Chapter 5

COINTEGRATION AND TRADING

5.1 Stationarity and Unit Root Tests

In the MSCF Financial Time Series Analysis course, students were introduced to some of the basic models for univariate time series, e.g. ARMA models. The models that were studied were designed for stationary time series. Referring to Section 1.5 of Anthony Brockwell's Financial Time Series notes:

Definition:
If $\{X_t, t \in T\}$ is a time series, then it is *weakly stationary* or *covariance stationary* if

1. $\forall t \in T,\ E(X_t^2) < \infty$,
2. $\forall t \in T,\ E(X_t) = m$,
3. $\forall r, s, u \in T,\ \mathrm{Cov}(X_r, X_s) = \mathrm{Cov}(X_{r+u}, X_{s+u})$

If a process is covariance stationary, then the covariance between two random variables in the stochastic process depends only on the time increment between them. That is, the process is characterized by its *covariance function*:

$$\gamma_X(h) = \mathrm{Cov}(X_t, X_{t+h}).$$

Note also that

$$\sigma^2 = \gamma_X(0) = \mathrm{Cov}(X_t, X_t) = \mathrm{Var}(X_t).$$

Unless otherwise stated, in the Statistical Arbitrage course, the word "stationary" will be assumed to mean "covariance stationary." A stationary process should also be mean reverting. We note in

passing that there is a more restrictive definition of stationarity, *strict stationarity* or *full stationarity.*

Definition:
If $\{X_t, t \in T\}$ is a time series, then it is *strictly stationary* if for integers k and time points $t_1, t_2, \ldots, t_k$and $\in T$, the joint distribution of $(X_{t_1}, \ldots, X_{t_k})$ is the same as $(X_{t_1+h}, \ldots, X_{t_k+h})$.

This definition is very restrictive and is very unlikely to be useful as a model for financial time series. Thus we will restrict attention to covariance stationary processes.

One special class of processes of particular importance is the class of autoregressive processes, AR(p). Consider an autoregressive process with $p = 1$. Then $\{X_t\}$ is given by

$$X_t = \alpha + \beta X_{t-1} + \epsilon_t,$$

where $\{\epsilon_t\}$ is a sequence of i.i.d. random variables with mean 0 and finite variance σ^2. For this process to be stationary, we must have $|\beta| < 1$. By taking first moments and second moments across the above equation, one finds:

$$E(X_t) = \alpha + \beta E(X_{t-1}),$$

and

$$E(X_t^2) = \alpha^2 + \beta^2 E(X_{t-1}^2) + \sigma^2 + 2\alpha\beta E(X_{t-1}).$$

Thus both the first and second moments of the time series satisfy first order difference equations that can be easily solved. For the process to be stationary, we require $\forall t$, $E(X_t)$ and $E(X_t^2)$ must be finite constants independent of t. A necessary condition for this to occur is that $|\beta| < 1$.

To see this we first iterate the difference equation for the mean to find

$$E(X_t) = \alpha \sum_{i=0}^{t-1} \beta^i + \beta^t E(X_0).$$

The $\beta^t E(X_0)$ term converges to 0 if and only if $|\beta| < 1$ or $E(X_0) = 0$. Similarly, $\alpha \sum_{i=0}^{t-1} \beta^i$ converges if and only if $|\beta| < 1$. If $|\beta| < 1$, then $E(X_t) \to \frac{\alpha}{1-\beta}$ as $t \to \infty$.

A similar argument applies to the second moment. Iterating the difference equation and substituting $E(X_t) = \frac{\alpha}{1-\beta}$, we find

$$\begin{aligned} E(X_t^2) &= \alpha^2 + \sigma^2 + \beta^2 E(X_{t-1}^2) + 2\alpha\beta E(X_{t-1}) \\ &= \alpha^2 + \sigma^2 + \beta^2 E(X_{t-1}^2) + 2\alpha\beta \frac{\alpha}{1-\beta} \\ &= \frac{\alpha^2(1+\beta) + \sigma^2(1-\beta)}{1-\beta} + \beta^2 E(X_{t-1}^2) \\ &= C + \beta^2 E(X_{t-1}^2), \end{aligned}$$

where $C = \frac{\alpha^2(1+\beta)+\sigma^2(1-\beta)}{1-\beta}$. The resulting difference equation is of the same form as the difference equation for $E(X_t)$ with α and β replaced by C and β^2 respectively. It can be iterated to find

$$E(X_t^2) = C\sum_{i=0}^{t-1}\beta^{2i} + \beta^{2t}E(X_0^2).$$

Again, $|\beta| < 1$ is required for $E(X_t^2)$ to be independent of t. In this case, $E(X_t^2) \to \frac{\alpha^2}{(1-\beta)^2} + \frac{\sigma^2}{1-\beta^2}$ as $t \to \infty$. Note that v_t is the second moment of X_t, thus the stationary value for the $\mathrm{Var}(X_t)$ is given by $\frac{\sigma^2}{1-\beta^2}$.

An alternative way of expressing the stationarity requirement is that the root of the linear equation $1-\beta x = 0$ must have absolute value larger than 1. The polynomial $1-\beta x$ is called the autoregressive polynomial.

The case in which $\beta = 1$ corresponds to a random walk model in which the increments are i.i.d. random variables with mean α and variance σ^2. This model is non-stationary, and the variance grows at rate t, just like Brownian motion. In fact, properly scaled in time and space, the random walk model converges to a Brownian motion. The random walk model is commonly used to model the log of a stock price under the efficient market hypothesis. The case in which $|\beta| > 1$ corresponds to an explosive process. This case is not especially interesting, so tests of stationarity are generally formulated as $H_0 : |\beta| = 1$ vs $H_1 : |\beta| < 1$.

This approach to determining stationarity extends to the case of an AR(p) process in which the process, $\{X_t\}$, satisfies

$$X_t = \alpha + \sum_{i=1}^{p}\beta_i X_{t-i} + \epsilon_t.$$

In this case, one considers the roots of the autoregressive polynomial:

$$0 = 1 - \sum_{i=1}^{p}\beta_i x^i.$$

The process is stationary if all p roots (some of which may be complex) have modulus larger than 1. That is, when they are plotted in the complex plane, they must lie *outside* the unit circle. The estimation of the coefficients of the time series was considered in the Financial Time Series course. Given those estimates, one can estimate the roots and determine whether the series is stationary. Tests of stationarity are usually referred to as *unit root tests*.

The concept of stationarity can be extended to the multivariate case. One could, for example, define a vector autoregressive process (sometimes denoted by VAR(p)) in a straightforward fashion. Suppose that $\boldsymbol{X}_t = (X_1, \cdots, X_m)$ is an m-dimensional vector time series defined by:

$$\boldsymbol{X}_t = \boldsymbol{a}_0 + \sum_{i=1}^{p}\boldsymbol{A}_i\boldsymbol{X}_{t-i} + \boldsymbol{\epsilon}_t,$$

where $\boldsymbol{a}_0$ is an m-dimensional vector and $(\boldsymbol{A}_1, \cdots, \boldsymbol{A}_p)$ are $m \times m$ matrices and $\{\boldsymbol{\epsilon}_t\}$ are i.i.d. m-dimensional random vectors with mean $\mathbf{0}$ and some covariance matrix $\boldsymbol{\Gamma}$.

One could address the stationarity of the vector process by considering each component individually; however, we are concerned with *joint stationarity*. Nevertheless, one would ordinarily test to ensure that each of the component series is stationary, since if at least one is non-stationary, then the vector process will also be non-stationary. Suppose we first consider the simplest case with $p = 1$ so

$$\boldsymbol{X}_t = \boldsymbol{a} + \boldsymbol{A}\boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_t.$$

or for m = 2:

$$\begin{pmatrix} X_{1,t} \\ X_{2,t} \end{pmatrix} = \begin{pmatrix} a_1 \\ a_2 \end{pmatrix} + \begin{pmatrix} a_{11}a_{12} \\ a_{21}a_{22} \end{pmatrix} \begin{pmatrix} X_{1,t-1} \\ X_{2,t-1} \end{pmatrix} + \begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{pmatrix}.$$

A necessary condition for stationarity is analogous to the one-dimensional ($m = 1$) case, except we now set up multivariate difference equations for the mean vector and the covariance matrix for the time series. Rather than requiring restrictions on the individual components of the matrix $\boldsymbol{A}$, we need to ensure that the eigenvalues of $\boldsymbol{A}$ (which may be complex) are inside the unit circle, i.e. have modulus strictly less than 1.

The covariance matrix of $\boldsymbol{X}_t$, $\boldsymbol{\Sigma}_t$, is defined by $\boldsymbol{\Sigma}_t = E(\boldsymbol{X}_t\boldsymbol{X}_t^\top)$ (assuming the means are 0). Recalling that the means of the $\boldsymbol{\epsilon}_t$ and $\boldsymbol{X}_t$ vectors are 0, one can find an expression for $\boldsymbol{\Sigma}_t$ in terms of $\boldsymbol{\Sigma}_{t-1}$ and iterate it as follows:

$$\begin{aligned}
\boldsymbol{\Sigma}_t &= E(\boldsymbol{X}_t\boldsymbol{X}_t^\top) \\
&= E((\boldsymbol{a} + \boldsymbol{A}\boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_t)(\boldsymbol{a} + \boldsymbol{A}\boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_t)^\top) \\
&= \boldsymbol{a}\boldsymbol{a}^\top + \boldsymbol{A}E(\boldsymbol{X}_{t-1}\boldsymbol{X}_{t-1}^\top)\boldsymbol{A}^\top + E(\boldsymbol{\epsilon}_t\boldsymbol{\epsilon}_t^\top) \\
&= \boldsymbol{a}\boldsymbol{a}^\top + \boldsymbol{A}\boldsymbol{\Sigma}_{t-1}\boldsymbol{A}^\top + \boldsymbol{\Gamma} \\
&= \boldsymbol{a}\boldsymbol{a}^\top + \boldsymbol{A}(\boldsymbol{a}\boldsymbol{a}^\top + \boldsymbol{A}\boldsymbol{\Sigma}_{t-2}\boldsymbol{A}^\top + \boldsymbol{\Gamma})\boldsymbol{A}^\top + \boldsymbol{\Gamma}.
\end{aligned}$$

If one continues this iteration back to $t = 0$, $\boldsymbol{\Sigma}_t$ will contain the term $\boldsymbol{A}^t\boldsymbol{\Sigma}_0(\boldsymbol{A}^\top)^t$ and a power series involving $\boldsymbol{\Gamma}$ and $\boldsymbol{A}\boldsymbol{A}^\top$ including terms of the form $\sum_{j=0}^{t-1}\boldsymbol{A}^j(\boldsymbol{a}\boldsymbol{a}^\top + \boldsymbol{\Gamma})(\boldsymbol{A}^\top)^j$. Now we need $\boldsymbol{A}^t$ to converge to 0 as $t \to \infty$ so that the power series converges and the term with $\boldsymbol{\Sigma}_0$ vanishes. This is where the eigenvalues of $\boldsymbol{A}$ play a role. If $\boldsymbol{A}$ can be diagonalized, then $\boldsymbol{A} = \boldsymbol{R}\boldsymbol{D}\boldsymbol{L}$, where $\boldsymbol{D}$ is a diagonal matrix of the eigenvalues and $\boldsymbol{R}$ and $\boldsymbol{L}$ are matrices of the eigenvectors which are inverses of each other. Then $\boldsymbol{A}^t$ can be written as $\boldsymbol{A}^t = \boldsymbol{R}\boldsymbol{D}^t\boldsymbol{L}$. Obviously if one of the eigenvalues is on the unit circle, it will not vanish as it is powered up. If any of the eigenvalues lies outside the unit circle, then the corresponding term on the diagonal of $\boldsymbol{D}^t$ will explode. So all eigenvalues must be inside the unit circle for the second moment to behave properly.

If we simplified by setting $\boldsymbol{a} = 0$ and $\boldsymbol{\Gamma} = \sigma^2\boldsymbol{I}$, then

$$\boldsymbol{\Sigma}_t \to \sigma^2(\boldsymbol{I} - \boldsymbol{A}\boldsymbol{A}^\top)^{-1}.$$

This is the multivariate analog to the result we found earlier for the univariate case for which $\mathrm{Var}(X_t) \to \frac{\sigma^2}{1-\beta^2}$.

The eigenvalues are given as the roots of the equation $\det(\boldsymbol{A} - \lambda \boldsymbol{I}) = 0$. If one were to define $x = \frac{1}{\lambda}$, then x would be one of the roots of the equation $\det(\boldsymbol{I} - \boldsymbol{A}x) = 0$. In this case, for stationarity we require that the modulus of λ be smaller than 1 or equivalently that the modulus of x be greater than 1.

The stationarity condition for a VAR(p) process can be stated similarly. Stationarity requires that the roots of the polynomial equation

$$\det(\boldsymbol{I} - \sum_{i=1}^{p} \boldsymbol{A}_i x^i) = 0$$

to have modulus outside the unit circle. (The *R* library `vars` has a function `VAR` that will estimate VAR models. All *R* libraries can be found on the web site
`http://cran.r-project.org/`)

5.2 Differencing to Obtain Stationarity

If the time series in question represents a stock price or stock returns process, then for purposes of statistical arbitrage and trading, we are interested not in a single process but multiple processes considered simultaneously. In addition, we cannot reasonably expect a stock price or returns series to be stationary. In the case in which a time series is non-stationary, the most obviously first step is to transform it so that the transformed series is stationary.

Methods to transform a non-stationary time series into a stationary time series are discussed in Section 1.6 of the Financial Time Series notes. For purposes of the statistical arbitrage class, the most important uses differencing. Specifically, define the backshift (or lag) operator, B, and the differencing operator, ∇, by

$$B^k(X_t) = B^k X_t \triangleq X_{t-k},$$

and

$$\nabla X_t \triangleq X_t - X_{t-1} = X_t - BX_t = (1 - B)X_t.$$

One can iterate the differencing operator. For example:

$$\begin{aligned}
\nabla^2 X_t &= \nabla(\nabla X_t) \\
&= \nabla(X_t - X_{t-1}) \\
&= (X_t - X_{t-1}) - (X_{t-1} - X_{t-2}) \\
&= X_t - 2X_{t-1} + X_{t-2} \\
&= X_t - 2BX_t + B^2 X_t \\
&= (1 - 2B + B^2)X_t \\
&= (1 - B)^2 X_t
\end{aligned}$$

This calculation can be iterated to find

$$\nabla^k X_t = (1 - B)^k X_t.$$

For a given non-stationary time series, $\{X_t\}$, it may be possible to difference it repeatedly some number of times and have it become stationary.

Definition
A univariate time series, $\{X_t\}$, is *integrated of order k* (denoted by $I(k)$), if $\{\nabla^{k-1}\}$ is non-stationary (hence $\nabla^j X_t$ is also non-stationary for $j = 0, \ldots, k-1$); however, $\{\nabla^k X_t\}$ is stationary.

We now turn to the multivariate case. Suppose $\boldsymbol{X}_t = (X_1, X_2, \cdots, X_m)^\top$ is an m-dimensional random vector, thus $\{X_t\}$ is an m-dimensional time series. This time series is *integrated of order k* (denoted $I(k)$) provided at least one of its coordinates is $I(k)$ and all the others are $I(j)$ for $j \leq k$.

5.3 Dickey-Fuller Unit Root Test

In this section, we consider tests of stationarity of a time series. Recall we have associated stationarity with eigenvalues having modulus less than 1 or roots of a autoregressive polynomial equation being larger than 1 in modulus.

