SAPIENZA
UNIVERSITÀ DI ROMA

Dipartimento di Scienze Statistiche
Sezione di Statistica Economica ed Econometria

Søren Johansen

The analysis of nonstationary time series using regression, correlation and cointegration - with anapplication to annual mean temperature and sea level

DSS Empirical Economics and Econometrics
Working Papers Series

DSS-E3 WP 2011/4

DSS Empirical Economics and Econometrics
Working Papers Series

2011/1 Massimo Franchi, Paolo Paruolo "Normal forms of regular matrix polynomials via local rank factorization"

2011/2 Francesca Di Iorio, Stefano Fachin "A Sieve Bootstrap range test for poolability in dependent cointegrated panels"

2011/3 Maria Grazia Pittau, Shlomo Yitzhaki, Roberto Zelli "The make-up of a regression coefficient: An application to gender"

Dipartimento di Scienze Statistiche
Sezione di Statistica Economica ed Econometria
"Sapienza" Università di Roma
P.le A. Moro 5 – 00185 Roma - Italia
http://www.dss.uniroma1.it

The analysis of nonstationary time series using regression, correlation and cointegration – with an application to annual mean temperature and sea level

Søren Johansen*
University of Copenhagen
and CREATES University of Aarhus

November 10, 2011

Abstract

There are simple well-known conditions for the validity of regression and correlation as statistical tools. We analyse by examples the effect of nonstationarity on inference using these methods and compare them to model based inference. Finally we analyse some data on annual mean temperature and sea level, by applying the cointegrated vector autoregressive model, which explicitly takes into account the nonstationarity of the variables.

Keywords: Regression correlation cointegration, model based inference, likelihood inference, annual mean temperature, sea level

JEL Classification: C32.

1 Introduction

The purpose of this chapter is to conduct a statistical analysis of two time series measuring annual mean temperature anomalies and sea level from 1881 to 1995, using a cointegration analysis. We start, however, with a discussion of regression and correlation which are commonly applied statistical techniques, and emphasize the assumptions

*The author gratefully acknowledges support from Center for Research in Econometric Analysis of Time Series, CREATES, funded by the Danish National Research Foundation, and would like to thank Torben Schmith and Peter Thejll for many discussions and useful comments.
Address: Department of Economics, University of Copenhagen, Øster Farimagsgade 5, DK-1353 Copenhagen K. Denmark, Email: Soren.Johansen@econ.ku.dk

underlying the analysis in order to point out some instances, where these method cannot be used in a routinely fashion, namely when the variables are nonstationary, either because they contain a deterministic trend or a random walk.

Thus we consider two time series X_t and $Y_t, t = 1, \ldots, T$, and a substantive theory that X influences Y in a linear fashion formulates as $Y = \beta X$. For given data such a relation does not hold and there is most often no substantive theory for the deviations, and to quote Haavelmo (1943) 'we need a stochastic formulation to make simplified relations elastic enough for applications'. We therefore introduce the error term ε_t and write the relation as a statistical or semi-empirical relation

$$Y_t = \beta X_t + \varepsilon_t, \ t = 1, \ldots, T. \tag{1}$$

We want to estimate the parameter β and evaluate its uncertainty in order to be able to test hypotheses, for instance that $\beta = 0$, which means that there is no influence of X_t on Y_t.

For notational reasons we formulate the discussion of regression and correlation for two variables only. As a general reference we use the textbook by von Storch and Zwiers (2002), referred as (SZ 1998), for statistical concepts and the basic results for regression, correlation and stationary (ergodic) time series.

2 Two approaches to inference

There are two common approaches to deal with inference in linear regression and correlation analysis

- *The method based approach*

  Regression is used to estimate the effect of X on Y by calculating the least squares estimators and the residual error variance using the formulae

$$\hat{\beta} = \frac{\sum_{t=1}^{T} X_t Y_t}{\sum_{t=1}^{T} X_t^2}, \tag{2}$$

$$\hat{\sigma}^2 = T^{-1} \sum_{t=1}^{T} (Y_t - \hat{\beta} X_t)^2. \tag{3}$$

  These are then used to conduct asymptotic inference by comparing the t-ratio

$$t_{\beta=\beta_0} = (\sum_{t=1}^{T} X_t^2)^{1/2} \frac{\hat{\beta} - \beta_0}{\hat{\sigma}}, \tag{4}$$

  with the quantiles of a standard normal distribution. Regression works well if the estimates $\hat{\beta}$ and $\hat{\sigma}^2$ are close to their theoretical counterparts, β and σ^2, and if the asymptotic distribution of $t_{\beta=\beta_0}$ is close to the Gaussian distribution. We discuss below some examples, where there is no relation between the empirical regression estimates and the theoretical values.

Correlation is used to describe the linear relation between two observed variables Y and X. We define the *theoretical* correlation coefficient between Y and X as

$$\rho = \frac{Cov(X,Y)}{\sqrt{Var(Y)Var(X)}}, \tag{5}$$

and the *empirical* correlation coefficient between two time series Y_t and X_t is calculated as

$$\hat{\rho} = \frac{\sum_{t=1}^{T}(X_t - \bar{X})(Y_t - \bar{Y})}{\sqrt{\sum_{t=1}^{T}(X_t - \bar{X})^2 \sum_{t=1}^{T}(Y_t - \bar{Y})^2}}. \tag{6}$$

both (5) and (6) are commonly called correlation, which causes some confusion. We distinguish here using the qualifications *empirical* and *theoretical*, and we discuss below some examples where the empirical correlation is not related to the theoretical correlation.

