\top(0, I_{k-r}), \tag{7.27}$$

with α_{22} a $((k-r)\times(k-r))$ matrix. The number of parameters in A, β, and α_{22} equals the number of elements in Π, namely k^2. Therefore, the parametric decomposition of Π in (7.27) can be used to estimate the parameters of (7.10) under the hypothesis H_k. Note that (7.27) can also be used to estimate the parameters of (7.10) under the null hypothesis H_r. This is seen by setting $\alpha_{22} = 0$, which results in $\Pi = AB^\top$, with A and B of full column rank. Therefore, (7.27) can be used to reformulate the hypotheses of interest as $H_r' : \alpha_{22} = 0$ versus $H_k' : \alpha_{22} \neq 0$. Similar decompositions are found in Phillips (1991a) and Kleibergen and van Dijk (1994), who both use $K_6^\top = (0, I_{k-r})$.

Define the vector of parameters θ to be $\operatorname{vec}((\beta_2, \alpha_{22}^\top)^\top, A, \Gamma, \Omega_{11})$, where $\Gamma = (\Phi_1, \ldots, \Phi_p)$. The hypothesis $\alpha_{22} = 0$ can now be formulated as $H\theta = 0$, with

$H = (I_{k-r}, 0) \otimes (0, I_{k-r})$. Let $\tilde{\theta}_{r,T}$ denote the estimator of θ under the hypothesis $\alpha_{22} = 0$ and let $\hat{\theta}_T$ denote the estimator under the alternative. Furthermore, let θ_T denote the true parameters for the local alternative specification in (7.16) and let θ_0 denote the true parameters for $A_1 B_1^\top = 0$. Note that θ_T approaches θ_0 as the sample size tends to infinity. Using these definitions, the following lemma can be proved.

Lemma 7.3

$$\lim_{T\to\infty} T(\theta_T - \theta_0) = \mathrm{vec}\left(\tilde{A}^{-1} A_1 B_1^\top B_\perp, A_1 B_1^\top (0, I_{k-r})^\top, 0\right),$$

where $\tilde{A}$ was defined earlier as $\tilde{A} = (A, K_6)$.

Proof. The final block of zeros is a trivial consequence of the fact that Γ and Ω_{11} are identical under the null hypothesis and under the local alternatives. Now let

$$\bar{\Pi} = \begin{pmatrix} \bar{\alpha}_{11} & K_6 \bar{\alpha}_{22} \\ \bar{\alpha}_{21} & \end{pmatrix} \begin{pmatrix} I_r & \bar{\beta}^\top \\ 0 & I_{k-r} \end{pmatrix}$$

be such that $\bar{\Pi} = AB^\top + A_1 B_1^\top / T$. Multiplying $\bar{\Pi}$ from the right by $(I_r, 0)^\top$ yields $(\bar{\alpha}_{11}^\top, \bar{\alpha}_{21}^\top) = A^\top + (I_r, 0)^\top B_1 A_1^\top / T$, while right multiplication by the matrix $B_\perp = (-\beta, I_{k-r})^\top$ yields

$$(A, K_6) \begin{pmatrix} \bar{\beta}^\top - \beta^\top \\ \bar{\alpha}_{22} \end{pmatrix} = \tilde{A} \begin{pmatrix} \bar{\beta}^\top - \beta^\top \\ \bar{\alpha}_{22} \end{pmatrix} = A_1 B_1^\top B_\perp / T.$$

□

The key convergence results are given in the following lemma.

Lemma 7.4 *Given the conditions of Theorem 7.1,*

$$D \frac{\partial \ell_T(\theta_T)}{\partial \theta} \Rightarrow -\begin{pmatrix} \int U \otimes d\tilde{A}^\top W_2 \\ \xi_1 \end{pmatrix},$$

$$D \frac{\partial^2 \ell_T(\theta_T)}{\partial \theta^\top \partial \theta} D \Rightarrow \begin{pmatrix} \int U U^\top \otimes \tilde{A}^\top C_1 \tilde{A} & 0 \\ 0 & \Xi_1 \end{pmatrix},$$

where $\xi_1 = O_p(1)$, $\Xi_1 = O_p(1)$, $U(s)$ is the Ornstein-Uhlenbeck process defined in Lemma 7.1, and