To develop the ideas, we begin with the simplest case, an AR(1) process:

$$X_t = \beta X_{t-1} + \epsilon_t,$$

where $\{\epsilon_t\}$ is a sequence of i.i.d. random variables with mean 0 and variance σ^2. Recall that stationarity of the time series is equivalent to $|\beta| < 1$, i.e. we want to test the hypothesis $H_0 : |\beta| = 1$ versus $H_1 : |\beta| < 1$, non-stationarity versus stationarity. If one had time series data, $\{X_t,\ 0 \leq t \leq T\}$, one could formulate a hypothesis by considering either of two regression problems: (1) regressing X_t on X_{t-1} and setting up a "t-statistic," or (2) subtracting X_{t-1} from both sides to obtain

$$\nabla X_t = (\beta - 1) X_{t-1} + \epsilon_t,$$

then regressing ∇X_t on X_{t-1} and setting up a "t-statistic."

It turns out that both approaches give identical estimates of β and identical "t-statistics;" hence either approach can be taken. However, under the random walk model ($\beta = 1$), which has been set up as the null hypothesis, the sampling distribution of this "t-statistic" is **not** given by the standard Student-t distribution (hence the use of quotation marks). The appropriate sampling distribution, the critical values for the test, and the power calculations must be determined either through Monte Carlo simulation or from an asymptotic limiting distribution as $T \to \infty$.

To be specific, the least squares estimate of the coefficient β is given by

$$\hat{\beta} = \frac{\sum_{t=1}^{\top} X_{t-1} X_t}{\sum_{t=1}^{\top} X_{t-1}^2}.$$

If we define the s_T^2 to be the usual least squares estimator of σ^2, i.e.

$$s_T^2 = \frac{1}{T-1}\sum_{t=1}^{\top}(X_t - \hat{\beta}X_{t-1})^2,$$

then the usual t-statistic for testing $H_0 : \beta = 1$ versus $H_1 : \beta < 1$ would be given by

$$t_T = \frac{\hat{\beta} - 1}{\sqrt{s_T^2 / \sum_{t=1}^{\top} X_{t-1}^2}}.$$

It can be shown that under H_0:

$$\sqrt{T}(\hat{\beta} - 1) \to 0 \text{ in probability,}$$

so β is a "superconsistent" estimate. Ordinarily one expects a consistent estimator to converge to the true value (1 in this case) in probability fast enough so that when the difference, $\hat{\beta} - 1$, when multiplied by the square root of the sample size, $\sqrt{T}$, converges to some *bona fide* limiting probability distribution like a normal or a Student-t. However, in this case, $\sqrt{T}(\hat{\beta}-1)$ also converges in probability to 0 indicating that the convergence of $\hat{\beta}$ to 1 is faster than $\sqrt{T}$. This is the reason for the "superconsistent" characterization.

It turns out that the asymptotic distribution of t_T is given by:

$$t_T \Rightarrow \frac{\frac{1}{2}(W(1)^2 - 1)}{\sqrt{\int_0^1 W(x)^2 dx}},$$

where $\{W(t)\}$ are a standard Wiener process. The Wiener processes in the numerator and denominator is the same, thus the distribution of this random variable, which is needed to determine p-values or critical values, must be calculated by simulation. Of course, students who have taken the Monte Carlo simulation course have substantial experience simulating the paths of a Brownian motion. One need only simulate a path and calculate the ratio, then repeat this many times to determine the distribution.

It is interesting to note that the distribution is skewed to the left. Since the denominator is positive, the ratio is negative if and only if the numerator is negative. Hence, the probability that the t_T statistic is negative is given by $\mathrm{P}(|W(1)| < 1)$ which is approximately .68. A Student-t distribution is symmetric about 0, hence this probability would be .5.

Dickey and Fuller presented a number of different test statistics, all of which go by the name "Dickey-Fuller tests for unit roots." A more common version for this same problem is to use the related, but different test statistic, $T(\hat{\beta}-1)$. The asymptotic distribution is similar to the t-statistic version, namely:

$$T(\hat{\beta} - 1) \Rightarrow \frac{\frac{1}{2}(W(1)^2 - 1)}{\int_0^1 W(x)^2 dx}.$$

The asymptotic distribution will provide approximate p-values. For exact work, one must use Monte Carlo simulation tailored to the specific problem (i.e. the sample size and the structure of the

$\{\epsilon_t\}$ process). The calculation of the p-value associated with a particular data set of size T is done under the least favorable point in the null hypothesis, namely $\beta = 1$. Using Monte Carlo simulation (described next), generate a large number, n, of time series of the form $X_t = X_{t-1} + \epsilon_t, 0 \leq t \leq T$ where the ϵ_t are generated independently with some distribution, typically Normal$(0, \sigma^2)$. For each simulated series, one computes the value of the "t-statistic." The n such values generated by simulation give the distribution of the "t-statistic." Under H_0, the p-value of the actual time series can be determined by finding its quantile in the simulated distribution.

One possible conclusion of the Dickey-Fuller (DF) hypothesis test is that the null hypothesis cannot be rejected. Therefore, one would not treat the series as being stationary, i.e. it is not $I(0)$. We could next check to see if it is $I(1)$, i.e. we could test $H_0 : X_t$ is $I(2)$ versus $H_1 : X_t$ is $I(1)$. In this case, we would regress $\nabla^2 X_t$ on ∇X_t, i.e.

$$\nabla^2 X_t = \alpha + \beta \nabla X_{t-1} + \epsilon_t,$$

and one would use the same Dickey-Fuller test. This approach could be continued if the null hypothesis continues to not be rejected.

The above presentation has been restricted to the case of AR(1) processes. The Dickey-Fuller methodology has been extended to handle AR(k) processes with $k > 1$. If we begin with the model

$$X_t = \alpha + \sum_{i=1}^{k} \beta_i X_{t-i} + \epsilon_t,$$

then by subtracting X_{t-1} from both sides, rearranging terms and redefining the model coefficients, we arrive at the model:

$$\nabla X_t = \alpha + \sum_{i=1}^{k-1} a_i \nabla X_{t-i} + a_k X_{t-k} + \epsilon_t.$$

Here, the new coefficients are $a_i = -1 + \sum_{j=1}^{i} \beta_j$. One would calculate the ordinary least squares estimate of a_k and test whether it is larger or smaller than 1 in modulus. This test is called the augmented Dickey-Fuller test. Again critical values can be calculated using Monte Carlo simulation. The *R* library `urca` contains the function `ur.df` that will compute critical values for the test at levels of significance 0.01, 0.05 and 0.10.

5.4 Cointegration

We now turn to cointegration, a possible reference for which is Chapter 12 of *Market Models* by Alexander[1]. In addition, the paper by Engle and Granger[2] on the Blackboard course site lays the basic foundations for this topic.

As Engle and Granger state in the introduction to their paper:

[1] Alexander, C., *Market Models: A Guide to Financial Data Analysis*, Wiley, 2001.

[2] Engle, R.F. and Granger, C.W.J., "Co-Integration and Error Correction: Representation, Estimation, and Testing." *Econometrica*, **55**, 1987, 251-276.

> An individual economic variable, viewed as a time series, can wander extensively and yet some pairs of series may be expected to move so that they do not drift too far apart. Typically economic theory will propose forces which tend to keep such series together. Examples might be short and long term interest rates, capital appropriations and expenditures, household income and expenditures, and prices of the same commodity in different markets or close substitutes in the same market.

In statistical arbitrage, the variables in question are stock prices. These price processes cannot be expected to be stationary but will vary, potentially widely. The crucial idea is to find a subset of these processes that "move together" in the sense that some linear combination is a stationary process. If this is the case, then over the long run the processes are "tied together," and make good candidates for a pairs trading strategy.

In their paper (p253), Engle and Granger introduce the following definition of cointegration.

Definition:
"The components of the vector X_t are said to be *co-integrated of order d, b,* denoted $X_t \sim CI(d,b)$, if (i) all components of X_t are $I(d)$; (ii) there exists a vector $\alpha \neq 0$ so that $Z_t = \alpha' X_t \sim I(d-b), b > 0$. The vector α is called the *cointegrating vector*."

Note that arbitrary constants can be added to the individual time series without altering the cointegration relationship since the addition of a constant does not alter stationarity. An illustrative, although abstract, example is given in the text by Alexander on page 351. She defines three time series X_t, Y_t, W_t from three independent sequences of i.i.d. random variables, $\{\epsilon_{X,t}\}, \{\epsilon_{Y,t}\}, \{\epsilon_t\}$. Define

$$X_t = W_t + \epsilon_{X,t},$$

$$Y_t = W_t + \epsilon_{Y,t},$$

$$W_t = W_{t-1} + \epsilon_t.$$

Thus X_t and Y_t are noisy versions of W_t, and W_t is a random walk. Note that $\nabla X_t = W_t + \epsilon_{X,t} - W_{t-1} - \epsilon_{X,t-1} = \epsilon_{X,t} - \epsilon_{X,t-1} + \epsilon_t$, which is stationary. Hence X_t is $I(1)$. Similarly $\nabla Y_t = W_t + \epsilon_{Y,t} - W_{t-1} - \epsilon_{Y,t-1} = \epsilon_{Y,t} - \epsilon_{Y,t-1} + \epsilon_t$, which is also stationary. Thus Y_t is also $I(1)$.

Clearly X_t and Y_t are cointegrated since both time series are $I(1)$ and $X_t - Y_t = \epsilon_{X,t} - \epsilon_{Y,t}$ which is a stationary noise process. Even though these two series are cointegrated, Alexander shows that they can have very low correlation when either $\mathrm{Var}(\epsilon_{X,t})$ or $\mathrm{Var}(\epsilon_{Y,t})$ is much larger than $\mathrm{Var}(\epsilon_t)$. In particular, $\mathrm{Var}(\nabla X_t) = \mathrm{Var}(\nabla W_t + \nabla \epsilon_{X,t}) = \sigma^2 + 2\sigma^2_{X,t}$. Similarly, $\mathrm{Var}(\nabla Y_t) = \mathrm{Var}(\nabla W_t + \nabla \epsilon_{Y,t}) = \sigma^2 + 2\sigma^2_{Y,t}$. Lastly, $\mathrm{Cov}(\nabla X_t, \nabla Y_t) = \mathrm{Cov}(\nabla W_t + \nabla \epsilon_{X,t}, \nabla W_t + \nabla \epsilon_{Y,t}) = \mathrm{Cov}(W_t, W_t) = \sigma^2$. It follows that the correlation between ∇X_t and ∇Y_t is

$$\frac{\sigma^2}{\sqrt{\left[\sigma^2\left(1+\frac{\sigma^2_{X,t}}{\sigma^2}\right)\right]\left[\sigma^2\left(1+\frac{\sigma^2_{Y,t}}{\sigma^2}\right)\right]}} = \frac{1}{\sqrt{\left(1+\frac{\sigma^2_{X,t}}{\sigma^2}\right)\left(1+\frac{\sigma^2_{Y,t}}{\sigma^2}\right)}}. \tag{5.1}$$

If either ratio $\frac{\sigma^2_{X,t}}{\sigma^2}$ or $\frac{\sigma^2_{Y,t}}{\sigma^2}$ is large, the resulting correlation coefficient is small. Consequently, the fact that variables are cointegrated does not mean that they are highly correlated.

When two time series, $X_{1,t}$ and $X_{2,t}$, are cointegrated, there is a unique cointegrating vector of the form $(1 \;\; \alpha)$. The uniqueness follows, since if there were two such vectors, $(1, \alpha_1)$ and $(1, \alpha_2)$ then $(X_1 + \alpha_1 X_2) - (X_1 + \alpha_2 X_2) = (\alpha_1 - \alpha_2)X_2$ would be stationary in contradiction to the assumption that X_2 was not stationary. This uniqueness no longer holds when $m > 2$. There may be as many as $m - 1$ linearly independent cointegrating vectors. Those vectors will span the cointegration space; however, the choice of vectors representing the cointegration space is not unique.

In the statistical arbitrage course, we will use cointegration as a way of selecting pairs (or even portfolios) in pairs trading strategies. For the selection of pairs, there are two important issues: 1) the degree of cointegration and 2) the selection of a cointegrating vector that will lead to a market neutral pairs trading strategy. These two topics are discussed next.

5.5 Tests for Cointegration

There are two important approaches to testing for cointegration, one proposed by Engle and Granger (1987) and another proposed by Johansen (1988, 1991).

There are two parts to the definition of cointegration. First, the individual components must be $I(d)$. Second, there must be a linear combination of those components that is $I(d - b)$ for some integer $b > 0$. Testing of the first requirement that each component is $I(d)$ was considered in the previous section, say using a Dickey-Fuller test. Thus we turn to the second issue, finding a linear combination of variables that is $I(d - b)$.

The Engle and Granger method has two steps. Since cointegration refers to having a linear combination of the components that is stationary, the Engle and Granger approach uses linear regression to find the best fitting linear relationship. One of the m component variables is selected as the dependent variable, say X_j, and the model

$$X_{t,j} = \beta_0 + \sum_{i \neq j} \beta_i X_{t,i} + \epsilon_t,$$

is fit using ordinary least squares. One would then test whether the residuals were stationary using a unit root test like the Dickey-Fuller test. If so, then we would have a cointegrating vector estimated using least squares. This process could be repeated with other variables taken to be the dependent variable. The difficulty that arises is that the results from the different regressions can be inconsistent. Consider, for example, the case of two variables, $m = 2$. Assuming that both $X_{1,t}$ and $X_{2,t}$ are $I(1)$, one could use the regression

$$X_{2,t} = \alpha_2 + \beta_1 X_{1,t} + \epsilon_t.$$

or the regression

$$X_{1,t} = \alpha_1 + \beta_2 X_{2,t} + \epsilon_t,$$

One can fit either regression model. Before doing so, to simplify the situation we first standardize the X_1 and X_2 data so that $\sum_{i=1}^{\top} X_{1,t} = \sum_{i=1}^{\top} X_{2,t} = 0$ and $\sum_{i=1}^{\top} X_{1,t}^2 = \sum_{i=1}^{\top} X_{1,t}^2 = 1$. In this case, $\hat{\beta}_1 = \hat{\beta}_2 = \sum_{i=1}^{\top} X_{1,t} X_{2,t}$. Furthermore, the residual sum of squares is given by $1 - \hat{\beta}_1^2 = 1 - \hat{\beta}_2^2$ for the two regressions. Consequently, the R^2 value will be identical for the two regressions, and both regressions will give identical statistical inferences concerning whether cointegration is present or not. Nevertheless, the two different approaches would give $(\hat{\beta}_1 \quad 1)$ and $(1 \quad \hat{\beta}_2)$ respectively as cointegrating vectors. This would lead to the formation of two different pairs in a pairs trading strategy, only one of which would be market neutral except in the relatively rare case in which $\hat{\beta}_1 = \hat{\beta}_2 = 1$. Consequently, the reader should exercise some caution when reading the following quote from Alexander on page 355 as it might pertain to pairs trading: "When $n = 2$ it does not matter which variable is taken as the dependent variable. There is only one cointegrating vector, which is the same when estimated by a regression of x on y as when estimated by a regression of y on x." Alexander goes on to say "But when there are more than two $I(1)$ series the Engle-Granger method can suffer from a serious bias. That is, different estimates of a cointegrating vector are obtained depending on the choice of dependent variable, and only one estimate is possible even though there can be up to $n - 1$ cointegrating vectors."