- *The model based approach*

  In the model based approach we first formulate a hypothetical mechanism for how the data is generated and then derive the relevant statistical methodology by an analysis of the likelihood function (SZ p. 88). One such model, which also specifies how X_t is generated, is

  $$\begin{aligned} Y_t &= \beta X_t + \varepsilon_{1t}, & (7)\\ X_t &= \xi X_{t-1} + \varepsilon_{2t}, & (8) \end{aligned}$$

  where $\varepsilon_t = (\varepsilon_{1t}, \varepsilon_{2t})$ are i.i.d. Gaussian with variances σ_1^2 and σ_2^2 and covariance σ_{12}. We then conduct inference using the method of maximum likelihood and likelihood ratio test. These methods, however, require that the assumptions of the model are carefully checked in any particular application in order to show that the model describes the data well, so that the results of asymptotic inference, which are derived under the assumptions of the model, can be applied.

It is well known that linear regression analysis can be derived as the Gaussian maximum likelihood estimator provided that ε_t in (1) are i.i.d. $N(0, \sigma^2)$, and X_t is nonstochastic, see (SZ 1998, p. 151). Similarly if (X_t, Y_t) are i.i.d. Gaussian with variances σ_1^2, σ_2^2 and covariance σ_{12}, then the theoretical correlation is $\rho = \sigma_{12}/\sigma_1\sigma_2$, and the maximum likelihood estimator of ρ is $\hat{\rho}$ given in (6). Thus there is no clear-cut distinction between the method based approach and the model based approach, but a difference of emphasis, in the sense that regression and correlation are often applied uncritically by "pressing the button on the computer", and the model based method requires more discussion and checking of assumptions.

We discuss below some examples where regression analysis and correlation analysis cannot be used, and hence one has to take properties of the data into account in order to avoid incorrect inference.

3 Regression and Correlation

We specify a set of conditions under which regression and correlation methods work well, and then analyse some examples where the methods do not work.

3.1 Regression

We formulate the statistical assumptions of the regression model (1) as

Assumption 1 *We assume that*

- $\varepsilon_1, \ldots, \varepsilon_T$ *are innovations in the sense that they are i.i.d.* $(0, \sigma^2)$ *and* ε_t *is independent of* $X_1, \ldots, X_t,\ t = 1, \ldots, T$
- $X_1, \ldots, X_T$ *are stochastic (or deterministic) variables for which the normalized sum of squares is convergent to a deterministic limit*

$$n_T^{-1} \sum_{t=1}^{T} X_t^2 \xrightarrow{P} \Sigma > 0,$$

for some sequence $n_T \to \infty$.

These assumptions are enough to show that

$$E(n_T^{-1/2} \varepsilon_t X_t | X_1, \ldots, X_t) = 0, \tag{9}$$

$$n_T^{-1} \sum_{t=1}^{T} Var(\varepsilon_t X_t | X_1, \ldots, X_t) \xrightarrow{P} \sigma^2 \Sigma > 0. \tag{10}$$

Apart from a technical assumptions on the third moment, these relations show that $n_T^{-1/2} \varepsilon_t X_t$ is a so-called martingale difference sequence, and that the sum of its successive conditional variances converges to a deterministic limit. This again implies that one can apply the Central Limit Theorem for martingales, see Hall and Heyde (1980). The theorem shows, in this particular case, that

$$\sum_{t=1}^{T} n_T^{-1/2} \varepsilon_t X_t \xrightarrow{d} N(0, \sigma^2 \Sigma), \tag{11}$$

where $\xrightarrow{d}$ means convergence in distribution (SZ p. 46).

From (2) and (3) we find that

$$\hat{\beta} = \frac{\sum_{t=1}^{T} Y_t X_t}{\sum_{t=1}^{T} X_t^2} = \frac{\sum_{t=1}^{T} (\beta X_t + \varepsilon_t) X_t}{\sum_{t=1}^{T} X_t^2} = \beta + n_T^{-1/2} \frac{n_T^{-1/2} \sum_{t=1}^{T} \varepsilon_t X_t}{n_T^{-1} \sum_{t=1}^{T} X_t^2},$$

$$\hat{\sigma}^2 = T^{-1} [\sum_{t=1}^{T} \varepsilon_t^2 - \frac{(n_T^{-1/2} \sum_{t=1}^{T} \varepsilon_t X_t)^2}{n_T^{-1} \sum_{t=1}^{T} X_t^2}].$$

The result (11) then implies that

$$\hat{\beta} \xrightarrow{P} \beta, \tag{12}$$

$$\hat{\sigma} \xrightarrow{P} \sigma^2, \tag{13}$$

$$n_T^{1/2}(\hat{\beta} - \beta) \xrightarrow{d} N(0, \sigma^2 \Sigma^{-1}), \tag{14}$$

$$t_{\beta=\beta_0} = (\sum_{t=1}^{T} X_t^2)^{1/2} \frac{(\hat{\beta} - \beta_0)}{\hat{\sigma}} \xrightarrow{d} N(0,1). \tag{15}$$

The first two results state that the estimators are close to the theoretical values, that is, the estimators are consistent, and the third that $\hat{\beta}$ is asymptotically normally distributed. The last result is used to conduct asymptotic inference and test the hypothesis that $\beta = \beta_0$, by comparing a $t-$ratio with the quantiles of the normal distribution. In this sense the regression method works well when the above Assumption 1 is satisfied.

3.2 Correlation

We formulate the condition that guarantees that the theoretical correlation can be measured by the empirical correlation.

Assumption 2 *We assume that (Y_t, X_t) is a stationary and ergodic time series with finite second moments.*

It follows from the Law of Large Numbers for ergodic processes that if Assumption 2 is satisfied, then

$$\hat{\rho} \xrightarrow{P} \rho. \tag{16}$$

Thus in order for the calculation of an empirical correlation to make sense as an approximation to the theoretical correlation, it is important to check Assumption 2.