$$D = \begin{pmatrix} I_{k(k-r)}/T & 0 \\ 0 & I/T^{1/2} \end{pmatrix}.$$

Proof. It is straightforward to verify that

$$\frac{\partial \ell_T(\theta_0)}{\partial \theta^\top} = \sum_{t=1}^{T} (y_{2,t-1}^\top \otimes \psi_t^\top \tilde{A}, Z_{1t}^\top \otimes \psi_t^\top, Z_{2t}^\top),$$

where $\psi_t = \psi(\Omega_{11}^{-1/2} \varepsilon_t)$, and $y_{2,t-1}$ contains the last $k-r$ rows of y_{t-1}. The vectors Z_{1t} and Z_{2t} are defined as $Z_{1t}^\top = (y_{t-1}^\top B, \Delta y_{t-1}^\top, \ldots, \Delta y_{t-p}^\top)$, and

$$Z_{2t}^\top = -\frac{1}{2}(\mathrm{vec}(\Omega_{11}^{-1}))^\top - \psi_t^\top \Omega_{11}^{1/2} (\varepsilon_t^\top \otimes I_k) \frac{\partial \mathrm{vec}(\Omega_{11}^{-1/2})}{\partial (\mathrm{vec}(\Omega_{11}))^\top},$$

respectively. Note that Z_{1t} and Z_{2t} only contain stationary elements, which together with the i.i.d. assumption for ε_t and the existence of the appropriate moments, implies that $T^{-1/2}\sum_{t=1}^{T}(Z_{1t}^{\top}\otimes\psi_t^{\top}, Z_{2t}^{\top}) = O_p(1)$. Furthermore we have that

$$y_{2,t} = (0, I_{k-r})y_t = (0, I_{k-r})(B(B^{\top}B)^{-1}B^{\top} + B_{\perp}(B_{\perp}^{\top}B_{\perp})^{-1}B_{\perp}^{\top})y_t,$$

where $B_{\perp}^{\top} = (-\beta, I_{k-r})$. From Lemma 7.1 and the stationarity of $B^{\top}y_t$, it follows that $y_{2,\lfloor sT\rfloor}/T^{1/2} \Rightarrow U(s)$. The first part of the lemma now follows directly from Phillips (1988) and Hansen (1992).

Let $\psi_t' = \psi'(\Omega_{11}^{-1/2}\varepsilon_t)$, then

$$D\frac{\partial^2\ell_T(\theta_0)}{\partial\theta^{\top}\partial\theta}D = \sum_{t=1}^{T}\begin{pmatrix} Q_{11,t} & Q_{12,t} & Q_{13,t} \\ Q_{12,t}^{\top} & Q_{22,t} & Q_{23,t} \\ Q_{13,t}^{\top} & Q_{23,t}^{\top} & Q_{33,t} \end{pmatrix} + o_p(1), \tag{7.28}$$

with

$$\begin{aligned}
Q_{11,t} &= -\frac{y_{2,t-1}y_{2,t-1}^{\top}}{T^2}\otimes\tilde{A}^{\top}\psi_t'\tilde{A} \\
Q_{12,t} &= -\frac{y_{2,t-1}Z_{1t}^{\top}}{T^{3/2}}\otimes\tilde{A}^{\top}\psi_t' + \left(\frac{y_{2,t-1}\psi_t^{\top}}{T^{3/2}}\otimes I\right)\frac{\mathrm{vec}(\tilde{A}^{\top})}{\mathrm{vec}(A,\Gamma)^{\top}} \\
Q_{13,t} &= \left(\frac{y_{2,t-1}\varepsilon_t^{\top}}{T^{3/2}}\otimes\tilde{A}^{\top}\psi_t'\Omega_{11}^{1/2}\right)\frac{\mathrm{vec}(\Omega_{11}^{-1/2})}{\mathrm{vec}(\Omega_{11})^{\top}} \\
Q_{22,t} &= -\frac{Z_{1t}Z_{1t}^{\top}}{T}\otimes\psi_t' \\
Q_{23,t} &= \left(\frac{Z_{1t}\varepsilon_t^{\top}}{T}\otimes\psi_t'\Omega_{11}^{1/2}\right)\frac{\mathrm{vec}(\Omega_{11}^{-1/2})}{\mathrm{vec}(\Omega_{11})^{\top}} \\
Q_{33,t} &= -\partial Z_{2t}/\partial\mathrm{vec}(\Omega_{11})^{\top}.
\end{aligned}$$

It is easily checked that under the present conditions $\sum_{t=1}^{T}(Q_{12,t}, Q_{13,t})$ converges to zero for $T\to\infty$. Furthermore, the weak convergence of $\sum_{t=1}^{T}Q_{11,t}$ follows from Phillips and Durlauf (1986) and Hansen (1992). The convergence of the remaining blocks in (7.28) follows directly by applying the law of large numbers. Joint convergence also holds. □

The next lemma gives the appropriate convergence result for the Hessian of the pseudo likelihood.