In general, the regression analysis approach will identify linear combinations that minimize least squares deviations; however, minimization of variance is not, in general, associated with finding stationarity in the residual process. For this reason, the approach of Johansen, which is designed to find the linear combination of variables that gives stationarity has gained favor. The Johansen methodology will be discussed next.

The Johansen Methodology

What follows summarizes the Johansen methodology for the Vector Autoregressive case. The material is taken from Section 12.2.2 of *Market Models* and the 1988 paper by Johansen[3].

We begin with a VAR(1) process,

$$\boldsymbol{X}_t = \boldsymbol{a} + \boldsymbol{A}\boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_t.$$

Now, subtracting $\boldsymbol{X}_{t-1}$ from both sides, we find

$$\nabla \boldsymbol{X}_t = \boldsymbol{a} + \boldsymbol{A}\boldsymbol{X}_{t-1} - \boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_t = \boldsymbol{a} + (\boldsymbol{A} - \boldsymbol{I})\boldsymbol{X}_{t-1} + \boldsymbol{\epsilon}_t.$$

If each component of $\boldsymbol{X}$ is $I(1)$, then $\nabla \boldsymbol{X}$ is stationary. Thus the right side of the above equation must also be stationary. But the term $(\boldsymbol{A} - \boldsymbol{I})\boldsymbol{X}_{t-1}$ is stationary only if the rows of the matrix $(\boldsymbol{A} - \boldsymbol{I})$ are cointegrating vectors. That is, the term $(\boldsymbol{A} - \boldsymbol{I})\boldsymbol{X}_{t-1}$ must consist of linear combinations of $\boldsymbol{X}_t$ that are stationary. If $(\boldsymbol{A} - \boldsymbol{I})$ has rank 0 (and is the $\boldsymbol{0}$ matrix) there are no cointegrating vectors to be found. If $(\boldsymbol{A} - \boldsymbol{I})$ has rank R (for some $1 \leq R \leq n - 1$), then there are R linearly independent combinations of the components of $\boldsymbol{X}$ which are stationary. That is, there are R linearly independent cointegrating vectors. This formulation leads us to the question of determining the rank of the matrix $(\boldsymbol{A} - \boldsymbol{I})$, or put equivalently, determining the number of eigenvalues that

[3]Johansen, S., "Statistical analysis of cointegration vectors," *Journal of Economic Dynamics and Control*, **12**, 1988, 231-254.

are 0. One part of the Johansen methodology is based on making a statistical inference about the rank of this matrix.

The idea described for the vector AR(1) process easily carries over to the vector AR(p) case. We begin with the model:

$$\boldsymbol{X}_t = \boldsymbol{a} + \sum_{i=1}^{p} \boldsymbol{A}_i \boldsymbol{X}_{t-i} + \boldsymbol{\epsilon}_t.$$

In this model, p is the order of the vector autoregression, n is the number of components of $\boldsymbol{X}$, and T is the data sample size. Subtracting $\boldsymbol{X}_{t-1}$ from both sides, then sequentially subtracting and adding $(\sum_{j=1}^{k} \boldsymbol{A}_j - \boldsymbol{I})\boldsymbol{X}_{t-k}$ from the right side, we find:

$$\begin{aligned}
\nabla \boldsymbol{X}_t &= \boldsymbol{a} + (\boldsymbol{A}_1 - \boldsymbol{I})\nabla \boldsymbol{X}_{t-1} + (\boldsymbol{A}_1 - I)\boldsymbol{X}_{t-2} + \sum_{i=2}^{p} \boldsymbol{A}_i \boldsymbol{X}_{t-i} + \boldsymbol{\epsilon}_t \\
&= \ldots \\
&= \boldsymbol{a} + \sum_{j=1}^{p-1} \boldsymbol{B}_j \nabla \boldsymbol{X}_{t-j} + \boldsymbol{B}_p X_{t-p} + \boldsymbol{\epsilon},
\end{aligned}$$

where $\boldsymbol{B}_i = (\boldsymbol{A}_1 + \ldots + \boldsymbol{A}_i - \boldsymbol{I})$. We again want the left side to be stationary, but the term $\boldsymbol{B}_p \boldsymbol{X}_{t-p}$ must be consistent with stationarity. If the rank of $\boldsymbol{B}_p$ is R, then this implies there must be R linearly independent combinations of the components that are stationary, i.e. the $\boldsymbol{X}$ must be cointegrated with R linearly independent cointegrating vectors.

The Johansen procedure estimates the eigenvalues of the matrix $\boldsymbol{B}_p$ (ordered by size, call them $\hat{\lambda}_{(1)} \geq \hat{\lambda}_{(2)}, \ldots, \geq \hat{\lambda}_{(n)}$), and then makes a series of one-sided hypothesis tests: $H_0 : r \leq R$ vs. $H_1 : r > R$. The test statistic that is used is $-T \sum_{i=R+1}^{n} \ln(1 - \hat{\lambda}_{(i)})$. The null-hypothesis (of r or fewer cointegration relations) is rejected if the test statistic is large, indicating that at least one of the smaller eigenvalues is not very near 0.

Once the number of cointegration relations has been established, the Johansen methodology goes on to estimate the specific cointegrating vectors. Of course, students in the statistical arbitrage class will rely on an appropriate statistical package (like the the `urca` library for R) to automate the calculation. It is, however, appropriate to caution that Johansen's work and the common statistical package implementations of it will assume that the errors in the statistical model have a multivariate normal distribution. Thus the user should be a bit cautious as many financial time series are known to have tails that are heavier than the normal distribution would imply.

5.6 Cointegrated Price Series

When we discussed pairs trading in Chapter 4, we illustrated a methodology for choosing pairs based on how close the series of standardized log-prices were to each other. This approach is rather *ad hoc* and not based on any theoretical foundation. For example, suppose that the logarithm of

the ratio of two price series is a random walk. To be more specific, suppose that the two price series are X and W. Suppose that for each time i,

$$\begin{aligned}\log(X_i) &= \log(X_{i-1}) + Y_i + \epsilon_{x,i},\\ \log(W_i) &= \log(W_{i-1}) + Y_i + \epsilon_{w,i},\end{aligned}$$

where the $\epsilon_{x,i}$ and $\epsilon_{y,i}$ are all independent random variables. This model says that the returns on both stocks are a common component (represented by Y) and a specific component. Then $\log(X_i/W_i)$ is a random walk. Regardless of how small the variance is of this random walk, there is no reason to expect it to behave in any predictable way. The two series X and W would appear to move together rather closely, but it would be pure luck if you could make any money trading them.

On the other hand, if $\log(X_i/W_i)$ is a stationary process, then one will expect it to return to a certain level eventually after it diverges from that level. One cannot always hope for such a simple relationship, but it may be the case that some linear function, e.g., $a_x \log(X_i) + a_w \log(W_i)$, is a stationary process. In this case, the two log price processes are cointegrated, and we could try to make use of this fact to trade.

Unfortunately, there is no position that we could open whose value is $a_x \log(X_i) + a_w \log(W_i)$. Nor is there any function of this stationary process that could be the value of a position. A linear function of X and W could be the value of a position, but log-prices have greater potential for cointegration due to the prevalence of heavy tails in actual stock prices. Also, the methods for identifying cointegrating relationships are based on an assumption that the input series have nearly normal distributions.

One approach would be to linearize the logarithm. Suppose that we open the position at $i = t_0$. Since

$$\log(x_t) \approx \log(x_{t_0}) + \frac{x_t - x_{t_0}}{x_{t_0}},$$

we can argue that

$$a_x \log(X_t) + a_w \log(W_t) \approx a_x \log(X_{t_0}) + a_w \log(W_{t_0}) + a_x \frac{X_t - X_{t_0}}{X_{t_0}} + a_w \frac{W_t - W_{t_0}}{W_{t_0}}. \tag{5.2}$$

We could open a position with ca_x/X_{t_0} shares of X and ca_w/W_{t_0} shares of W at time t_0, and its value would be approximately stationary by (**??**). An alternative linear approximation is to regress $a_x \log(X) + a_w \log(W)$ on X and W and use the resulting fitted process as an approximation to the supposedly stationary process.

Sometimes, one finds a set of three or more stocks that are cointegrated. One could extend the previous approximation to create a portfolio of more than two stocks to trade at once. For example, if $a_x \log(X_t) + a_w \log(W_t) + a_v \log(V_t)$ is a stationary linear combination of three log price series, we can approximate it by

$$a_x \log(X_{t_0}) + a_w \log(W_{t_0}) + a_v \log(V_{t_0}) + a_x \frac{X_t - X_{t_0}}{X_{t_0}} + a_w \frac{W_t - W_{t_0}}{W_{t_0}} + a_v \frac{V_t - V_{t_0}}{V_{t_0}}.$$

Similarly, we could regress $a_x \log(X) + a_w \log(W) + a_v \log(V)$ on (X, W, V) as a linear approximation.

5.7 An Example

Consider the drug company data from Chapter 4. One of the pairs that we traded was TGEN and INSM. The two log-price series for these stocks both appear to have unit roots, meaning that they are not stationary. The most popular unit root test is called the augmented Dickey-Fuller test (ADF). This test is performed by *R* using the *urca* library.

```
> library(urca)
> TGEN=PRICE[253:504,which(names(PRICE)=="TGEN")]
> INSM=PRICE[253:504,which(names(PRICE)=="INSM")]
> ur.tgen=ur.df(log(TGEN),lags=0,type="drift")
> attr(ur.tgen,"teststat")
               tau2     phi1
statistic -1.171476 1.370057
> attr(ur.tgen,"cval")
      1pct  5pct 10pct
tau2 -3.44 -2.87 -2.57
phi1  6.47  4.61  3.79
> ur.insm<-ur.df(log(INSM),lags=0,type="drift")
> attr(ur.insm,"teststat")
               tau2     phi1
statistic -1.344748 2.599871
> attr(ur.insm,"cval")
      1pct  5pct 10pct
tau2 -3.44 -2.87 -2.57
phi1  6.47  4.61  3.79
```

The ADF test does not reject the null hypothesis that there is a unit root. Plots of the ACF and PACF (Figures **??** and **??** for TGEN) suggest the existence of a unit root as well. The plots for INSM look pretty much the same.

Let's examine this pair for cointegration. The Granger-Engle method has two steps. First, we estimate the relationship

$$a_x \log(X_i) + a_w \log(W_i) = \epsilon_i. \tag{5.3}$$

Second, we examine the series $\epsilon_1, \epsilon_2, \ldots$ for stationarity. There are two natural ways to estimate the relationship, namely by regressing $\log(X_i)$ on $\log(W_i)$ and by regression $\log(W_i)$ on $\log(X_i)$. Although the estimated regression relationships are related, the residuals are not so nicely related. Hence, one typically does both regressions and checks both sets of residuals for stationarity.

To be more clear, in order for (**??**) to be a cointegration relationship, we must have both a_x and a_w nonzero. Hence, (**??**) is stationary if and only if $\log(X_i) + (a_w/a_x)\log(W_i)$ and $(a_x/a_w)\log(X_i) + \log(W_i)$ are both stationary. So, a regression of $\log(X_i)$ and $\log(W_i)$ estimates (a_w/a_x) and a

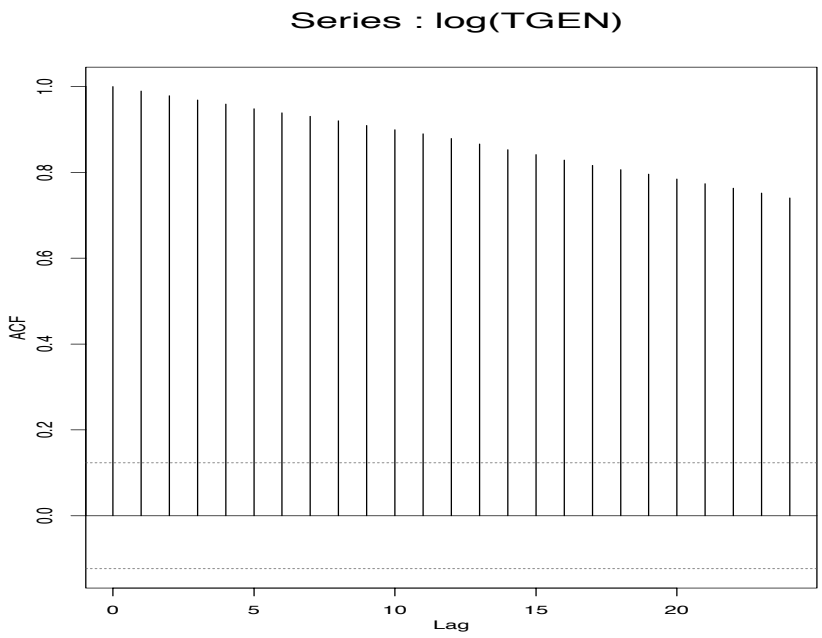

Figure 5.1: ACF plot for TGEN, 2003

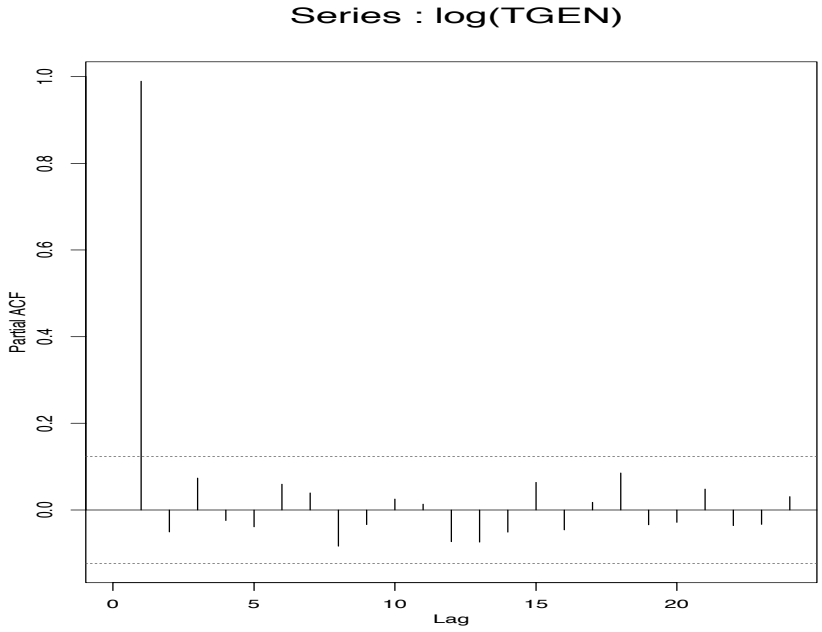

Figure 5.2: PACF plot for TGEN, 2003

regression of $\log(W_i)$ on $\log(X_i)$ estimates (a_x/a_w). Because the relationship (**??**) is not exact, the estimated regression coefficients are not reciprocals of each other, but we are not concerned with that right now. Let $e_{x,i}$ $(i = 1, \ldots, n)$ be the residuals from the regression of $\log(X_i)$ on $\log(W_i)$, and let $e_{w,i}$ be the residuals from the reverse regression. We then run a test of the null hypotheses that each residual series has a unit root.

```
> regTGENonINSM=lm(log(TGEN)~log(INSM))
> ur.reg1=ur.df(regTGENonINSM$resid,lags=0,type="drift")
> attr(ur.reg1,"teststat")
               tau2     phi1
statistic -5.842696 17.08189
> attr(ur.reg1,"cval")
      1pct  5pct 10pct
tau2 -3.44 -2.87 -2.57
phi1  6.47  4.61  3.79
> regINSMonTGEN=lm(log(INSM)~log(TGEN))
> ur.reg2=ur.df(regINSMonTGEN$resid,lags=0,type="drift")
> attr(ur.reg2,"teststat")
               tau2     phi1
statistic -5.850039 17.13296
> attr(ur.reg2,"cval")
      1pct  5pct 10pct
tau2 -3.44 -2.87 -2.57
phi1  6.47  4.61  3.79
```

Notice how the unit root seems to go away after regressing each series on the other. This is somewhat deceptive because the P-values calculated above assume that the cointegrating vectors are known, but we used regressions to estimate them. Nevertheless, Figures **??** and **??** show the ACF and PACF for the residuals of the regression of TGEN on INSM. Those of the reverse regression look equally like plots for stationary series.