This problem was pointed out by Yule (1926) in his presidential address to The Royal Statistical Society, and he introduced the concept of spurious or nonsense correlation, and showed by simulation that for some nonstationary processes, the empirical correlation seems not to converge in probability, even if the processes are independent. This was later discussed by Granger and Newbold (1974), and Phillips (1986) found the limit distributions in terms of Brownian motion.

3.3 Examples

The first example shows that we have to choose different normalizations depending on which regressor variable we have.

Example 1. (Regression) If $X_t = 1$ we have $\sum_{t=1}^{T} X_t^2 = T$ and we choose $n_T = T$, and if $X_t = t$, then $\sum_{t=1}^{T} X_t^2 = \sum_{t=1}^{T} t^2 \approx \frac{1}{3}T^3$, and we choose $n_T = T^3$. If X_t is

an ergodic process with $E(X_t^2) < \infty$, see (SZ 2002, p. 202), then the Law of Large Numbers for ergodic processes shows that $T^{-1}\sum_{t=1}^{T} X_t^2 \xrightarrow{P} E(X_t^2)$. Hence we use the normalization $n_T = T$ in this case. This, however, is not enough to apply the regression method because we also need ε_t to be independent of the regressor, see Assumption 1.

Consider for instance the model defined in (7) and (8) for $|\xi| < 1$, which defines an ergodic process X_t. Then $T^{-1}\sum_{t=1}^{T} X_t^2 \xrightarrow{P} Var(X_t) = \sigma_2^2/(1-\xi^2)$, but note that (9) fails because

$$E(\varepsilon_{1t}X_t|X_1,\ldots,X_t) = X_t E(\varepsilon_{1t}|\varepsilon_{2t}) = \sigma_{12}\sigma_2^{-2}X_t\varepsilon_{2t} = \sigma_{12}\sigma_2^{-2}X_t(X_t - \xi X_{t-1}) \neq 0,$$

when ε_{1t} is not independent of the regressor, and we cannot apply the asymptotic theory unless $\sigma_{12} = 0$. Thus even for stationary processes an autocorrelated regressor variable is enough to invalidate the simple regression.

If, however, we take the model based approach we can analyse model (7) and (8) as follows. We first find the conditional mean of Y_t given $X_1,\ldots,X_t$:

$$E(Y_t|X_1,\ldots,X_t) = \beta X_t + E(\varepsilon_{1t}|X_1,\ldots,X_t) = \beta X_t + \sigma_{12}\sigma_2^{-2}(X_t - \xi X_{t-1}).$$

This means we can replace (7) and (8) by the equations

$$\begin{aligned} Y_t &= \beta X_t + \sigma_{12}\sigma_2^{-2}(X_t - \xi X_{t-1}) + \varepsilon_{1t} - \sigma_{12}\sigma_2^{-2}\varepsilon_{2t}, && (17)\\ X_t &= \xi X_{t-1} + \varepsilon_{2t}. && (18) \end{aligned}$$

Because the error terms $\varepsilon_{1t} - \sigma_{12}\sigma_2^{-2}\varepsilon_{2t}$ and ε_{2t} are independent, we can analyse the equations separately and estimate ξ by regressing X_t on X_{t-1}, and determine $\beta + \sigma_{12}\sigma_2^{-2}$ and $-\sigma_{12}\sigma_2^{-2}\xi$ by regression of Y_t on X_t and X_{t-1}, and that allows one to derive consistent asymptotically Gaussian estimators for the parameter of interest β. Thus by analysing the model we can determine the relevant regression analysis. ■

*Example 2 (*Correlation) Let again the data be generated by (7) and (8) for $|\xi| < 1$. Then X_t, Y_t is an ergodic process and the empirical correlation, $\hat{\rho}$, will converge towards the theoretical correlation

$$\rho = \frac{Cov(\beta X_t + \varepsilon_{1t}, X_t)}{\sqrt{Var(\beta X_t + \varepsilon_{1t})Var(X_t)}} = \frac{\beta\sigma_2^2 + \sigma_{12}(1-\xi^2)}{\sqrt{[\beta^2\sigma_2^2 + \sigma_1^2(1-\xi^2) + 2\beta\sigma_{12}(1-\xi^2)]\sigma_2^2}},$$

using the results that $Var(X_t) = \sigma_2^2/(1-\xi^2)$ and $Cov(X_t, \varepsilon_{1t}) = \sigma_{12}$.

If X_t instead is generated by

$$X_t = \gamma t + \varepsilon_{2t}, \tag{19}$$

then

$$Y_t = \beta\tau t + \beta\varepsilon_{2t} + \varepsilon_{1t}$$

and correlation analysis does not work. We find $E(X_t) = \gamma t$ and $E(Y_t) = \beta\gamma t$, so that the theoretical correlation is

$$\rho = \frac{E(Y_t - \beta\gamma t)(X_t - \gamma t)}{\sqrt{E(Y_t - \beta\gamma t)^2 E(X_t - \gamma t)^2}} = \frac{E((\varepsilon_{1t} + \beta\varepsilon_{2t})\varepsilon_{2t})}{\sqrt{E(\varepsilon_{1t} + \beta\varepsilon_{2t})^2 E(\varepsilon_{2t}^2)}} = \frac{\sigma_{12} + \beta\sigma_2^2}{\sqrt{(\beta^2\sigma_2^2 + 2\beta\sigma_{12} + \sigma_1^2)\sigma_2^2}},$$

that is, the correlation between the stochastic error term of Y_t and X_t.