Lemma 7.5 *Let $\hat{\theta}_T$ with corresponding residuals $\hat{\varepsilon}_t$ be such that $\hat{\Omega}_{11}^{-1/2}\hat{\varepsilon}_t - \Omega_{11}^{-1/2}\varepsilon_t = o_p(1)$ uniformly in t, then*

$$D\left\{\frac{\partial^2\ell_T(\hat{\theta}_T)}{\partial\theta^{\top}\partial\theta} - \frac{\partial^2\ell_T(\theta_T)}{\partial\theta^{\top}\partial\theta}\right\}D \xrightarrow{p} 0,$$

with D and θ_T as defined earlier.

Proof. This follows straightforwardly from the Lipschitz continuity of ψ' and the convergence of (7.28) (compare Chapter 6). □

One of the major results of Lemmas 7.4 and 7.5 is that the appropriately normalized Hessian of the pseudo log likelihood is asymptotically block diagonal between the parameters β and α_{22} on the one hand, and the parameters A, Γ, and Ω_{11} on the other hand. This simplifies the proof of Theorem 7.1, because the effect of Ω_{11} being estimated rather than known can now be discarded. Moreover, without loss of generality, attention can be restricted to the case $p = 0$, i.e., the VAR(1) model. Therefore, with a slight abuse of notation, the sequel of this appendix only discusses the case $p = 0$ with fixed and known scaling matrix Ω_{11}.

Proof of Theorem 7.1. One has $\mathrm{vec}(\alpha_{22}) = H\theta$, with θ as defined above Lemma 7.3. Following Gallant (1987, Chapter 3, Theorems 13 and 15), one obtains

$$\begin{aligned}2(\ell_T(\tilde\theta) - \ell_T(\hat\theta)) \quad = \quad & \frac{\partial \ell_T(\theta_T)}{\partial\theta^\top} J^{-1}H^\top(HJ^{-1}H^\top)^{-1}HJ^{-1}\frac{\partial\ell_T(\tilde\theta)}{\partial\theta} + \\ & 2\frac{\partial \ell_T(\theta_T)}{\partial\theta^\top}J^{-1}H^\top(HJ^{-1}H^\top)^{-1}H(\theta_0-\theta_T) + \\ & (\theta_0-\theta_T)^\top H^\top(HJ^{-1}H^\top)^{-1}H(\theta_0-\theta_T) + o_p(1),\end{aligned} \tag{7.29}$$

where $J = \partial^2\ell_T(\theta_T)/(\partial\theta^\top\partial\theta)$. Note that $H \cdot D = H/T$. Using this fact and the Lemmas 7.3 through 7.5, one can prove that the last term of (7.29) converges weakly to

$$\begin{aligned}&\Big(\mathrm{vec}((0,I_{k-r})\tilde A^{-1}A_1B_1^\top B_\perp)\Big)^\top \Big((\int UU^\top)\otimes((0,I_{k-r})(\tilde A^\top C_1\tilde A)^{-1}(0,I_{k-r})^\top)^{-1}\Big) \\ &\Big(\mathrm{vec}((0,I_{k-r})\tilde A^{-1}A_1B_1^\top B_\perp)\Big) = \\ &(\mathrm{vec}(C_2))^\top(\int UU^\top\otimes K_0)\mathrm{vec}(C_2) = \\ &\mathrm{tr}(K_0C_2(\int UU^\top)C_2^\top).\end{aligned} \tag{7.30}$$

Similarly, one can show that $H(DJD)^{-1}D\partial\ell_T(\theta_T)/\partial\theta$ converges weakly to

$$\begin{aligned}&(I_{k-r}\otimes(0,I_{k-r}))((\int UU^\top)\otimes\tilde A^\top C_1\tilde A)^{-1}(\int U\otimes d\tilde A^\top W_2) = \\ &((\int UU^\top)^{-1}\otimes(0,I_{k-r})\tilde A^{-1}C_1^{-1})(\int U\otimes dW_2).\end{aligned}$$