Now that it appears that we have a cointegration relationship, we might try to estimate it. The two sets of regression coefficients are

```
> regTGENonINSM$coef
 (Intercept) log(INSM)
    1.939416  1.100124
> regINSMonTGEN$coef
 (Intercept) log(TGEN)
   -1.677275 0.8740295
```

It is not clear which of the two suggested relationships we should use

$$\log(\text{TGEN}) - 1.1\log(\text{INSM}) \quad \sim \quad \text{stationary, or}$$

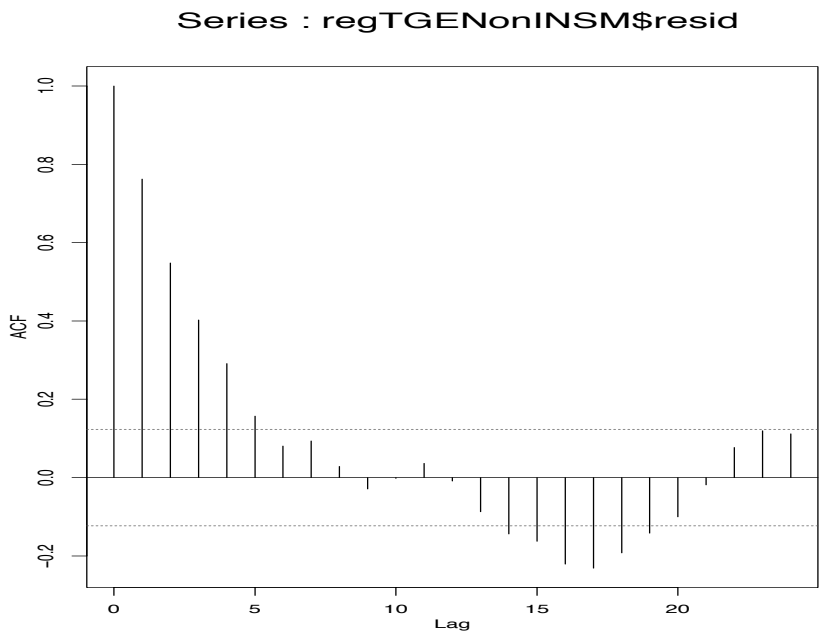

Figure 5.3: ACF plot of residuals of regression of TGEN on INSM, 2003

$$\log(\text{INSM}) - 0.874\log(\text{TGEN}) \quad \sim \quad \text{stationary.}$$

Fortunately, they aren't much different, and there are alternatives for choosing the relationship.

The Johansen test is not so intuitive to describe, but it provides an easier choice of cointegrating vector when it finds a cointegration relationship. First, one fits VAR models to see what AR order to use.

```
> library(vars)
> var2=VAR(log(cbind(TGEN,INSM)),2)
> summary(var2)

VAR Estimation Results:
=========================
Endogenous variables: TGEN, INSM
Deterministic variables: const
Sample size: 250
Log Likelihood: 597.115
Roots of the characteristic polynomial:
0.9942 0.7038 0.194 0.01627
Call:
VAR(y = log(cbind(TGEN, INSM)), p = 2)
```

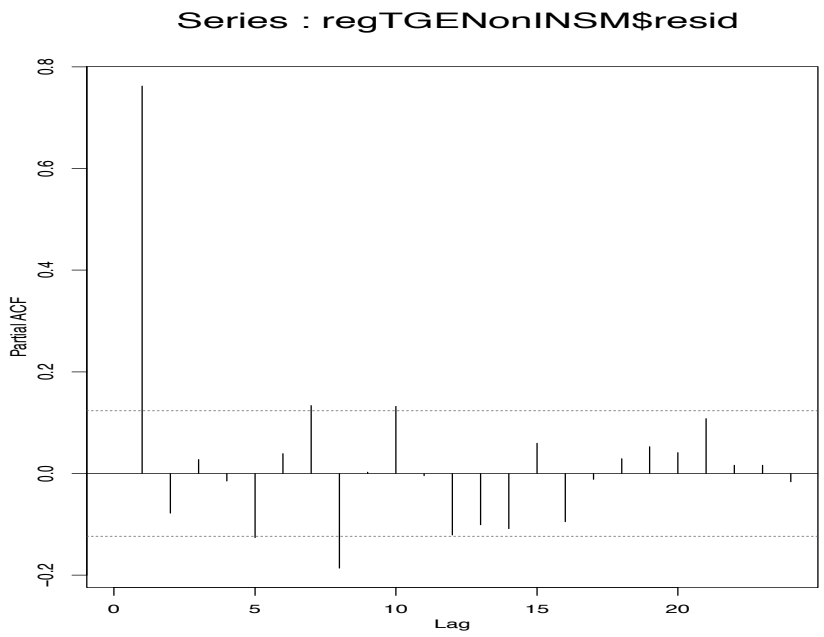

Figure 5.4: PACF plot of residuals of regression of TGEN on INSM, 2003

```
Estimation results for equation TGEN:
=====================================
TGEN = TGEN.l1 + INSM.l1 + TGEN.l2 + INSM.l2 + const

        Estimate Std. Error t value Pr(>|t|)
TGEN.l1  0.98447    0.06329  15.556  < 2e-16 ***
INSM.l1  0.22803    0.09353   2.438   0.0155 *
TGEN.l2 -0.15333    0.06361  -2.410   0.0167 *
INSM.l2 -0.04434    0.09233  -0.480   0.6315
const    0.33385    0.07003   4.767 3.20e-06 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 0.08961 on 245 degrees of freedom
Multiple R-Squared: 0.9891,     Adjusted R-squared: 0.989
F-statistic:  5582 on 4 and 245 DF,  p-value: < 2.2e-16


Estimation results for equation INSM:
=====================================
INSM = TGEN.l1 + INSM.l1 + TGEN.l2 + INSM.l2 + const
```

```
        Estimate Std. Error t value Pr(>|t|)
TGEN.l1  0.10090    0.04333   2.329  0.02068 *
INSM.l1  0.92376    0.06403  14.426  < 2e-16 ***
TGEN.l2 -0.01973    0.04355  -0.453  0.65101
INSM.l2 -0.02010    0.06321  -0.318  0.75072
const   -0.14717    0.04794  -3.070  0.00238 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 0.06135 on 245 degrees of freedom
Multiple R-Squared: 0.9936,     Adjusted R-squared: 0.9934
F-statistic:  9441 on 4 and 245 DF,  p-value: < 2.2e-16



Covariance matrix of residuals:
          TGEN      INSM
TGEN 0.0080305 0.0004055
INSM 0.0004055 0.0037639

Correlation matrix of residuals:
        TGEN    INSM
TGEN 1.00000 0.07375
INSM 0.07375 1.00000
```

The t statistics for the AR(2) coefficients are not so large as those for the AR(1) coefficients, so we will assume that the AR(1) model is sufficient. Next, one performs the cointegration test specifying the AR order. For some reason, `K=2` specifies an AR(1) model.

```
> coint2=ca.jo(log(cbind(TGEN,INSM)),type="eigen",ecdet="none",K=2,spec="longrun")
> summary(coint2)

######################
# Johansen-Procedure #
######################

Test type: maximal eigenvalue statistic (lambda max) , with linear trend

Eigenvalues (lambda):
[1] 0.12698200 0.00623771
```

```
Values of teststatistic and critical values of test:

          test 10pct  5pct  1pct
r <= 1 |  1.56  6.50  8.18 11.65
r = 0  | 33.95 12.91 14.90 19.19

Eigenvectors, normalised to first column:
(These are the cointegration relations)

          TGEN.l2  INSM.l2
TGEN.l2  1.000000 1.000000
INSM.l2 -1.122423 1.166665

Weights W:
(This is the loading matrix)

           TGEN.l2      INSM.l2
TGEN.d -0.16630436 -0.002554965
INSM.d  0.08346252 -0.002285210
```

The key portions of the summary of the `teststatistic` part and the `Eigenvectors`. The test above can be interpreted as follows. The hypothesis `r = 0` is that there are no cointegrating vectors, and this hypothesis is rejected. The hypothesis `r <=1` is that there is one cointegrating vector, and this is not rejected. So, we will proceed as if there were one cointegration relation, and it is estimated as

$$\log(\text{TGEN}) - 1.1224 \log(\text{INSM}) \sim \text{stationary}. \tag{5.4}$$

We can switch the coefficient of 1 to the other variable by dividing by the other coefficient. This gives $-0.8909 \log(\text{TGEN}) + \log(\text{INSM})$ as stationary. Both of these are nearly the same as the relationships that we found from the Engle-Granger method.

In order to use this estimated relationship for pairs trading, we could monitor $\log(\text{TGEN}) - 1.1224 \log(\text{INSM})$ during the trading period. It would make sense to standardize by subtracting its historical mean M and dividing by its historical standard deviation D, and open a position when

$$\frac{\log(\text{TGEN}) - 1.1224 \log(\text{INSM}) - \text{M}}{D} \tag{5.5}$$

leaves some interval around 0, such as $[-2, 2]$ or $[-3, 3]$. If, for example, (**??**) were to become greater than 2 on day t_0, signaling us to open a position, we could go short c/TGEN_{t_0} shares of TGEN and long $c\ 1.1224/\text{INSM}_{t_0}$ shares of INSM and wait until (**??**) crossed 0 again.

Alternatively, we could use linear regression to form a linear approximation by regressing (**??**) on TGEN and ISNM.

```
> dualr=lm((log(TGEN)-1.12241*log(INSM))~TGEN+INSM)
```

```
> dualr$coef
(Intercept)        TGEN        INSM
 -0.3025022   0.4732447  -0.4014992
```

This calculation suggests that

$$\log(\text{TGEN}) - 1.1224 \log(\text{INSM}) \approx 0.4732447\text{TGEN} - 0.4014992\text{INSM}.$$

Then, if (**??**) gets large, we would open a position short $0.4732447c/\text{TGEN}_{t_0}$ shares of TGEN and long $0.4014992c/\text{INSM}_{t_0}$ shares of INSM. This is noticably different from the other approximation.

Figure **??** contains a plot to help assess the fit of the two approximations. It is pretty clear that

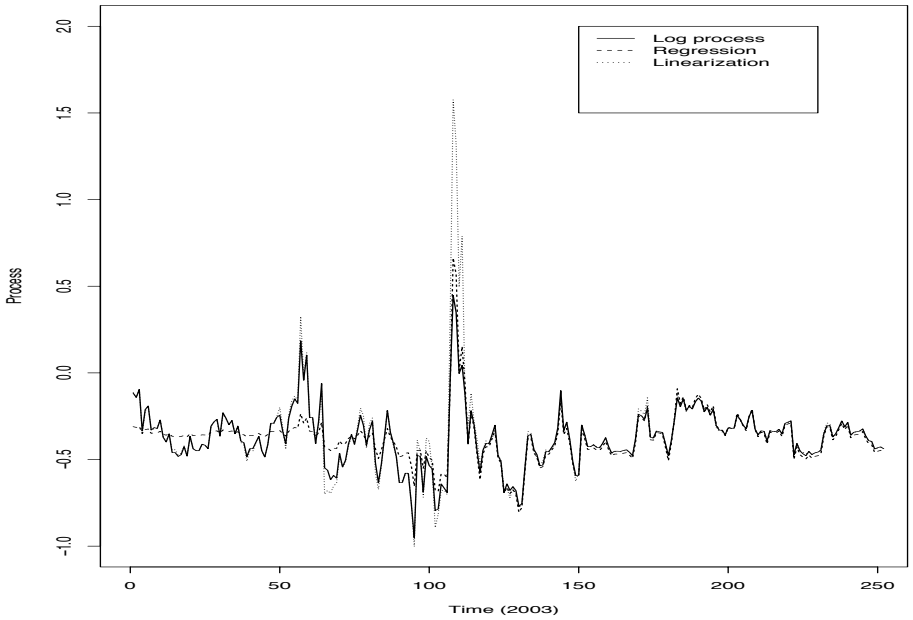

Figure 5.5: Time series plots of (**??**), called "Log process", linear regression fitted values, and (**??**) five days after a position opens, called "Linearization" for 2003

both linear approximations are pretty close most of the time, but linearization (**??**) can diverge after a few days if there is significant movement.

Choosing which pairs to trade is more challenging based on cointegration because it is more difficult and time consuming to measure the degree of cointegration. A simple surrogate for the degree of cointegration might be the size of one of the Johansen test statistics.

In the following example, we used the trace statistic for the test of no cointegrating vectors vs. one cointegrating vector, the test of H(0), as our measure of how cointegrated a pair was. We ranked all pairs by this statistic and chose the 50 top pairs to trade. (We could have chosen some other

number of pairs, and we could have chosen a different measure of cointegration.) The set of stocks used was that part of the industry group 283 with complete data for 2003. This gave us 291 stocks. Trading was done for 2004. Table **??** gives the summary of the trades over the year 2004. The profit numbers have been adjusted for the fact that the trades were not all net zero investment as they were in Chapter 4. This is due to the fact that the coefficients in the cointegrating vectors are not always close to 1 and -1. Each trade would either require a small investment with a cost that depends on the current interest rate or would result in a small amount of extra cash that could produce an income. We used an additional restriction in Table **??** which was to trade no pair unless the two coefficients were close together. In this case, we required that the absolute value of the logarithm of the ratio of the (absolute values of the) two coefficients had to be less than 0.223. This corresponds to the smaller coefficient being no more than 20% less the larger coefficient.

5.8 An Example with Three Stocks

Cointegration can occur between more than two series at a time. Three stocks that seem to move together are AVNR, ALO, and EBIO. We will check their cointegration status next based on 2003 data.

First, we fit a vector autoregression model `VAR` to see how many lags the processes need.

```
> AVNR=PRICE[253:756,which(names(PRICE)=="AVNR")]
> ALO=PRICE[253:756,which(names(PRICE)=="ALO")]
> EBIO=PRICE[253:756,which(names(PRICE)=="EBIO")]
> mat3var2=VAR(log(cbind(AVNR,ALO,EBIO))[1:252,],2)
> summary(mat3var2)
VAR Estimation Results:
========================
Endogenous variables: AVNR, ALO, EBIO
Deterministic variables: const
Sample size: 250
Log Likelihood: 1254.222
Roots of the characteristic polynomial:
0.9733 0.8636 0.8119 0.07572 0.06069 0.01103
Call:
VAR(y = log(cbind(AVNR, ALO, EBIO))[1:252, ], p = 2)


Estimation results for equation AVNR:
=====================================
AVNR = AVNR.l1 + ALO.l1 + EBIO.l1 + AVNR.l2 + ALO.l2 + EBIO.l2 + const

        Estimate Std. Error t value Pr(>|t|)
```

Table 5.1: Positions and adjusted profits for all 50 traded pairs. Adjusted profit is the profit adjusted by applying the risk-free rate (1-month T Bills) to the investment.