The empirical correlation, however, measures something quite different. It contains the averages $\bar{X} = \gamma\bar{t} + \bar{\varepsilon}_2$, where $\bar{t} = T^{-1}\sum_{t=1}^{T} t = (T+1)/2$, so that $X_t - \bar{X} = \gamma(t - \bar{t}) + \varepsilon_{2t} - \bar{\varepsilon}_2$ and $Y_t - \bar{Y} = \beta(X_t - \bar{X}) + \varepsilon_{1t} - \bar{\varepsilon}_1 = \beta\gamma(t - \bar{t}) + \beta(\varepsilon_{2t} - \bar{\varepsilon}_2) + \varepsilon_{1t} - \bar{\varepsilon}_1$ are dominated by the linear trend and we have

$$\hat{\rho} \xrightarrow{P} \frac{\beta}{|\beta|} = \pm 1,$$

if $\beta \neq 0$. Thus, if the regressor is trending with a linear trend, there is no relation between the empirical correlation, which is often very close to ± 1, and the theoretical correlation which measures a correlation between the error terms. The mistake made is of course that $\bar{X}$ and $\bar{Y}$ do not measure the expectation of X_t and Y_t.

The model based approached leads to estimating $(\beta\gamma, \gamma)$ from a regression of (Y_t, X_t) on t and that gives consistent asymptotically Gaussian estimators of the parameters of interest without using or misusing any measure of correlation.

A good check of the relevance of the empirical correlation is very simply to calculate it recursively, that is, define $\hat{\rho}_t$ based on date up to time t, and then plot it and check if it is reasonably constant in t. ■

Next we give an example where one cannot normalize $\sum_{t=1}^{T} X_t^2$ so that the limit exists as a deterministic limit, and hence that simple regression analysis may fail.

Example 3. (Random walk regressor) A very special situation occurs in example (7) and (8) if $\xi = 1$, so that X_t is stochastic and nonstationary in the sense that,

$$X_t = \sum_{i=1}^{t} \varepsilon_{2i} + X_0.$$

In this case $E(X_t|X_0) = X_0$ and $Var(X_t|X_0) = \sigma_2^2 t$ which increases to infinity, and something completely different happens. Let us first find out how to normalize $E(\sum_{t=1}^{T} X_t^2|X_0)$, because such a normalization could be a good candidate for the normalization of $\sum_{t=1}^{T} X_t^2$. We find

$$E(\sum_{t=1}^{T} X_t^2|X_0) = \sum_{t=1}^{T} E(X_t^2|X_0) = \sigma_2^2 \sum_{t=1}^{T} t = \frac{1}{2}\sigma_2^2 T(T+1).$$

Thus a good choice seems to be $n_T = T^2$, which at least makes sure that the mean converges when normalized by T^2.

Unfortunately $T^{-2}\sum_{t=1}^{T} X_t^2$ does not converge to a deterministic limit but to a stochastic variable. The detailed theory of this is quite complicated because it involves Brownian motion.

Brownian motion is a continuous stochastic process defined on the unit interval for which $B(0) = 0, B(u)$ is distributed as $N(0, u)$ and for $0 \leq u_1 < u_2 < u_3 \leq 1$ we have that $B(u_2) - B(u_1)$ is independent of $B(u_3) - B(u_2)$. The main reason for this to be interesting in the present context, is that we can approximate Brownian motion by random walks, because

$$T^{-1/2} \sum_{1 \leq i \leq Tu} \begin{pmatrix} \varepsilon_{1i} \\ \varepsilon_{2i} \end{pmatrix} \xrightarrow{d} \begin{pmatrix} \sigma_1 B_1(u) \\ \sigma_2 B_2(u) \end{pmatrix}, \ 0 \leq u \leq 1 \tag{20}$$

Thus a Brownian motion can be thought of as a random walk with a very large number of steps, and that is how its properties are studied using stochastic simulation. The two Brownian motions in (20) are correlated with correlation $\rho = \sigma_{12}/\sigma_1\sigma_2$.

Two fundamental results about Brownian motion are

$$T^{-2} \sum_{t=1}^{T} X_t^2 \xrightarrow{d} \sigma_2^2 \int_0^1 B_2(u)^2 du,$$
$$T^{-1} \sum_{t=1}^{T} X_t \varepsilon_{1t} \xrightarrow{d} \sigma_2 \sigma_1 \int_0^1 B_2(u)(dB_1).$$

These limits are stochastic variables, and for our purpose the main result is that the product moments should be normalized by T^2 and T respectively to get convergence. It follows that Assumption 1 is not satisfied because the limit of $T^{-2}\sum_{t=1}^{T} X_t^2$ is stochastic, and we cannot count on the results (12) to (16) to be correct.

If we run a regression anyway, we can calculate the t-ratio and find its limit

$$(\sum_{t=1}^{T} X_t^2)^{1/2}(\hat{\beta} - \beta) = \frac{T^{-1}\sum_{t=1}^{T} \varepsilon_{1t} X_t}{\sqrt{T^{-2}\sum_{t=1}^{T} X_t^2}} \xrightarrow{d} \frac{\sigma_1 \int_0^1 B_2(u) dB_1(u)}{\sqrt{\int_0^1 B_2(u)^2 du}}. \tag{21}$$

If ε_{1t} and ε_{2t} are independent, one can show that the limit distribution (21) is $N(0, \sigma_1^2)$, and therefore (12) and (15) hold anyway, whereas (14) is different, because we get instead a so-called mixed Gaussian distribution of the limit of $T(\hat{\beta} - \beta)$. So despite the fact the $\hat{\beta}$ is not asymptotically normally distributed one can still test hypotheses on β using the usual t-ratio, but the independence of ε_{1t} and ε_{2t} is crucial for this last result. A simulation is show in Figure 1. It is seen that for $\rho = 0$, where there is independence between the regressor and the error term in the regression, the distribution of the t-ratio is very close to Gaussian, but the distribution of $T(\hat{\beta} - \beta)$ is centered around zero, but far from Gaussian.