As a result, the first and second terms in (7.29) converge weakly to

$$\begin{aligned}&(\int U\otimes dW_2)^\top((\int UU^\top)^{-1}\otimes(0,I_{k-r})\tilde A^{-1}C_1^{-1})^\top\cdot \\ &((\int UU^\top)\otimes((0,I_{k-r})(\tilde A^\top C_1\tilde A)^{-1}(0,I_{k-r})^\top)^{-1})\cdot\end{aligned}$$

$$((\int UU^\top)^{-1} \otimes (0, I_{k-r})\tilde{A}^{-1}C_1^{-1})(\int U \otimes dW_2) =$$

$$(\int U \otimes dA_\perp^\top C_1^{-1}W_2)^\top ((\int UU^\top)^{-1} \otimes K_0)(\int U \otimes dA_\perp^\top C_1^{-1}W_2) =$$

$$\mathrm{tr}(K_0(\int Ud(A_\perp^\top C_1^{-1}W_2)^\top)^\top (\int UU^\top)^{-1}(\int Ud(A_\perp^\top C_1^{-1}W_2)^\top)), \qquad (7.31)$$

and

$$2 \cdot (\int U \otimes dA_\perp^\top C_1^{-1}W_2)^\top (I_{k-r} \otimes K_0)\mathrm{vec}(C_2) =$$

$$2 \cdot \mathrm{tr}(K_0 C_2 \int Ud(A_\perp^\top C_1^{-1}W_2)^\top), \qquad (7.32)$$

respectively. Now replacing U and $A_\perp^\top C_1^{-1}W_2$ in (7.30), (7.31), and (7.32) by

$$\hat{U}(s) = S_1^\top (A_\perp^\top \Omega_{11} A_\perp)^{-1/2}(A_\perp^\top \Psi B_\perp)U(s),$$

and

$$\hat{W}_2(s) = S_2^\top (A_\perp^\top C_1^{-1}\Omega_{22}C_1^{-1}A_\perp)^{-1/2}A_\perp^\top C_1^{-1}W_2(s),$$

respectively, one obtains

$$\mathrm{tr}\left(\tilde{C}_2^\top S_1^\top (A_\perp^\top \Omega_{11} A_\perp)^{1/2} K_0 (A_\perp^\top \Omega_{11} A_\perp)^{1/2} S_1 \tilde{C}_2 (\int \hat{U}\hat{U}^\top)\right), \qquad (7.30$$

')

$$\mathrm{tr}\left(\tilde{K}_0(\int \hat{U}d\hat{W}_2^\top)^\top (\int \hat{U}\hat{U}^\top)^{-1}(\int \hat{U}d\hat{W}_2^\top)\right), \qquad (7.31$$

')and

$$2 \cdot \mathrm{tr}\left(\tilde{K}_0 S_2^\top \bar{K}_0^{-1/2}(A_\perp^\top \Omega_{11} A_\perp)^{1/2} S_1 \tilde{C}_2 (\int \hat{U}d\hat{W}_2^\top)\right). \qquad (7.32$$

') It is easily checked that, given the result of Lemma 7.1, $\hat{U}(s)$ satisfies the stochastic differential equation presented in Theorem 7.1. Moreover, let $\hat{W}_1(s) = A_\perp^\top W_1(s)$, then $E(\hat{W}_1(s)\hat{W}_2(s)^\top) = S_1^\top S_0 S_2 = R$. □

Proof of Theorem 7.2. The result follows directly from Theorem 7.1 and the fact that $\hat{W}_2 = R\hat{W}_1(s) + \hat{W}_3(s)$. □

Proof of Corollary 7.1. From the fact that $\psi(\Omega_{11}^{-1/2}\varepsilon_t) = \Omega_{11}^{-1}\varepsilon_t$, it follows that $\Omega_{22} = C_1 = \Omega_{11}^{-1}$. Therefore, R, S_0, S_1, S_2, and $\tilde{K}_0$ reduce to the unit matrix, $\hat{W}_1 = \hat{W}_2$, and $K_0^{-1} = (A_\perp^\top \Omega_{11} A_\perp)$. The result now follows directly from Theorem 7.1. □

Proof of Corollary 7.2. Under the conditions stated in the corollary, the information matrix equality holds, meaning that

$$E(d^2 \ln(f(\varepsilon_t))/(d\varepsilon_t^\top d\varepsilon_t)) = -E((d\ln(f(\varepsilon_t))/d\varepsilon_t)(d\ln(f(\varepsilon_t))/d\varepsilon_t)^\top).$$

For the specific form of $f(\cdot)$ given in the corollary, this implies that $C_1 = \Omega_{22}$. Using this fact and the the fact that $\tilde{C}_2 = 0$, the result follows directly from Theorem 7.2. □

Proof of Lemma 7.2. The only stochastic variable in the first term of (7.18) is K_3, which has expectation $S_1^\top \bar{K}_3 S_1$. Using some elementary matrix manipulations, it is easily shown that the expectation of the first term equals $\mathrm{tr}(K_0\bar{C}_2\bar{K}_3\bar{C}_2^\top)$, with $\bar{C}_2$ as defined in the lemma.