	Cutoff 2				Cutoff 3			
Stocks	Open	Bail	Close	Adj. Profit	Open	Bail	Close	Adj. Profit
KV.2/KV.1	1	1	0	-0.0212	1	1	0	-0.0141
ADLR/EZEM	1	1	0	0.0155	1	1	0	0.0155
ALO/BIIB	1	1	0	-0.1839	1	1	0	-0.1215
INSM/TGEN	3	0	3	1.3199	1	0	1	0.7719
LCI/EBIO	2	1	1	-0.0376	2	1	1	0.1662
CRL/SIAL	1	1	0	-0.1578	1	1	0	-0.0740
ALO/SIAL	1	1	0	-0.2738	1	1	0	-0.2220
WCRX.1/SNUS	2	1	1	-0.0408	1	1	0	-0.1718
SLXA/STEM	3	1	2	0.5616	1	1	0	-0.2840
IVD/ORCH	2	0	1	0.6745	1	0	1	0.7125
WFHC/REGN	1	1	0	-0.1764	1	1	0	-0.1764
WCRX.1/QLTI	1	1	0	-0.2359	1	1	0	-0.1084
PDLI/ELN	1	1	0	-0.4467	1	1	0	-0.2625
EBIO/SCLN	1	1	0	-0.3042	1	1	0	-0.2156
EBIO/IMNR	2	1	1	-0.0475	1	1	0	-0.2186
EBIO/ENZ	3	0	2	0.4471	1	0	1	0.2968
DOVP/OSIP	1	1	0	0.0153	1	1	0	0.0153
ABT/ALO	2	1	1	-0.0713	1	1	0	0.0429
DRRX/VPHM	1	1	0	-0.3929	1	1	0	-0.1809
HGSI/EBIO	3	0	2	0.4596	1	0	0	-0.0108
BMY/ALO	3	0	2	0.7379	1	0	1	0.4648
ONTY/AVNR	1	1	0	-0.5623	1	1	0	-0.2579
EBIO/AVNR	2	1	1	0.0739	1	1	0	-0.1275
FRX/CHTT	2	1	1	0.0033	1	1	0	-0.1347
TPPH/IMCL	0	0	0	0.0000	0	0	0	0.0000
LGND/IDEV	2	1	1	0.2874	2	1	1	0.3609
AIMM/IDEV	1	1	0	-0.3454	1	1	0	-0.2159
CRIS/CYTR	1	0	0	-0.1209	0	0	0	0.0000
LCI/AVNR	2	1	1	-0.1056	1	1	0	-0.2024
VVUS/BTX	2	1	1	0.0603	2	1	1	0.1814
ONTY/CHTT	1	0	1	1.1852	1	0	1	1.7961
AGEN/NTEC	1	1	0	-0.2053	1	1	0	-0.1147
WCRX.1/PRCS	1	1	0	-0.2463	1	1	0	-0.1592
WFHC/AVII	1	0	0	-0.5774	1	0	0	-0.3220
IDBE/SEPR	1	1	0	-0.6002	1	1	0	-0.0444
DRRX/ELN	1	1	0	-0.8490	1	1	0	-0.5600
VXGN/LJPC	1	0	0	0.4381	1	0	0	0.8962
CTIC/ALSE	1	1	0	-0.3718	1	1	0	-0.2384
ISPH/OLAB	3	1	2	0.6584	2	1	1	0.3610
NEXM/GTCB	0	0	0	0.0000	0	0	0	0.0000
ACAM/EBIO	1	1	0	-0.1955	1	1	0	-0.1276
VPHM/ONTY	2	0	1	0.5697	1	0	0	0.3230
ISIS/LCBM	1	0	0	0.0314	1	0	0	0.3062
ENDP/EBIO	2	0	2	0.9463	1	0	1	0.5457
AVNR/ENZ	1	1	0	-0.1750	1	1	0	-0.1997
INSM/ENMD	1	0	1	0.3420	1	0	1	0.8345
AIMM/AVAN	1	0	0	-0.3227	1	0	0	-0.1591
ELI/BCII	2	1	1	-0.0316	1	1	0	-0.1789
VPHM/MCHM	1	1	0	-0.2739	1	1	0	-0.1241
IMNR/ENZ	1	1	0	-0.2385	1	1	0	-0.1629
Totals	73	34	29	1.2154	51	34	11	2.7001

```
AVNR.l1  0.84151    0.06443  13.061  < 2e-16 ***
ALO.l1   0.04125    0.13846   0.298  0.76601
EBIO.l1  0.18641    0.06423   2.902  0.00405 **
AVNR.l2  0.02569    0.06454   0.398  0.69094
ALO.l2   0.03802    0.13448   0.283  0.77766
EBIO.l2 -0.11828    0.06309  -1.875  0.06204 .
const   -0.04830    0.13154  -0.367  0.71382
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 0.05951 on 243 degrees of freedom
Multiple R-Squared: 0.9361,     Adjusted R-squared: 0.9345
F-statistic: 593.4 on 6 and 243 DF,  p-value: < 2.2e-16


Estimation results for equation ALO:
====================================
ALO = AVNR.l1 + ALO.l1 + EBIO.l1 + AVNR.l2 + ALO.l2 + EBIO.l2 + const

        Estimate Std. Error t value Pr(>|t|)
AVNR.l1  0.01579    0.03010   0.525    0.600
ALO.l1   0.87638    0.06470  13.546  < 2e-16 ***
EBIO.l1  0.01361    0.03001   0.454    0.651
AVNR.l2  0.01452    0.03016   0.481    0.631
ALO.l2   0.02051    0.06283   0.326    0.744
EBIO.l2 -0.01574    0.02948  -0.534    0.594
const    0.25340    0.06146   4.123 5.13e-05 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 0.0278 on 243 degrees of freedom
Multiple R-Squared: 0.9386,     Adjusted R-squared: 0.9371
F-statistic: 619.1 on 6 and 243 DF,  p-value: < 2.2e-16


Estimation results for equation EBIO:
====================================
EBIO = AVNR.l1 + ALO.l1 + EBIO.l1 + AVNR.l2 + ALO.l2 + EBIO.l2 + const

         Estimate Std. Error t value Pr(>|t|)
AVNR.l1  0.097947   0.064882   1.510  0.13244
ALO.l1   0.105213   0.139437   0.755  0.45124
EBIO.l1  0.904831   0.064680  13.989  < 2e-16 ***
```

```
AVNR.l2 -0.016978   0.064995  -0.261  0.79415
ALO.l2   0.003343   0.135425   0.025  0.98033
EBIO.l2 -0.033080   0.063536  -0.521  0.60309
const   -0.367694   0.132460  -2.776  0.00593 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1


Residual standard error: 0.05992 on 243 degrees of freedom
Multiple R-Squared: 0.9468,     Adjusted R-squared: 0.9455
F-statistic: 721.1 on 6 and 243 DF,  p-value: < 2.2e-16
```

The lag-two coefficients for all three series appear to be not very significant, so we will use a VAR(1) model. Then we run a Johansen test. This time, we will use the eigenvalue test. For some reason, `K=2` means to fit a VAR(1) model.

```
> mat3coint<-ca.jo(log(cbind(AVNR,ALO,EBIO))[1:252,],type="eigen",ecdet="none",K=2)
> summary(mat3coint)

######################
# Johansen-Procedure #
######################

Test type: maximal eigenvalue statistic (lambda max) , with linear trend

Eigenvalues (lambda):
[1] 0.09747590 0.09374891 0.02020870

Values of teststatistic and critical values of test:

          test 10pct  5pct  1pct
r <= 2 |  5.10  6.50  8.18 11.65
r <= 1 | 24.61 12.91 14.90 19.19
r = 0  | 25.64 18.90 21.07 25.75

Eigenvectors, normalised to first column:
(These are the cointegration relations)

           AVNR.l2     ALO.l2    EBIO.l2
AVNR.l2   1.000000  1.0000000  1.0000000
ALO.l2  -10.853807 -0.4439529 -0.5312446
EBIO.l2   2.665811 -0.7587888  0.8935871
```

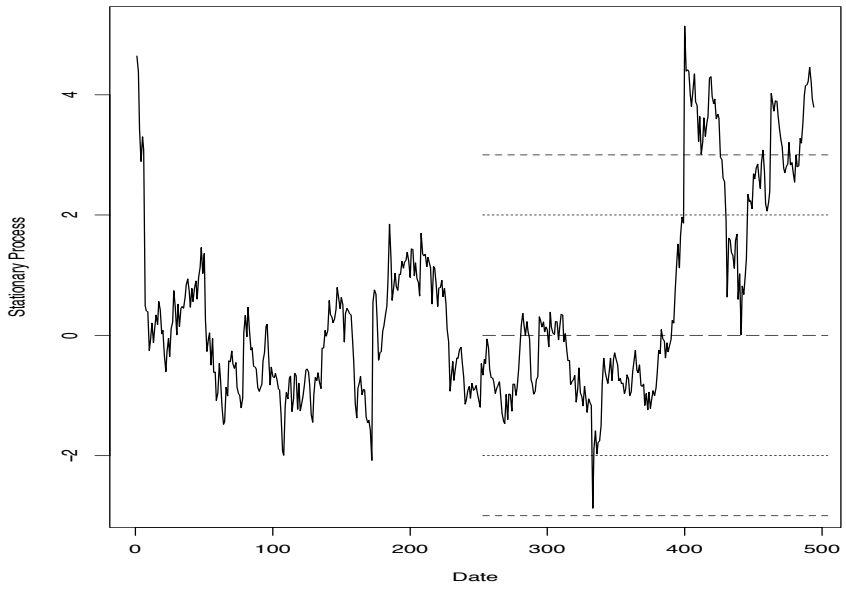

Figure 5.6: Three-stock trading based on cointegration with both formation (2003) and trading (2004) periods. The time series is the normalized version of process (**??**). The horizontal lines are trading boundaries during the trading period.

```
Weights W:
(This is the loading matrix)

            AVNR.l2      ALO.l2      EBIO.l2
AVNR.d -0.001817130 -0.11499602 -0.015986383
ALO.d   0.008657042  0.02695847 -0.005310815
EBIO.d -0.013781754  0.10662088 -0.011870153
```

It appears that there is at least one cointegrating vector for the three of these series together. One process that appears to be stationary is

$$\log(\text{AVNR}) - 10.8538\log(\text{ALO}) + 2.6658\log(\text{EBIO}). \tag{5.6}$$

We could monitor this process during 2004 and open positions when it gets far from its historical mean. When the value gets large, we short AVNR and EBIO and long ALO. When the value gets small, we do the opposite. Figure **??** shows a plot of (**??**) for both the formation and trading period with trading boundaries. It indicates that, with a cutoff of 2, two positions would be opened and one closed. (There was almost a third position, but the second one does not quite hit 0.) Figure **??** shows the three separate series of cointegration residuals for the three pairs of trades. There is nothing that stops one from trading both a triple and some of the pairs.

In the example above, there were two potential cointegrating vectors. We could have traded either of

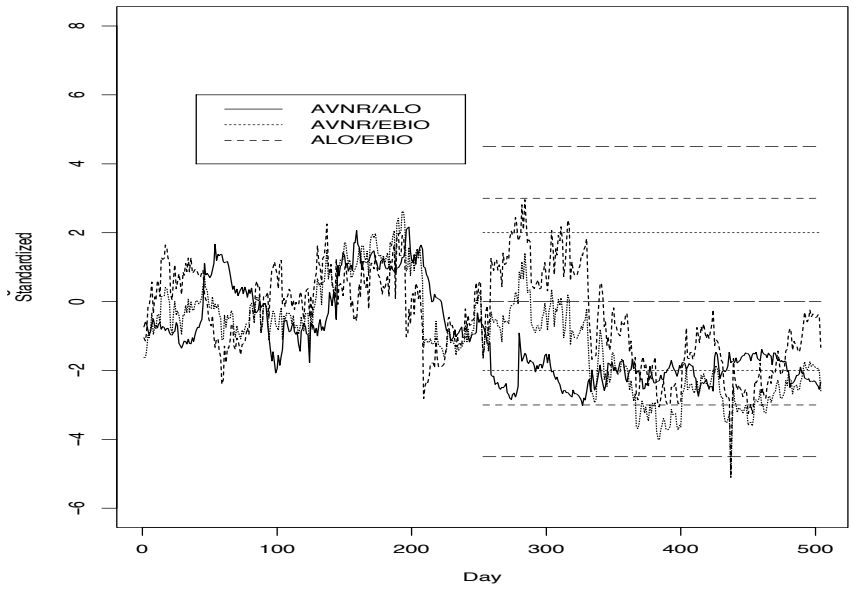

Figure 5.7: Three pairs of cointegration-based trades. The time series are the normalized cointegration residuals. The horizontal lines are trading boundaries during the trading period.

them or any linear combination of them. In particular, we could have chosen a linear combination that had the property that the sum of the coefficients is equal to 0. This would make the net investment at the start of each trade equal to 0, which simplifies the trade.

5.9 Choosing Multiple Stocks

There is a Machine Learning technique that is adaptable to the formation of portfolios consisting of multiple stocks exhibiting cointegration. With 291 stocks, there are over four million potential portfolios with three stocks. Overall there are more than 10^{57} potential portfolios. The *graphical lasso* is a technique that was designed for finding sparse networks with a large number of nodes. We can think of each stock in the universe of stocks that we want to trade as the nodes in a network. The edges of the network connect those nodes that we should consider trading. If we imagine that the prices or log-prices have a joint multivariate normal distribution with a 291×291 covariance matrix Σ, then $\Omega = \Sigma^{-1}$ expresses the conditional independence relationships. A zero entry in Ω means that the row and column variables are conditionally independent given the other variables. If a row of Ω has only a few nonzero entries, then the row variable is conditionally independent of most of the other variables given just a few that might then make up a portfolio.

The graphical lasso gives us a way to prefilter the set of all portfolios before we test the remaining

ones for cointegration. The graphical lasso was originally developed as a method to fit sparse graphical models to samples of a large number ℓ of continuous random variables. Here "sparse" means that the number of edges in the graph should be small compared to $\binom{\ell}{2}$, the maximal number. A "graphical model" is a way of specifying which random variables are conditionally independent of each other given which other random variables.

For our purposes, we would make use of the dependencies to suggest possible portfolios. A sparse graphical model would allow us to extract a manageable number of candidate portfolios that could then be tested for cointegration.

The graphical lasso is a constrained maximum likelihood method for estimating the inverse of a covariance matrix. The constraint forces a large number of the entries of the matrix to be 0. Each 0 corresponds to a pair of variables that are conditionally independent given the other variables. The nonzero entries can then be used to identify candidate portfolios.

The first step is to fit the graphical lasso. To that end, we briefly describe how that is done and why it is expected to do what is advertised above.