The result in (21) shows that applying a simple regression analysis, without checking Assumption 1, can be seriously misleading, and we next want to show how we can solve the problem of inference by analysing the model, that generated the data.

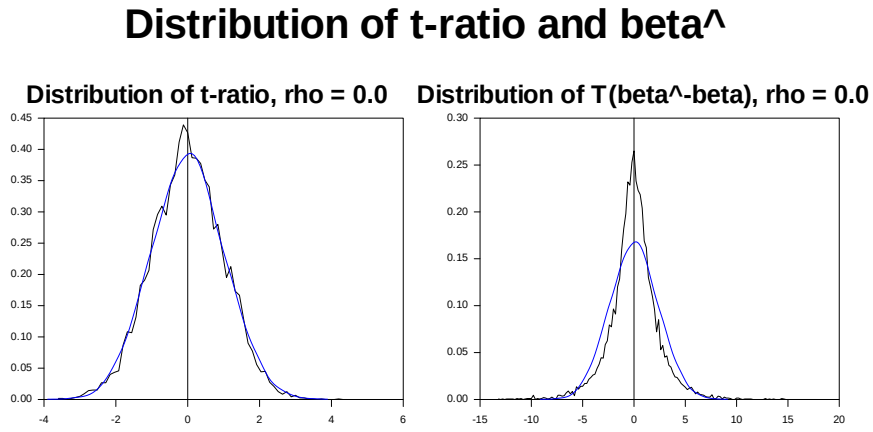

Figure 1: The densities are based upon 10.000 simulations of $T = 100$ observations. The plots show simulations of the t-ratio, (15), and $T(\hat{\beta} - \beta)$, (14), in the regression of $Y_t = \beta X_t + \varepsilon_{1t}$, when X_t is a random walk, $\Delta X_t = \varepsilon_{2t}$, see Example 3, and ε_{1t} is independent of ε_{2t}. Each plot contains a Gaussian density for comparison. It is seen that the t-ratio has approximately a Gausisan distirbution and that the estimator normlized by T has a distribution with longer tails than the Gaussian.

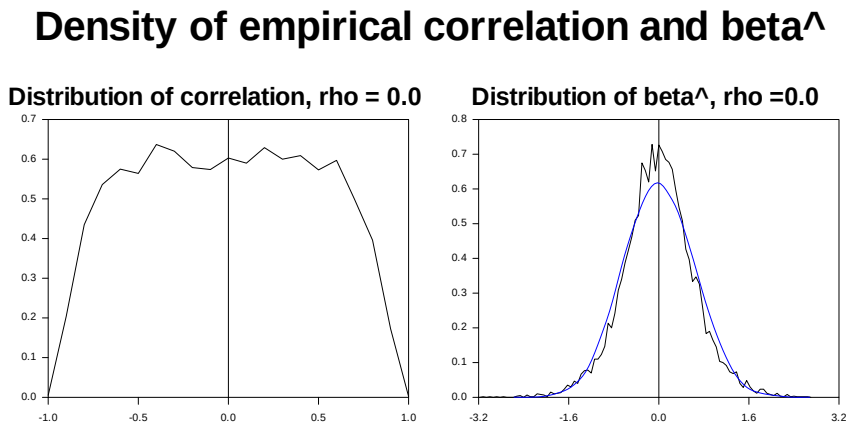

Figure 2: The densities are based upon 10.000 simulations of $T = 100$ observations. The left panel shows the distribution of the empirical correlation between two independent random walks. The results are the same for higher values of T, thus there is no tendency to converge to $\rho = 0$. The right panel shows the similar results for the empirical regression coefficient.

If $\xi = 1$, then $\Delta X_t = \varepsilon_{2t}$, and we find the equations, see (17) and (18)

$$\begin{aligned} Y_t &= \beta X_t + \sigma_{12}\sigma_2^{-2}\Delta X_t + \varepsilon_{1t} - \sigma_{12}\sigma_2^{-2}\varepsilon_{2t}, \\ \Delta X_t &= \varepsilon_{2t}. \end{aligned} \tag{22}$$

Here the errors are independent and

$$Var(\varepsilon_{1t} - \sigma_{12}\sigma_2^{-2}\varepsilon_{2t}) = Var(\varepsilon_{1t}|\varepsilon_{2t}) = \sigma_1^2 - \sigma_{12}^2\sigma_2^{-2} = \sigma_{1|2}^2.$$

Equation (22) is analysed by regression of Y_t on X_t and ΔX_t to find an asymptotically Gaussian estimator for β. This simple modification of the regression problem solves the inference problem. We still get an expression like (21)

$$(\sum_{t=1}^{T} X_t^2)^{1/2}(\hat{\beta} - \beta) \xrightarrow{d} \frac{\sigma_{1|2}\int_0^1 B_2(u)dB_{1|2}(u)}{\sqrt{\int_0^1 B_2(u)^2du}}, \tag{23}$$

where $B_{1|2}(u) = B_1(u) - \rho B_2(u)$ is independent of B_2, so the limit is mixed Gaussian and inference can be conducted using the usual t-ratio and comparing it to the quantiles of the Gaussian distribution.