The second term in (7.18) consists of two parts. The first part vanishes, because $E(\int \hat{U}d\hat{W}_3^\top) = 0$ through the independence of $\hat{W}_1$ and $\hat{W}_3$. The second part can be rewritten as

$$2 \cdot \mathrm{tr}\left(K_0\bar{C}_2\bar{K}_1(A_\perp^\top\Omega_{11}A_\perp)^{-1/2}(A_\perp^\top\Omega_{12}C_1^{-1}A_\perp)\right).$$

Let u_1 be a univariate Ornstein-Uhlenbeck process generated by the standard Brownian motion w_1. Let w_2 denote a standard Brownian motion that is uncorrelated with w_1. Then the elements of $\bar{K}_1$ are either of the form $E(\int u_1dw_1)$ or $E(\int u_1dw_2)$. From the independence of w_1 and w_2, it follows easily that $E(\int u_1dw_2) = 0$. Moreover, using the results of Bobkoski (1983), it follows that

$$E(\int u_1dw_1) = -\partial\Lambda(0,0)/\partial s = 0,$$

with Λ as defined in Example 7.2. So one obtains $\bar{K}_1 = 0$.

Similar to the first component of the second term, the second component in the third term in (7.18) vanishes in expectation. The other component can easily be rewritten as

$$\mathrm{tr}(\bar{K}_0^{1/2}K_0\bar{K}_0^{1/2}S_0^\top\bar{K}_2S_0),$$

which equals

$$\mathrm{tr}\left(K_0(A_\perp^\top C_1^{-1}\Omega_{21}A_\perp)(A_\perp^\top\Omega_{11}A_\perp)^{-1/2}\bar{K}_2(A_\perp^\top\Omega_{11}A_\perp)^{-1/2}(A_\perp^\top\Omega_{12}C_1^{-1}A_\perp)\right).$$

In order to compute the expectation of the fourth term, define the sigma algebras $\mathcal{F}_{1s}$ for $0 \leq s \leq 1$ that are generated by $\hat{W}_1(s)$. Due to the independence of $\hat{W}_1$ and $\hat{W}_3$, the conditional distribution of the fourth term in (7.18) given $\mathcal{F}_{1s}$ is χ^2. Therefore the expectation of this term equals $(k-r)\mathrm{tr}(\tilde{K}_0(I_{k-r}-R^2))$, which can be rewritten as $(k-r)\mathrm{tr}(K_0P)$. □

Proof of Theorem 7.3. The Euler-Lagrange equations for the maximization problem are given by

$$\partial L/\partial\psi_i = \sum_{j=1}^{k}\partial^2L/(\partial x_j\partial\psi'_{ij}),$$

for $i = 1,\ldots,k$. Working out these conditions, one obtains the set of equations

$$2fA_\perp K_0A_\perp^\top\Omega_{21}A_\perp(K_5^{-1/2}\bar{K}_2K_5^{-1/2} - (k-r)K_5^{-1})A_\perp^\top\varepsilon+$$
$$2(k-r)fA_\perp K_0A_\perp^\top\psi - f\Lambda_1 + \Lambda_2\dot{f} = 0, \qquad (7.33)$$

where $f = f(\varepsilon)$, $\dot{f} = df/d\varepsilon$, $\psi = \psi(\Omega_{11}^{-1/2}\varepsilon)$, and $\varepsilon = \varepsilon_t$. Integrating (7.33) and using the fact that $E(\psi) = E(\varepsilon) = 0$, one obtains $\Lambda_1 = 0$. Furthermore, premultiplying

(7.33) by $A^\top$, it follows that $A^\top \Lambda_2 = 0$. Finally, premultiplying (7.33) by $A_\perp^\top / f$, one obtains

$$A_\perp^\top \Omega_{21} A_\perp (K_5^{-1/2} \bar{K}_2 K_5^{-1/2} - (k-r) K_5^{-1}) A_\perp^\top \varepsilon + \\ (k-r) A_\perp^\top \psi = \frac{1}{2} A_\perp^\top \Lambda_2 (\dot{f}/f). \tag{7.34}$$