Let $X_1, \ldots, X_n$ be a random sample of random variables having common multivariate normal distribution with covariance matrix Σ. (In trading, each X_i is a vector of the stock universe at a single time i.) Let $\Omega = \Sigma^{-1}$, which is often called the *precision matrix*. The log-likelihood function is n times

$$\ell(\Omega) = -0.5 \left[\log |\Omega| + \operatorname{trace}(S\Omega) \right],$$

where S is the sample covariance matrix:

$$S = \frac{1}{n} \sum_{i=1}^{n} (X_i - \overline{X})(X_i - \overline{X})^\top.$$

Instead of maximizing the log-likelihood, we combine it with an L^1 penalty and minimize

$$-\ell(\Omega) + \rho \sum_{j \neq k} |\Omega_{j,k}|, \tag{5.7}$$

where $\rho \geq 0$ is a tuning parameter. We divide the log-likelihood by n so that ρ in the penalty (**??**) can be chosen without reference to n. Using $\rho = 0$ gives the MLE, while $\rho \to \infty$ forces all off-diagonal entries of $\hat{\Omega}$ to go to 0. We want something in between. It should be noted that minimizing (**??**) is equivalent to maximizing $\ell(\Omega)$ subject to a constraint of the form $\sum_{j \neq k} |\Omega_{j,k}| \leq c$, where c and ρ are related. In particular, letting c go to 0 corresponds to letting ρ go to ∞ and vice-versa. This latter, constrained, version of the problem makes it somewhat easier to see why the solution tends to be sparse. Figure **??** illustrates why the constrained solution tends to be sparse.

We don't need a graphical model for interpretation. All we need is a collection of portfolios to test for cointegration. Each row of the fitted matrix $\hat{\Omega}$ that has more than one nonzero entry corresponds to a potential portfolio with stocks corresponding to those nonzero entries. If a potential portfolio has too many stocks, it will probably incur too many transaction costs to be profitable. It will also

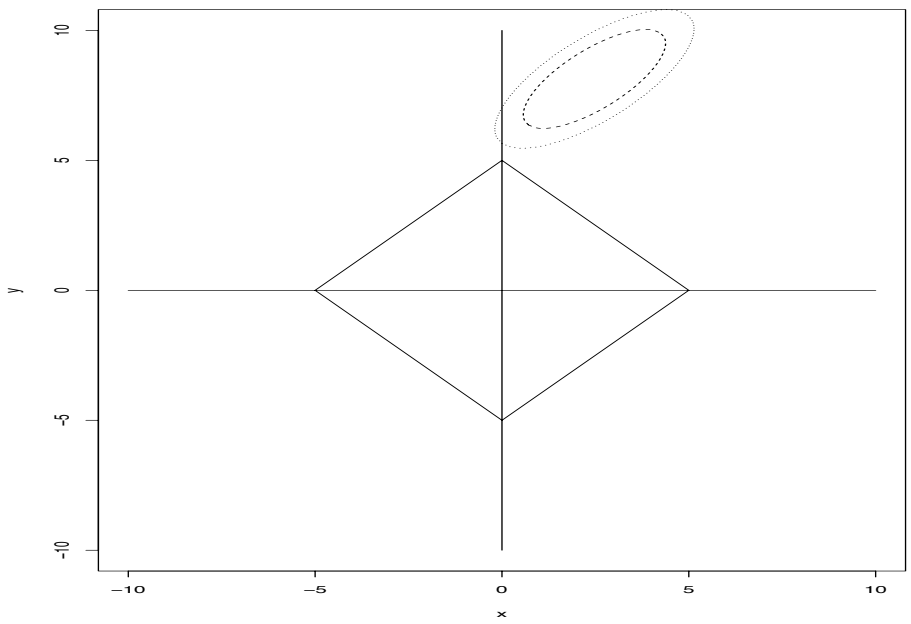

Figure 5.8: Illustration of L^1-constrained likelihood maximization. As the ellipse gets larger, the log-likelihood decreases. As the diamond gets larger, the L^1 constraint weakens. The two convex sets tend to meet at a corner of the diamond (or an edge in higher dimensions). Corners and edges are where some (or many) of the variables are 0. In high dimensions, there are lots of corners and edges, making it very common for the ellipse to meet the diamond at a corner or edge.

be subject to more random shocks than we would like. So, we might choose ρ (equivalently c) so that the average number of nonzero entries in each row is some nice value, like 2 to 6. It might make more sense to run the graphical lasso on the correlation matrix because $|\Omega_{j,k}|$ will be on the same scale for all j and k. In this case, there is a theorem that says that $\rho \geq 1$ will cause the estimate of Ω to be the identity matrix (all off-diagonal entries equal to 0), hence one never need use $\rho \geq 1$ for a correlation matrix.

The *R* package `glasso` (use `library(glasso)` after you download it from *CRAN*) has functions that will fit the graphical lasso. The function `glasso` has three important arguments. They are the sample covariance matrix, the value ρ, and `penalize`, which should be set to `F` (to avoid penalizing the diagonal elements of Ω). A more versatile function `glassopath` allows "simultaneous" fitting of the graphical lasso for several ρ values. That function has three similar arguments, but `rho` is replaced by `rholist`, which gives the list of ρ values. There is one additional argument `trace` which should be set to 0, unless you want a slew of "information" printed. Both functions output a list with several items, the most important of which is `wi`, estimated inverse matrix (or matrices).

As an example, consider the technology stocks with SIC code 367x. We look at one year of data from September 2006 to August 2007. There are 240 stocks with SIC code 367x, but only 224 of them have complete data for the year in question.

```
library(glasso)
nomiss=!apply(sic367$price[1:251,],2,function(y){any(is.na(y))})
sum(nomiss)
[1] 24
```

The 251 dates in that year are not enough data to get a good estimate of the entire covariance matrix for these 240 stocks, let alone the inverse of that matrix. But with L^1 penalty, we can get a less variable estimate.

```
glexamp=glassopath(cor(sic367$price[1:251,nomiss]),pen=F,trace=0,
rholist=c(1:9)/10)
```

The array `glexamp$wi` is 224 by 224 by 9, the last index corresponding to the 9 values of ρ.

The estimated matrices are too big to print on the page, but we can summarize the effect of increasing ρ. With $\rho = 0$, we get the inverse of the sample covariance matrix with all 224 rows having no zero entries.

```
max(apply(glexamp$wi[,,1],1,function(x){sum(x!=0)}))
[1] 42
max(apply(glexamp$wi[,,9],1,function(x){sum(x!=0)}))
[1] 20
mean(apply(glexamp$wi[,,9],1,function(x){sum(x!=0)}))
[1] 2.848214
```

With $\rho = 1$, no row has more than 42 nonzero entries, and the average row has 28 nonzero entries rather than 224. As ρ increases toward 1, the average number of nonzero entries per row decreases to 1, with the average being 11.2 at $\rho = 0.8$ and 2.8 at $\rho = 0.9$. For each stock, we could find a value of ρ that gives a nice number of nonzero entries in the corresponding row. Then we could pass the variables with nonzero entries to a cointegration test to look for linear combinations that appear stationary. In our example, the first stock has 6 nonzero entries with $\rho = 0.7$.

```
sum(glexamp$wi[1,,7]!=0)
[1] 6
stock1=which(nomiss)[which(glexamp$wi[1,,7]!=0)]
stock1
[1]   1  80 142 159 166 190
library(urca)
glexamp$jo=ca.jo(as.data.frame(sic367$price[1:251,stock1]),type="eigen",ecdet="none",
K=2,spec="longrun")
```

```
summary(glexamp$jo)

######################
# Johansen-Procedure #
######################

Test type: maximal eigenvalue statistic (lambda max) , with linear trend

Eigenvalues (lambda):
[1] 0.1715728393 0.1368881290 0.0712006569 0.0505873625 0.0267103760
[6] 0.0003542622

Values of teststatistic and critical values of test:

          test 10pct  5pct  1pct
r <= 5 |  0.09  6.50  8.18 11.65
r <= 4 |  6.74 12.91 14.90 19.19
r <= 3 | 12.93 18.90 21.07 25.75
r <= 2 | 18.39 24.78 27.14 32.14
r <= 1 | 36.66 30.84 33.32 38.78
r = 0  | 46.87 36.25 39.43 44.59

Eigenvectors, normalised to first column:
(These are the cointegration relations)

            V1.l2      V2.l2      V3.l2     V4.l2     V5.l2      V6.l2
V1.l2  1.00000000  1.0000000  1.0000000  1.000000  1.000000  1.0000000
V2.l2 -2.57918321 10.4979326  2.9966105 -3.817049 13.393064 -1.4953857
V3.l2  0.15949700 -4.1954074 -1.1984106 -5.217411 -1.144969 -0.7810016
V4.l2 -0.78261291 -1.8134309  1.6910033  1.521727  1.171225  0.6609377
V5.l2 -0.08284749  0.5066976 -0.8946559 -2.146302 -6.342189  3.5901805
V6.l2 -0.15996440 -0.2698508 -0.6932976  1.106452  1.700489  0.0235387
```

The Johansen procedure suggests that there might be one or two cointegrating vectors. The first one is

$$X_1 - 2.58X_{80} + 0.159X_{142} - 0.783X_{159} - 0.0882X_{166} - 0.160X_{190},$$

which we compute and plot in the code that follows. The plot appears in Figure **??** over the period in question plus the next year.

```
stock1.port=sic367$price[,stock1]%*%matrix(attributes(glexamp$jo)$V[,1],6,1)
plot(stock1.port[1:502],type="l",xlab="Date",ylab="Portfolio")
lines(c(0,502),rep(mean(stock1.port[1:251]),2),lty=2)
lines(c(0,502),rep(mean(stock1.port[1:251])
```

```
+2*sqrt(var(stock1.port[1:251])),2),lty=3)
lines(c(0,502),rep(mean(stock1.port[1:251])
-2*sqrt(var(stock1.port[1:251])),2),lty=3)
title("Six-stock Portfolio: Sep. 2006-Aug. 2008")
lines(c(251,251),c(-30,-10),lty=4)
```

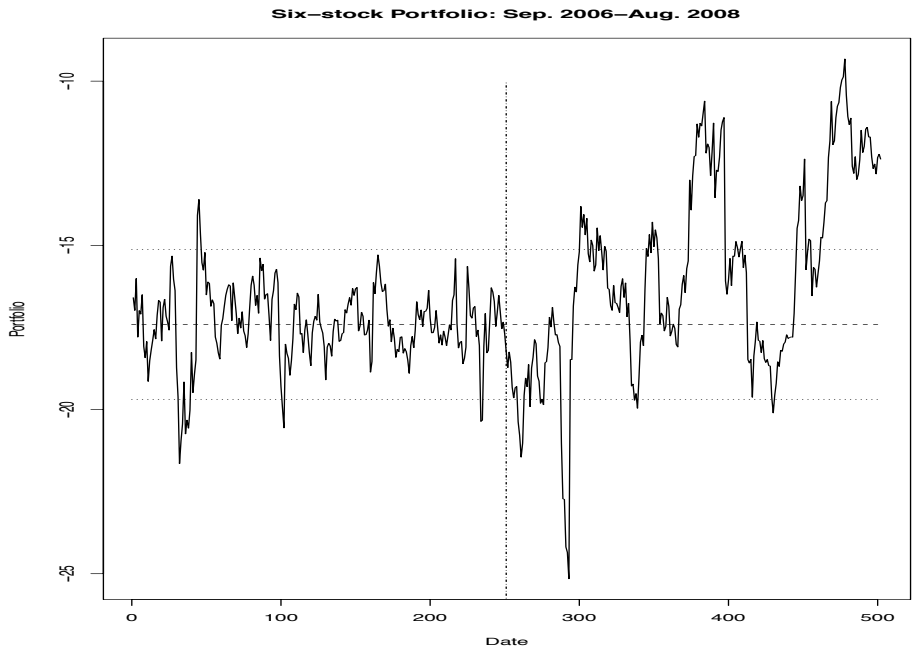

Figure 5.9: Six-stock portfolio chosen by graphical lasso. The horizontal lines indicate the mean and plus/minus 2 standard deviations of the process (computed from the first year). The vertical line indicates the end of the first year.

The variance increases dramatically in the second year. If we anticipate either a change in variance or a drift after we form the portfolio, hedging with options would be a natural insurance strategy. Unfortunately, only four of the six stocks in this portfolio have options available. The four stocks with options still have some degree of cointegration. Figure **??** shows how the four-stock portfolio goes bad in the second year.

5.10 Additional Points

In the TGEN/INSM example, the coefficients of the estimated conintegrating vector were close to 1 and -1. An alternative to using the Johansen or Engle/Granger methodologies to find a cointegrating relationship is the following. Just look for pairs where the log-ratio of prices appears to be approximately stationary. That is, perform a unit root test on a series such as log(TGEN/INSM), and choose pairs that have no evidence of unit root.

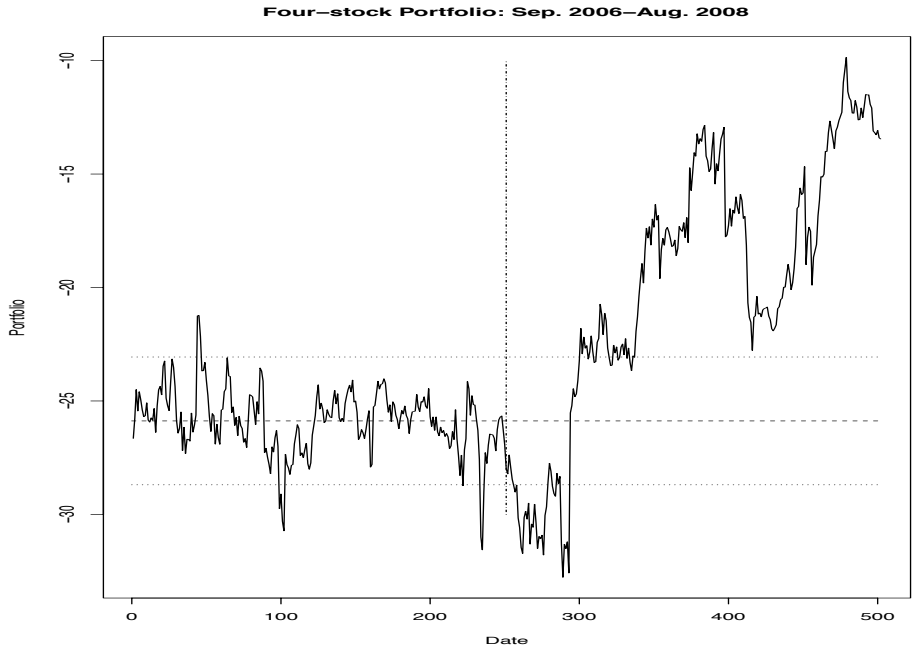

Figure 5.10: Four-stock portfolio consisting of those stocks from the six-stock portfolio that had available options.

Having no evidence of unit root suggests that the series is stationary, but not every stationary series is suitable for trading. For example, a white-noise process is stationary but useless for trading. An AR(1) process with coefficient in $(-1, 1)$ is stationary, but if the coefficient is too close to 0, the opportunities for trading would be limited. The AR(1) coefficient needs to be large enough so that, when we recognize a divergence we still have time to open a position before it goes away. But the coefficient shouldn't be so large that the divergence is essentially permanent on the time scale on which we want to trade (as with a random walk).

We also want to stress that cointegration is not the same as correlation. Figure **??** shows a plot of sample correlation against the Johansen cointegration test statistic for the companies in the drug industry data set. You can see that there is not a very strong empirical relationship between the two quantities. Although the pairs with the very largest cointegration test statistics have pretty high correlation, there are several pairs with pretty high test statistics and rather low correlation. And there are loads of pairs with very high correlation and low test statistics.

We showed why cointegrated vectors may not be highly correlated in the discussion surrounding (**??**). For the other direction, consider two series X and Y with $Y_t = X_t + \epsilon_t$ with X and ϵ being random walks that are independent of each other. Assume that the innovations of the ϵ process have variance (σ_2^2) that is very small compared to the variance (σ_1^2) of the innovations of the X process. Then Y and X are both $I(1)$ processes and so is every linear combination of Y and X.