The correlation analysis of Y_t and X_t leads to a theoretical correlation (conditional on X_0)

$$\rho_t = \frac{Cov(\beta X_t + \varepsilon_{1t}, X_t|X_0)}{\sqrt{Var(\beta X_t + \varepsilon_{1t}|X_0)Var(X_t|X_0)}} = \frac{\beta\sigma_2^2 t + \sigma_{12}}{\sqrt{[\beta^2\sigma_2^2 t + \sigma_1^2 + 2\beta\sigma_{12})]\sigma_2^2 t}} \to \frac{\beta}{|\beta|} = \pm 1,$$

if $\beta \neq 0$. Thus for large t we find a value ± 1 depending on the sign of β.

The empirical correlation coefficient has the same limit

$$\hat{\rho} = \frac{\beta\sum_{t=1}^T (X_t - \bar{X})^2 + \sum_{t=1}^T (\varepsilon_{1t} - \bar{\varepsilon}_1)(X_t - \bar{X})}{\sqrt{\sum_{t=1}^T (\beta(X_t - \bar{X}) + \varepsilon_{1t})^2 \sum_{t=1}^T (X_t - \bar{X})^2}} \xrightarrow{P} \frac{\beta}{|\beta|} = \pm 1,$$

if $\beta \neq 0$, so that it estimates the limit of the theoretical correlation for $T \to \infty$.

This model with $\xi = 1$ is an example of two nonstationary variables with a stationary linear combination, that is, a model for cointegration. ■

Example 4. (Spurious correlation and regression)
Assume (X_t, Y_t) are generated by the equations

$$\begin{aligned} \Delta Y_t &= \varepsilon_{1t}, \\ \Delta X_t &= \varepsilon_{2t}, \end{aligned}$$

where we assume that $\sigma_{12} = 0$, so X_t and Y_t are independent of each other. The theoretical correlation is, conditioning on initial values,

$$\rho_t = \frac{Cov(Y_t, X_t|Y_0, X_0)}{\sqrt{Var(Y_t|Y_0)Var(X_t|X_0)}} = \frac{t\sigma_{12}}{\sqrt{t\sigma_1^2 t\sigma_2^2}} = \frac{\sigma_{12}}{\sigma_1\sigma_2} = 0.$$

If we calculate the empirical correlation, (6), all product moments should be normalized by T^{-2} and we find the limit

$$\hat{\rho} \xrightarrow{d} \frac{\int_0^1 (B_2(u) - \bar{B}_2)(B_1(u) - \bar{B}_1)du}{\sqrt{\int_0^1 (B_1(u) - \bar{B}_1)^2 du \int_0^1 (B_2(u) - \bar{B}_2)^2 du}}.$$

Thus $\hat{\rho}$ does not converge to zero or any other value but is stochastic even for infinitely many observations. This phenomenon was observed by Yule (1926) who simulated the limit distribution by producing random uniform integers from -10 to 10, using a deck of cards and found a distribution between 0 and 1, see Figure 2 for a simulation of the distribution. He called this "nonsense correlation".

A regression of Y_t on X_t gives similarly

$$\hat{\beta} \xrightarrow{d} \frac{\int_0^1 B_2(u)B_1(u)du}{\int_0^1 B_2(u)^2 du},$$

where the stochastic limit is totally unrelated to any theoretical measure of the effect of X_t on Y_t. Thus by calculation of a correlation or a regression coefficient one may infer an effect of X_t on Y_t, when absolutely no effect is present because they are independent, see Figure 2.

If the independent random walks contain a trend, we model them as

$$\Delta Y_t = \varepsilon_{1t} + \mu_1, \; Y_t = \sum_{i=1}^{t} \varepsilon_{1i} + \mu_1 t + Y_0, \tag{24}$$

$$\Delta X_t = \varepsilon_{2t} + \mu_2, \; X_t = \sum_{i=1}^{t} \varepsilon_{2i} + \mu_2 t + X_0, \tag{25}$$

where we again assume $\sigma_{12} = 0$. In this case, the trend is dominating the random walk, and we find that for instance

$$T^{-1}(X_t - \bar{X}) = T^{-1}\sum_{i=1}^{t} \varepsilon_{2i} - T^{-1}\sum_{t=1}^{T}[T^{-1}\sum_{i=1}^{t} \varepsilon_{2i}] + \mu_2(\frac{t}{T} - \frac{T+1}{2T}) \xrightarrow{P} \mu_2(u - 1/2),$$

for $t/T \to u$, because $T^{-1}\sum_{i=1}^{t} \varepsilon_{2i} \xrightarrow{P} 0$. It follows that because $\sum_{t=1}^{T}(t - \bar{t})^2 \approx T^3/3$ we get

$$\hat{\rho} \xrightarrow{P} \frac{\mu_2\mu_1}{|\mu_2\mu_1|} = \pm 1,$$

if $\mu_1\mu_2 \neq 0$. Thus, despite the fact that Y_t and X_t are stochastically independent, an empirical correlation suggests something quite different.

The regression coefficient satisfies similarly

$$\hat{\beta} \xrightarrow{P} \frac{\mu_1}{\mu_2},$$

which is the ratio of the slopes of the trends, which makes some sense, but an analysis of the data, using the model (24) and (25), would find a linear trend in each variable and estimates of μ_1 and μ_2 which would contain more information.

It is therefore very easy to calculate an empirical correlation between two variables that are completely uncorrelated, but which each depend on the same third variable, like here a time trend. It is important in the calculation of correlations to replace $E(X_t)$ and $E(Y_t)$ by reasonable estimates, not use averages.