Differentiating (7.34) with respect to $\varepsilon^\top$, taking expectations, and using the fact that $E(\psi') = I_k$, one obtains

$$A_\perp^\top \Omega_{21} A_\perp (K_5^{-1/2} \bar{K}_2 K_5^{-1/2} - (k-r) K_5^{-1}) A_\perp^\top + (k-r) A_\perp^\top = \frac{1}{2} A_\perp^\top \Lambda_2 \mathcal{I}, \tag{7.35}$$

with $\mathcal{I} = -E((d^2 \ln(f))/(d\varepsilon^\top d\varepsilon))$ being the Fisher information matrix. Combining the equations $A^\top \Lambda_2 = 0$ with (7.35), one can solve for Λ_2 and obtain

$$\Lambda_2 = 2 A_\perp K_0 ((k-r) I_{k-r} - \bar{K}_4) A_\perp^\top \mathcal{I}^{-1},$$

with

$$\bar{K}_4 = -A_\perp^\top \Omega_{21} A_\perp (K_5^{-1/2} \bar{K}_2 K_5^{-1/2} - (k-r) K_5^{-1}).$$

Substituting this solution back into (7.34) and solving for ψ, one obtains the result

$$(k-r) A_\perp^\top \psi = \bar{K}_4 A_\perp^\top \varepsilon + (\bar{K}_4 - (k-r) I_k) A_\perp^\top \mathcal{I}^{-1} (\dot{f}/f). \tag{7.36}$$

Note that ψ in (7.36) enters on the right-hand side as well as on the left-hand side, namely in the matrix Ω_{21} (or $\bar{K}_4$). Multiplying (7.36) from the right by $\varepsilon^\top A_\perp$ and taking expectations, one can solve for the value of Ω_{21} (or $\bar{K}_4$). When doing this, note that $E(\dot{f}\varepsilon^\top / f) = -I_k$ due to the fact that f vanishes on the edge of its support (see Assumption 7.4). Substituting the solution for Ω_{21} back into (7.36), the final result is established. □

Proof of Corollary 7.3. Let $\tilde{\psi}^*(\Omega_{11}^{-1/2} \varepsilon_t) = \partial \rho^*(\Omega_{11}^{-1/2} \varepsilon_t) / \partial \varepsilon_t$, then

$$\tilde{\psi}^*(\Omega_{11}^{-1/2} \varepsilon_t) = a \mathcal{I} \varepsilon_t + b \frac{d \ln f(\varepsilon_t)}{d \varepsilon_t}.$$

Define $(k-r)\psi^*(\Omega_{11}^{-1/2} \varepsilon_t) = \mathcal{I}^{-1} \tilde{\psi}^*(\Omega_{11}^{-1/2} \varepsilon_t)$, then the first order conditions implied by the functions ψ^* and $\tilde{\psi}^*$ have the same solution(s). Therefore, they define the same PML estimator. One has that

$$(k-r)\psi^*(\Omega_{11}^{-1/2} \varepsilon_t) = a \varepsilon_t + b \mathcal{I}^{-1} \frac{d \ln f(\varepsilon_t)}{d \varepsilon_t}.$$

Following Section 7.5, $\bar{K}_2 = \bar{k}_2 I_{k-r}$. Therefore,

$$K_4 = -(\bar{k}_2 - (k-r)) A_\perp^\top \Omega_{21}^* A_\perp K_5^{-1}.$$

Using the fact that $A_\perp^\top \mathcal{I}^{-1} A_\perp = k_3 K_5$ for some $k_3 \neq 0$, one further obtains

$$(k-r) A_\perp^\top \Omega_{21}^* A_\perp = -A_\perp^\top \Omega_{21}^* A_\perp (\bar{k}_2 - (k-r)) + k_3 (k-r) K_5 + A_\perp^\top \Omega_{21}^* A_\perp (\bar{k}_2 - (k-r)) k_3,$$

such that

$$A_\perp^\top \Omega_{21}^* A_\perp = k_3 (k-r) K_5 / (\bar{k}_2 - k_3 (\bar{k}_2 - (k-r)))$$

and $K_4 = k_4 I_{k-r}$, with

$$k_4 = -k_3 (k-r)(\bar{k}_2 - (k-r)) / (\bar{k}_2 - k_3 (\bar{k}_2 - (k-r))).$$

The result now follows by setting $a = k_4$ and $b = k_4 - (k-r)$. □