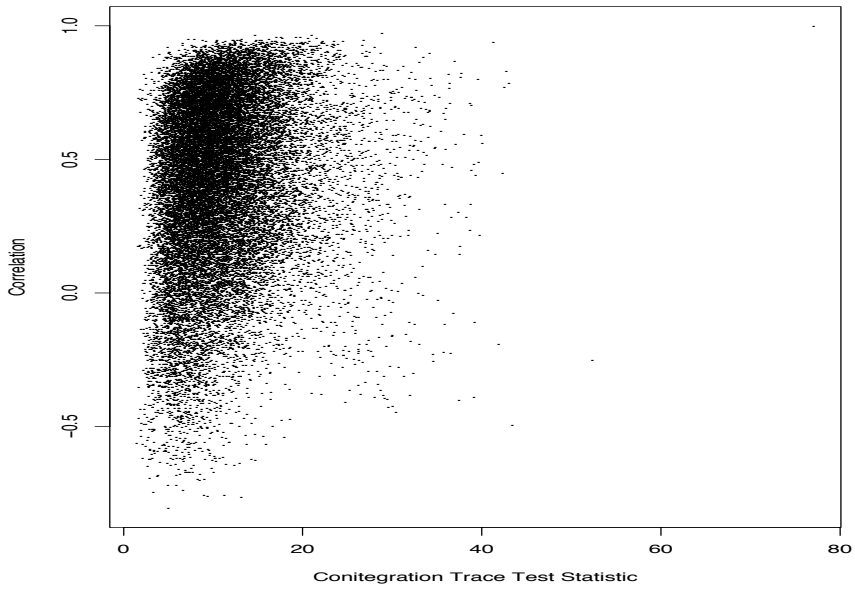

Figure 5.11: Correlation versus cointegration test statistic for drug industry data.

Hence, they are not cointegrated. The correlation, however, is

$$Cor(X_t, Y_t) = \frac{t\sigma_1^2}{t\sigma_1^2 + t\sigma_2^2} = \frac{1}{1 + \frac{\sigma_2^2}{\sigma_1^2}}.$$

This will be very close to 1 when σ_2^2 is much less than σ_1^2. Hence, highly correlated series need not be cointegrated.

Several of the pairs that were traded in the example in Chapter 4 are highly correlated, but the differences between the log-prices for the pairs still look like an $I(1)$ process. This explains why so many of the pairs opened one position that never closed during the trading period.

5.11 M. Avellaneda and J.-H. Lee (2010)

Reading Notes
"Statistical arbitrage in the US equities market"[4]

[4]In *Quantitative Finance*, **10(7)**, 2010, 761-782.

5.11.1 Overview

This paper discusses two approaches to pairs trading, except rather than trading one security against another, it explores trading one security against a group of securities. The group of securities is constructed either by using principal components analysis (PCA) or using using sector exchange traded funds (ETFs). The underlying theory is based on equation (3) on page 762 of the paper:

$$\frac{dP(t)}{P_t} = \alpha dt + \sum_{j=1}^{m} \beta_j F_j(t) + dX(t),$$

where α is an excess above market rate of return and is often set to 0, $F_j(t)$ are returns of the systematic risk-factors associated with the market under consideration ($F_j(t) = dU_j(t)/U_j(t)$ where $U_j(t)$ is the price of the jth risk factor), m in number, while $X(t)$ is the idiosyncratic component. This idiosyncratic component is not only assumed to be stationary (so the price and risk factors are cointegrated), but it is specifically modeled using an Ornstein-Uhlenbeck (linear mean-reverting) process:

$$dX(t) = \kappa(\mu - X(t))dt + \sigma dW(t),$$

where μ is the long run level of the process, κ is the speed of mean reversion, and σ is the volatility parameter. These parameters, $\alpha, \kappa, \mu, \sigma$ are specific to each stock under consideration for trading. The authors assert that these parameters may vary slowly, so they estimate them from data using a trailing 60 trading day time window. The Ornstein-Uhlenbeck process is well-understood and easy to work with. The conditional increments (i.e. the distribution of $X(t)$ given $X(s)$ for $s \leq t$) have a normal distribution with known parameters. The paper's appendix presents a method to estimate the regression parameters, β_i, for each stock as well as the parameters of the Ornstein Uhlenbeck process using equally spaced data for each of the stocks under consideration. The parameters are estimated daily using a 60-day trailing window (the authors chose the 60 trading day window, because, they say, it represents one earnings cycle).

The authors describe market-neutral portfolios in the context of their models. Letting $\{R_i,\ 1 \leq i \leq n\}$ represent the returns of the n stocks being considered for trading and consider the model

$$R_i = \sum_{j=1}^{m} \beta_{ij} F_j + \bar{R}_i,$$

where the $\{F_j,\ 1 \leq j \leq m\}$ are the systematic components of risk, while $\bar{R}_i$ is an uncorrelated idiosyncratic component.

Consider a portfolio in which $\{Q_i\ 1 \leq i \leq n\}$ represents the dollar amounts invested in each of the n stocks. Choose the Q_i to satisfy $\bar{\beta}_j = \sum_{i=1}^{n} \beta_{ij} Q_i = 0,\ 1 \leq j \leq m$. The $\bar{\beta}_j$ are the portfolio betas. A portfolio with 0 portfolio betas is called **market-neutral** in that it is uncorrelated with the market portfolio or the factors that drive the market. Simple substitution and some algebra shows that the portfolio returns, $\sum_{i=1}^{n} Q_i R_i = \sum_{i=1}^{n} Q_i \bar{R}_i$. This means that a market neutral portfolio depends only on the idiosyncratic returns.

A major issue in the paper is the extraction of the risk factors, $\{F_j,\ 1 \leq j \leq m\}$. The authors use two methods. The first is to take data on all the relevant stocks and conduct a principal components analysis (PCA). The 15 largest eigenvalues are used as the risk factors, and trades are set up accordingly.

An alternative approach taken by the authors is to use a collection of industry ETFs to represent collectively the market risk factors. Specifically, they choose 15 different ETFs, which they consider to be uncorrelated. For each ETF, they select a set of stocks each having market cap above \$1B, and each of which is represented by one of the industry ETFs. They consider pair-trading one stock against the corresponding ETF. At each time going forward, they estimate the parameters for the particular model and trade according to the strategy defined later. Their hope is that each industry ETF will appear both positively and negatively in the full set of pair trades, with an overall small net position in each of the ETFs. We reproduce the ETFs selected, their associated industries, and the number of stocks considered for pairs trading against each of the chosen ETFs. Insufficient information concerning which stocks are used is given for replicating of their study. Table 3 in the paper gives the ETF information some of which is summarized below:

Abbreviated Table 3: ETFs used in study

ETF	Sector	# of stocks
HHH	Internet	22
IYR	Real Estate	87
IYT	Transportation	46
OIH	Oil Exploration	42
RKH	Regional Banks	69
RTH	Retail	60
SMH	Semiconductors	55
UTH	Utility	75
XLE	Energy	75
XLF	Financial	210
XU	Industrial	141
XLK	Technology	158
XLP	Consumer Staples	61
XLV	Healthcare	109
XLY	Consumer Discretionary	207
Total		1417

The ETFs are not necessarily uncorrelated, and the authors take a simplified approach, namely each stock under consideration is associated with a particular single sector ETF. Then they set up a pairs trade involving the stock and its associated ETF. Once the set of stocks has been reduced by the \$1B requirement, the set of stocks is further reduced by requiring that the speed of mean reversion parameter, κ, is at least 1/2 of the estimation period. That is, the authors require the estimated κ to satisfy $\kappa > 252/30 = 8.4$.

For the idiosyncratic error process for stock i versus its associated ETF, $X_i(t)$ with speed of mean reversion parameter, κ_i, volatility parameter σ_i, and equilibrium mean μ_i, the authors define the equilibrium volatility by $\sigma_{i,eq} = \sigma_i\sqrt{\frac{\tau_i}{2}}$ where $\tau_i = \frac{1}{\kappa_i}$. They also define the **s-score** given by $s_i = \frac{X_i(t)-\mu_i}{\sigma_{i,eq}}$ associated with trading stock i versus the associated ETF. The behavior of the s-score function for the residuals from trading JPM against the corresponding financial ETF, XLF, is given in Figure 7, page 770.

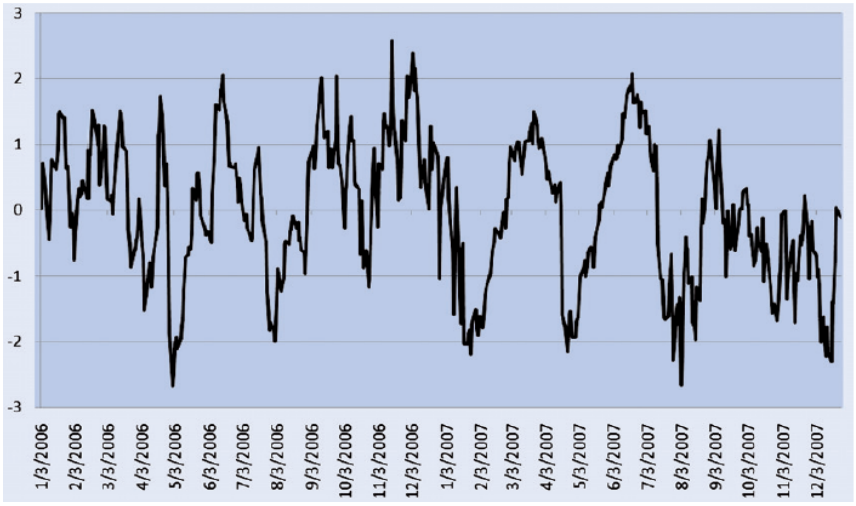

Figure 5.12: Figure 7, p770, s-score for JPM versus XLF

The pairs trading strategy has the usual form, namely:

- buy to open if $s_i < -\bar{s}_{bo}$,
- sell to open if $s_i > \bar{s}_{so}$,
- close short position if $s_i < \bar{s}_{bc}$,
- close long position if $s_i > -\bar{s}_{sc}$,

where the parameter values are chosen empirically. Here "buy to open" means buying one dollar of the stock and shorting β_i dollars of ETF_i, $1 \leq i \leq m$, so the portfolio is market neutral. The authors report they chose the following parameter values: $\bar{s}_{bo} = \bar{s}_{so} = 1.25$, $\bar{s}_{bc} = .75$ and $\bar{s}_{sc} = .50$. They also assume transaction costs of .05% or 5 basis points per trade, 10 basis points per round trip.

The authors backtested this strategy both for 15 actual ETFs from 2002-2007 and in a separate study for 15 synthetic ETFs over 1996-2007. Of course they also backtested using the PCA approach; however, that is not discussed in these notes. Synthetic ETFs were created because there were relatively few ETF products available for trading before 2002. These synthetic ETFs were created using cap-weighted industry-sector indices.

In section 6 of the paper, the authors introduce a method for taking volume into account. This is done by altering the time scale by weighting it by the volume. Instead of considering stock price changes per unit time, they consider stock price changes per unit volume. Let $V(t)$ represent the cumulative volume of shares traded over $[0, t]$, $S(t)$ represents the price at time t, and P(V(t)) represents the prices when $V(t)$ shares have been traded. Thus $P(V(t)) = S(t)$. From this relationship we find that the price change per share traded over some period of interest is given by:

$$\frac{P(V(t+h)) - P(V(t))}{V(t+h) - V(t)} = \frac{S(t+h) - S(t)}{V(t+h) - V(t)}.$$

This leads the authors to consider a volume-weighted modification of the usual daily returns:

$$\bar{R}(t) = \frac{S(t+h) - S(t)}{S(t)} \left(\frac{< \delta V >}{V(t+h) - V(t)} \right),$$

where $< \delta V >$ is the average daily trading volume calculated over the trailing 10 trading days. Thus returns are rescaled by the local relative trading volume. The authors observe that if trading volumes are relatively constant, then the returns in clock time and in trading time will be essentially equivalent. They go on to observe that if trading volume is low, the $\frac{<\delta V>}{V(t+h)-V(t)}$ factor will be relatively large (and larger than 1). Consequently, the trading time returns will be larger than the clock time returns. If trading volume is large compared with the 10 day average window, then trading time returns will be smaller than clock time returns. Thus, the trading volume at a time at which a trading signal would be generated under the standard clock time method will have an impact on whether to actually implement that trade if the trading time model is used. If the stock increases on high volume, using trading time, this may not be considered to be an "open and sell the stock" signal. Similarly, if a stock falls on high volume, under the trading time model, this may not be considered an "open and buy the stock" signal. Thus trades are relatively more likely to be opened on low volume and relatively less likely to be opened on high volume.

Table 4: Sharpe ratios using synthetic ETFs

Year	Portfolio
1996	1.7
1997	3.6
1998	3.4
1999	0.8
2000	0.3
2001	2.9
2002	2.0
2003	0.1
2004	0.8
2005	(1.3)
2006	(0.5)
2007	(0.5)
Total	1.1

Tables 5 and 9: Sharpe ratios using actual ETFs

Year	Portfolio Std. Time	Portfolio Trading Time
2002	2.7	
2003	0.8	0.9
2004	1.6	3.1
2005	0.1	1.6
2006	0.7	1.5
2007	(0.2)	0.4
Total	0.9	1.51

The PNL results for ETF trading using the standard time method (Tables 4 and 5 and Figure 9) show very good results from 1996-2002 for synthetic ETFs and 2002-2004 using actual ETFs. Beyond those time windows, the results are either mediocre or poor. It is interesting to note that the trading time results for actual ETFs show very good results over 2003-2007, averaging 1.51 over this entire period. Figure 19 summarizes the differences between clock time results and trading time results.