Sober (2001), considered the example of Venetian sea levels and British bread prices. He claims they are truly correlated but not causally connected by construction. The claim of "true correlation" is based on the calculation of the empirical correlation, which of course is very high, because both variables trend with time. ■

4 The cointegrated vector autoregressive model

Cointegration was introduced in econometrics by Granger (1981) because many macro variables show nonstationarity of the random walk type, but also clear co-movement. Engle and Granger (1987) contains the first statistical analysis of cointegration using regression methods, and Phillips (1991) modified the regression approach to allow for valid inference. The analysis of cointegration and model based inference in the vector autoregressive framework was initiated by Johansen (1988). The technique of cointegration is described in most text book on times series econometrics and many computer programs are available, see for instance Cats for Rats, (Dennis *et al.* 2005), which was used for the calculations in section 5. For a systematic account of the theory, see Johansen (1996), and for applications the monograph by Juselius (2006) is recommended. A recent survey is given in Johansen (2006).

Below we give a simple example of such a model and discuss briefly the statistical analysis of the model.

4.1 An example of a model for cointegration

We consider two variables X_t and Y_t which are generated by the equations

$$\Delta Y_t = \tau(Y_{t-1} - \gamma X_{t-1}) + \varepsilon_{1t}, \tag{26}$$

$$\Delta X_t = \eta(Y_{t-1} - \gamma X_{t-1}) + \varepsilon_{2t},\ t = 1, \ldots, T \tag{27}$$

The special choices of $\tau = -1$, $\eta = 0$, and $\gamma = \beta$ give the model (7) and (8) with a redefinition of the error term. Each equation is linear in past variables, but note that the levels Y_{t-1} and X_{t-1} enter only through the same linear combination $U_{t-1} = Y_{t-1} - \gamma X_{t-1}$ in both equations. We call U_{t-1} the disequilibrium error and think of the relation $Y = \gamma X$ as an equilibrium relation, to which the variables react with adjustment coefficients τ and η respectively.

It is seen that the equation for $U_t = Y_t - \gamma X_t$ is

$$\Delta U_t = (\tau - \gamma\eta)U_{t-1} + \varepsilon_{1t} - \gamma\varepsilon_{2t},$$

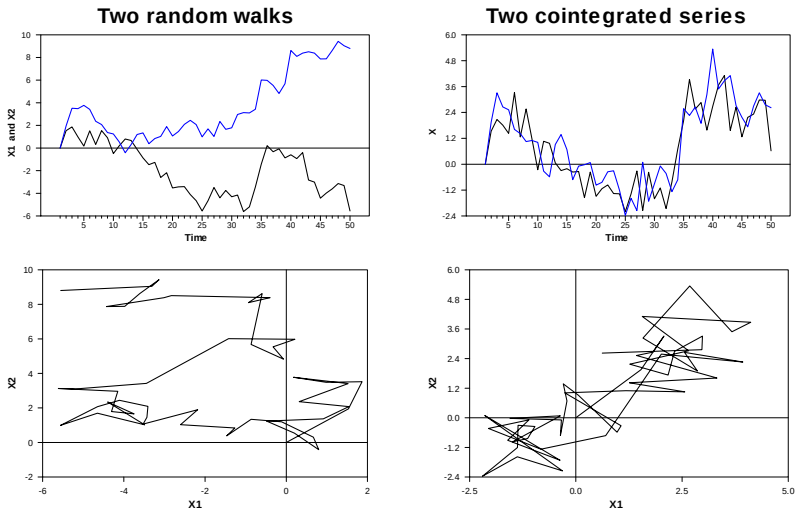

Figure 3: Plots of integrated series generated by the equations (26) and (27). To the left are two random walks ($\eta = \tau = 0$). To the right are two cointegrated random walks ($\gamma = 1, \tau = -1/2, \eta = 1/2$). Note how they follow each other in the upper panel and move around the line $Y - X = 0$ in the lower panel.

so that U_t is an autoregressive process with one lag, which is stationary if $|1+\tau-\gamma\eta| < 1$. By eliminating U_{t-1} from (26) and (27) we get

$$\eta\Delta Y_t - \tau\Delta X_t = \eta\varepsilon_{1t} - \tau\varepsilon_{2t},$$

which, by summation, shows that

$$\eta Y_t - \tau X_t = \sum_{i=1}^{t}(\eta\varepsilon_{1i} - \tau\varepsilon_{2i}) + \eta Y_0 - \tau X_0 = S_t,$$

where S_t, is a random walk and hence nonstationary.

The solution of the equations can be expressed as

$$\begin{pmatrix} X_t \\ Y_t \end{pmatrix} = \frac{1}{\eta\gamma - \tau}\begin{pmatrix} S_t - \eta U_t \\ \gamma S_t - \tau U_t \end{pmatrix} = \frac{1}{\eta\gamma - \tau}\left[\begin{pmatrix} 1 \\ \gamma \end{pmatrix}\begin{pmatrix} \eta \\ -\tau \end{pmatrix}' \sum_{i=1}^{t}\varepsilon_i - \begin{pmatrix} \eta \\ \tau \end{pmatrix} U_t\right] \tag{28}$$

which is a special case of the general formula below, see (30).

That is, the model produces nonstationary variables, each of which is composed of a stationary and a nonstationary variable. The linear combination $(1, -\gamma)$ eliminates the common random walk (common trend) and makes the linear combination stationary.

This is expressed by saying that (Y_t, X_t) is nonstationary but cointegrated with cointegrating vector $(1, -\gamma)$ and common stochastic trend S_t, see Granger (1981).

Note that the variables are both modelled and treated similarly, unlike in a regression of Y_t on X_t. Thus for instance if $Y_t - \gamma X_t$ is stationary then so is $\gamma^{-1}Y_t - X_t$, so we can normalize on either one of them, provided the coefficient is nonzero. A cointegration relation is a relation between variables.