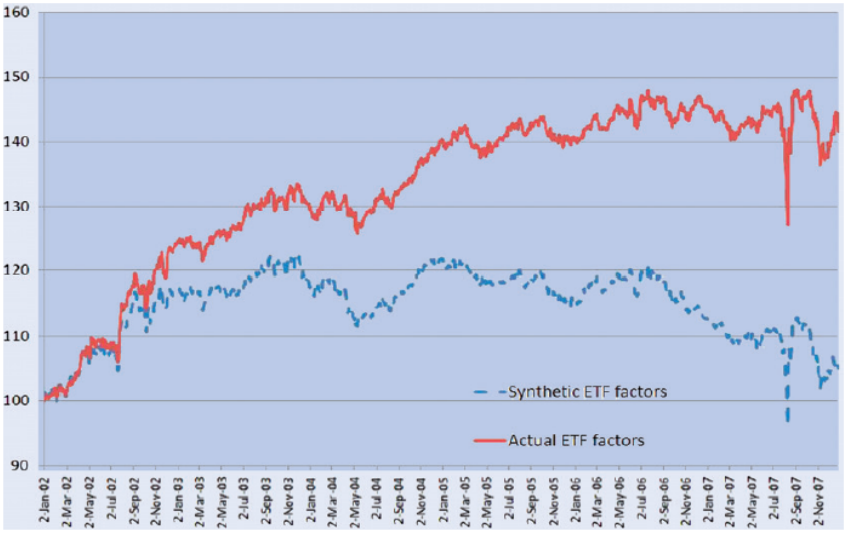

Figure 5.13: Figure 9, p772

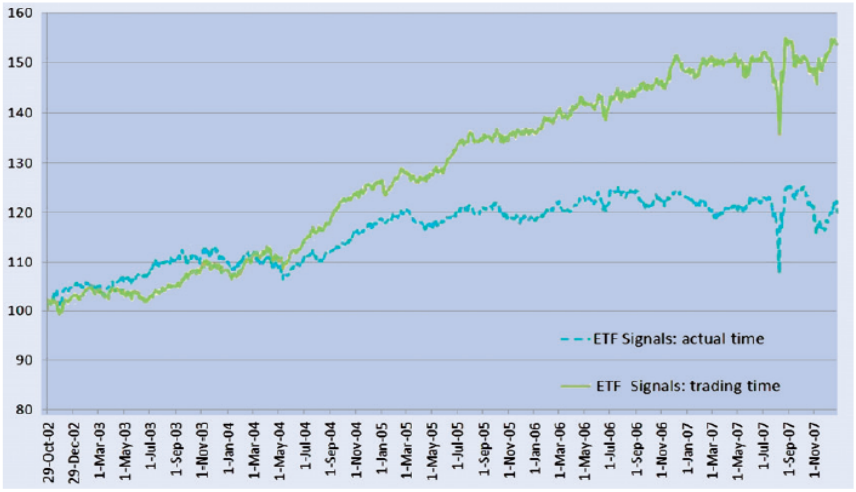

Figure 5.14: Figure 17, p776, PNL clock time vs trading time

5.11.2 Data Example

We can use the pharmaceutical data together with a number of pharmaceutical ETFs to illustrate some of Avellaneda's results. I have data for five ETFs: PJP, IXJ, PPH, XPH, and IHE. One is a HOLDR, two are from iShares, one is a SPDR, and one is from PowerShares. They move remarkably closely together, but they do offer trading opportunities with the stocks in SIC 283. There are three ways that we could make use of the ETF's in trading. One would be to throw them in with all of the other stocks and treat them like individual stocks in any of the algorithms that we have seen. A second would be to trade individual stocks against individual ETF's. A third would be to use the method described by Avellaneda and Lee.

Estimation of the Avellaneda and Lee model

Let $\{S_i(t) : t > 0\}$ be a price series, and let $\{U_j(t) : t > 0\}$ be indices, like ETF's for $j = 1, \ldots, N$. Then, the model relating S_i to the U_j's is

$$\frac{dS_i(t)}{S_i(t)} = \alpha_i dt + \sum_{j=1}^{N} \beta_{i,j} \frac{dU_j(t)}{U_j(t)} + dX_i(t),$$

where X_i is assumed to be a stationary Ornstein-Uhlenbeck process:

$$dX_i(t) = \kappa_i[\mu_i - X_i(t)]dt + \sigma_i dW_i(t).$$

Here, κ_i is the speed of mean-reversion, and μ_i is the mean of the stationary distribution. The variance of the stationary distribution is $\sigma_i^2/[2\kappa_i]$. The analysis that we are about to perform will be applied to an entire collection $S_1, \ldots, S_m$ of stocks in the same industry as the indices. Presumably, we observe the various series at discrete times only, so we see $(S_1(t_k), \ldots, S_m(t_k), U_1(t_k), \ldots, U_N(t_k))$, for $k = 1, \ldots, n$.

A discrete-time version of the above model can be expressed as

$$\log[S_i(t_{k+1})/S_i(t_k)] = \alpha_i + \sum_{j=1}^{N} \beta_{i,j} \log[U_j(t_{k+1})/U_j(t_k)] + \epsilon_{i,k+1},$$

where $\epsilon_{i,k+1} = X_i(t_{k+1}) - X_i(t_k)$. This makes $X_i(t_{k+1}) = \sum_{\ell=0}^{k+1} \epsilon_{i,\ell}$. One can recognize $R_{i,k} = \log[S_i(t_{k+1})/S_i(t_k)]$ as the log-return from time t_k to t_{k+1} for series S_i, and similarly for U_j. Let $R^U_{j,k}$ denote the log-return for index U_j Avellaneda and Lee fit the model in several stages. First, for each i perform a multiple regression of R_i on R^U_j, $j = 1, \ldots, N$. The data used in the regression are the previous 60 returns for the various series. There is no magic about 60, but it seemed like a long enough time to get decent estimates without introducing too much bias from the past. Save the coefficients $\hat{\beta}_i$ and residuals $\{e_{i,\ell}\}_{\ell=1}^{60}$ for later. Next, reconstruct

$$\hat{X}_i(t_{k+1}) = \sum_{\ell=0}^{k+1} e_{i,\ell}.$$

The discrete analog to the O-U process is an AR(1) process with drift. So, fit an AR(1) model $\hat{X}_i(t_{k+1}) = a_i + b_i\hat{X}_i(t_k) + \zeta_{i,k+1}$. Matching the parameters of the AR(1) model to those of the O-U process gives

$$\begin{aligned} a_i &= \mu_i[1 - \exp(-\kappa - i\Delta t)], \\ b_i &= \exp(-\kappa_i \Delta t), \\ Var(\zeta_{i,k}) &= \sigma_i^2 \frac{1 - \exp(-2\kappa_i \Delta t)}{2\kappa_i}. \end{aligned}$$

where Δt is the discrete time step, one day (1/252 years) if we use daily data. Solving, we get

$$\begin{aligned} \kappa_i &= -\log(b_i) \times 252, \\ \mu_i &= \frac{a_i}{1 - b_i}, \\ \sigma_i^2 &= \frac{Var(\zeta_{i,k}) \times 2\kappa_i}{1 - b_i^2}. \end{aligned}$$

This makes the equilibrium variance $\sigma_{i,eq}^2 = Var(\zeta_{i,k})/[1 - b_i^2]$.

In order to have fast mean reversion, we need large values of κ_i. The authors use $\kappa_i > 252/30$, so that the mean-reversion times should be about 1.5 months. This is the same as $0 < b_i < 0.9672$. Trading signals come from the deviation of the $\hat{X}_i$ process from its mean:

$$s_i = \frac{\hat{X}_i(t) - \mu_i}{\sigma_{i,eq}}.$$

At the last observation, $t = 60$, we get

$$s_i = \frac{-\mu_i}{\sigma_{i,eq}},$$

because $\hat{X}_i(60)$ has to be 0 due to the fact that it is the sum of the residuals from a regression. This can be "fixed" by using more than 60 days to fit the regression, but use only the last 60 residuals for the rest of the analysis. The authors prefer to replace μ_i by $\mu_i - \overline{\mu}$, where you average over different stocks, presumably only over stocks with $b_i < 0.9672$, otherwise the $\sigma_{i,eq}^2$ might be undefined. The authors also suggest that s_i be replaced by the modified score

$$s_{mod,i} = s_i - \frac{a_i}{\kappa_i \sigma_{i,eq}}.$$

Results

When s_i is big, the stock is temporarily overpriced relative to the index, so you should short the stock and long the index. If s_i is small, go the other way. Unwind when s_i gets back closer to 0. The position that the authors recommend opening has \$1 in the stock and $\hat{\beta}_{i,j}$ in the jth index. In the examples that I worked, I found a number of trades that involved very large amounts being invested in some of the indices. I decided to reject all trades if the number of shares of any stock

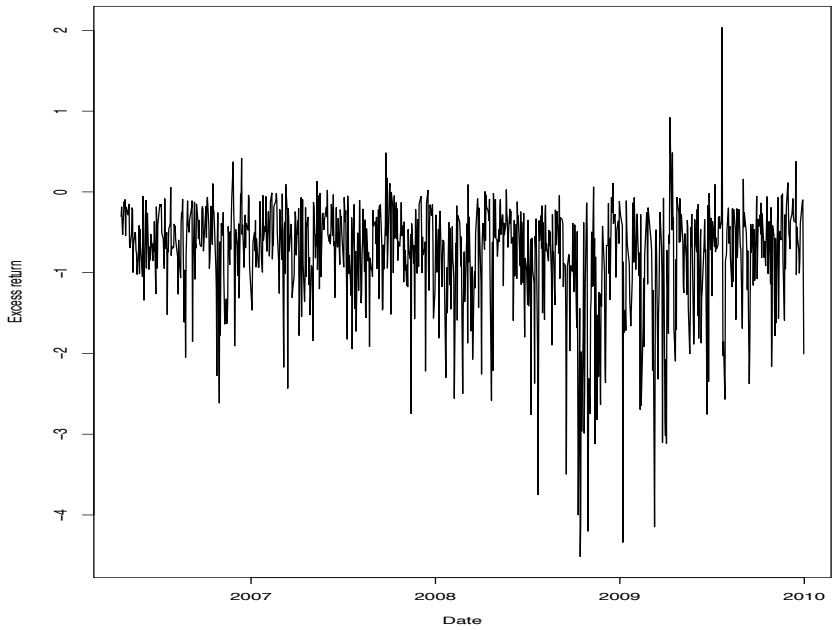

Figure 5.15: Time series of daily excess returns from Avellaneda and Lee strategy: 2006–2009

or index was more than 2. This tended to reject about 5% of the trades. Figure **??** shows the time series of excess returns from using the above strategy from 2006–2009. The data used were the same stocks from SIC 283 together with five ETF's. Their histogram is in Figure **??**. The daily excess returns were calculated for the strategy that uses 70 days to estimate the regression model but uses only the last 60 residuals to fit the AR(1) part. It also rejects all trades that want more than 2 shares of anything. The daily excess returns were computed as follows. For each trade that ends on a given day, start with the net cash flow: amount received upon closing minus amount used to open. Then, subtract 0.06/252 times the amount required to open the trade times the number of days the trade was open if that amount was positive. If the amount to open the trade was negaive (because the short had more than the long) then add 0.02/252 times the amount of excess cash times the number of days the trade was open. Then add together all results for each day separately. The Sharpe ratio for the illustrated strategy was only -0.7. Perhaps the strategy works better in another industry.

Since there were five ETF's we have the option of using various combinations of ETF's in the strategy. The strategy above used all five. At the other extreme, we could use only one ETF to trade against the industry. The Sharpe ratios for these five simpler strategies ranged from -0.42 to -0.46. If we run all five strategies simultaneously, we get a Sharpe ratio of -0.445.

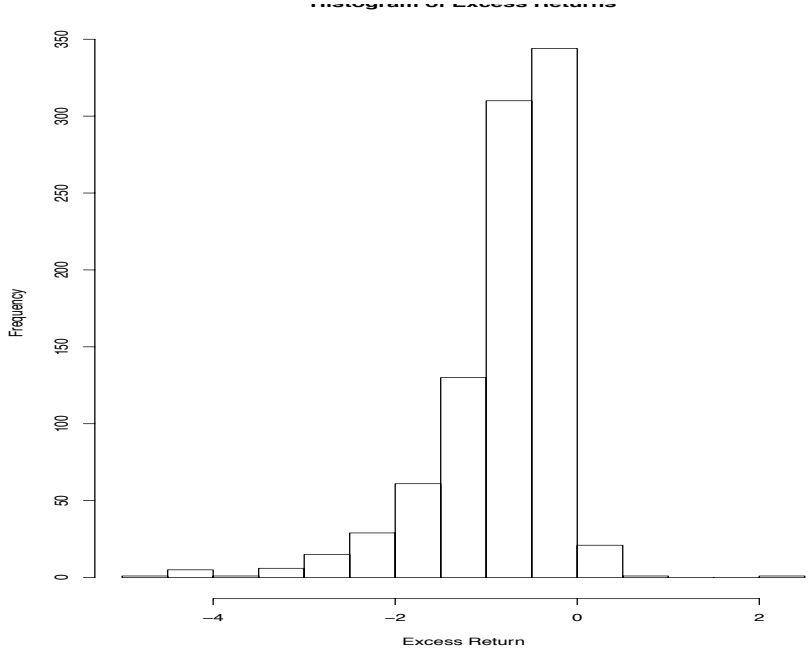

Figure 5.16: Histogram of daily excess returns from Avellaneda and Lee strategy: 2006–2009

5.11.3 Choosing ETFs

The number of ETFs has increased dramatically in recent years. The idea of forming a portfolio with one stock and all of the related ETFs is not so easy to implement as it was in the past. For example, there are almost 30 ETFs in the technology sector. Some ETFs may be more related to certain stocks than are others.

The lasso is another Machine Learning method for taking a large number of predictors and finding a small number to use in a regression model. Since the Avellaneda and Lee method uses regression on ETFs as a major step, we can use the lasso to choose which ETFs we want to associate with each stock. The lasso is an L^1-penalized (or constrained) least-squares method. Suppose that we observe pairs (x_i, Y_i) $(i = 1, \ldots, n)$, where each x_i is a p-dimensional vector of potential predictors and Y_i is the corresponding response. The solution $\hat{\beta}$ is the minimizer of

$$\sum_{i=1}^{n}(Y_i - \beta^\top x_i)^2 + \lambda \sum_{j=1}^{p} |\beta_j|.$$

The corresponding constrained minimization problem is to minimize $\sum_{i=1}^{n}(Y_i - \beta^\top x_i)^2$ subject to $\sum_{j=1}^{p} |\beta_j| \leq c$. The problems are equivalent so long as λ increases as c decreases in the appropriate way. As $\lambda \to \infty$, the L^1 penalty overwhelms the sum of squares, and all β_j eventually become 0. At $\lambda = 0$, we get the least-squares fit. The reason that the lasso produces sparse solutions is the same as for the graphical lasso. It is common to standardize the predictors before running the lasso, otherwise the β_j are on different scales and the L^1 penalty is less compelling.

For each stock, we can run the lasso for finding a model prediciting the stock returns based on the ETF returns, and choose a set of ETFs of an appropriate size. We can then apply the Avellaneda and Lee procedure with the chosen ETFs. The *R* package `glmnet` fits the lasso to linear regression models as well as other generalized linear models. It also fits the *elastic net* which is a combination of ridge regression and lasso. We will not deal with the elastic net here, but it could be worth pursuing as a portfolio selection tool.

The output from `cv.glmnet` is a *list* containing several useful items:

- `lambda` is a list of the λ values used. There is a known formula for the largest λ needed, the first λ that forces all coefficients to be 0. The entries in `lambda` run from largest to smallest.
- `nzero` is a list of the numbers of nonzero coefficients corresponding to each λ. The values of `nzero` increase from 0 to the number of potential predictors. We want the last one that is no greater than the desired number of ETFs.
- `glmnet.fit` is itself a list that contains the detailed fit information. The most important is `glmnet.fit$beta` which is a matrix containing all of the fitted regression coefficients. There is one column for each λ and one row for each predictor.

For example, for stock i, we would use the following commands if we wanted at most `k1` ETFs in the portfolio and if we were using returns on days `d1` to `d2`,

```
nomiss.etf=!apply(etf$price[(d1-1):d2,],2,function(y){any(is.na(y))})
etf.ret=log(etf$price[d1:d2,nomiss]/etf$price[(d1-1):(d2-1),nomiss])
stock.ret=log(stock$price[d1:d2,i]/stock$price[(d1-1):(d2-1),i])
lassofit=cv.glmnet(etf.ret,stock.ret)
k2=which(lassofit$nzero>k1)[1]-1
portf=which(lassofit$glmnet.fit$beta[,k2]!=0)
coef=lassofit$glmnet.fit$beta[portf,k2]
resid=price.ret-etf.ret[,portf]%*%matrix(coef,k1,1)-lassofit$glmnet.fit$a0
```

Then `portf` tells us which columns of `etf$price[,nomiss]` have the ETFs we want to use. The original indices of these ETFs are `which(nomiss.etf)[portf]`. The residuals from the lasso fit are `resids`, in the above code. These can then be used in the Avellaneda and Lee procedure.