4.2 The general vector autoregressive model and its solution

The vector autoregressive model with two lags and a constant term for a $p-$dimensional process X_t is given by the equations,

$$\mathcal{H}_r : \Delta X_t = \alpha\beta' X_{t-1} + \Gamma\Delta X_{t-1} + \mu + \varepsilon_t,\ \varepsilon_t \text{ i.i.d.} N_p(0,\Omega), t = 1,\ldots,T \tag{29}$$

where α and β are $p \times r$ matrices. Note that we need the values X_{-1} and X_0 as initial values in order to be able to determine the process recursively.

We define the polynomial

$$\phi(z) = \det((1-z)I_p - \alpha\beta' z - \Gamma(1-z)z).$$

In order to avoid explosive processes we assume that the roots of $\phi(z) = 0$ satisfy either $|z| > 1$ or $z = 1$, see Tables 1 and 2 where the reciprocal roots (which should satisfy $|z| < 1$ or $z = 1$) are given for the two models analysed. Under a further regularity condition, the solution is nonstationary with stationary differences and given by

$$X_t = C\sum_{i=1}^{t}\varepsilon_i + C\mu t + \sum_{n=0}^{\infty} C_i^*(\varepsilon_{t-i} + \mu). \tag{30}$$

The matrix C satisfies $\beta' C = 0$ and $C\alpha = 0$, and C_i^* are functions of α, β, and Γ. Note that the trend disappears if $\mu = \alpha\rho'$ because $C\alpha = 0$, and in this case there is no linear trend in the solution only a level $\sum_{n=0}^{\infty} C_i^*\mu$, and $E(\beta' X_t) = -\rho'$.The representation (28) is a special case of (30) for a bivariate system where $\beta = (1, -\gamma)'$ and $\alpha = (\tau, \eta)$, so that $\alpha_\perp' = (\eta, -\tau)$. The model defined in (7) and (8) is a special case of (29) for $\alpha_1 = -1,\ \alpha_2 = 0$ and $\beta' = (1, -\gamma)$.

Thus we have seen that

- X_t is nonstationary with linear trend, $C\mu t$, and ΔX_t is stationary.
- $\beta' X_t = \sum_{n=0}^{\infty}\beta' C_i^*(\varepsilon_{t-i} + \mu)$ is stationary, so that X_t is cointegrated with r cointegrating relations β and disequilibrium error $\beta' X_t - E(\beta' X_t)$.
- X_t has $p - r$ common stochastic trends, $\alpha_\perp' \sum_{i=1}^{t}\varepsilon_i$, where $\alpha_\perp$ is $p \times (p-r)$ of full rank and $\alpha'\alpha_\perp = 0$.

4.3 Statistical inference in the cointegrated VAR model

It is important to emphasize that before inference can be made in this model the assumptions of the model should be carefully checked. Thus we have to fit a lag length so that the residuals are close to being i.i.d. We therefore plot the residuals and their autocorrelation function. The Gaussian assumption is not so important for the analysis, but the assumption that the error term is i.i.d. is crucial for the application of the result from the asymptotic theory below.

Thus briefly summarize, we can conduct inference as follows

- First determine the lag length needed to describe the data and check the assumptions behind the model, in particular the independence of the residuals.
- Find the cointegration rank and estimate and interpret the cointegrating relation.
- Simplify the model by testing coefficients to zero.

4.4 The test for cointegrating rank

The rank of α and β is the number of cointegrating relations and it is important either to check ones knowledge of the rank, or estimate it from the data. The statistical formulation starts by considering the unrestricted vector autoregressive model

$$\mathcal{H}_p : \Delta X_t = \Pi X_{t-1} + \Gamma \Delta X_{t-1} + \mu + \varepsilon_t, \tag{31}$$

where ε_t i.i.d. $N(0, \Omega)$ and Π, Γ, μ, and Ω are unrestricted. If we denote

$$\varepsilon_t(\Pi, \Gamma, \mu) = \Delta X_t - \Pi X_{t-1} - \Gamma \Delta X_{t-1} - \mu,$$

then the conditional Gaussian log likelihood function, given the initial values X_{-1} and X_0, is apart from a constant,

$$\log L(\mathcal{H}_p) = -\frac{T}{2}[\log \det(\Omega) + tr\{\Omega^{-1} T^{-1} \sum_{t=1}^{T} \varepsilon_t(\Pi, \Gamma, \mu)\varepsilon_t(\Pi, \Gamma, \mu)'\}]. \tag{32}$$

Note that in (SZ 1998, p. 257) the likelihood function is based upon the joint density of the data. This is not possible for nonstationary variables, like random walks, as there is no joint density. We therefore condition on X_0 and X_{-1}, and consider the conditional density of $X_1, \ldots, X_T$ given X_0 and X_{-1}. It follows that, conditional on initial values, the (conditional) maximum likelihood estimators of $(\Pi, \Gamma, \mu, \Omega)$ in (31) can be found by multivariate regression of ΔX_t on X_{t-1}, ΔX_{t-1}, and a constant. The maximized likelihood function, $L_{\max}(\mathcal{H}_p)$, can be found from (32) by inserting the maximum likelihood estimators $(\hat{\Pi}, \hat{\Gamma}, \hat{\mu}, \hat{\Omega})$.

The hypothesis of r cointegrating relations is formulated as in model (29)

$$\Pi = \alpha\beta',$$

where α and β are $p \times r$ matrices. It turns that the maximum likelihood estimators can be calculated explicitly by an eigenvalue problem, even though this is a nonlinear maximization problem, see for instance Johansen (1996). This gives estimates $(\breve{\alpha}, \breve{\beta}, \breve{\Gamma}, \breve{\mu}, \breve{\Omega})$ and the maximized value, $L_{\max}(\mathcal{H}_r)$, calculated from (32). From this we calculate the likelihood ratio test

$$-2 \log LR(\Pi = \alpha\beta') = -2 \log \frac{L_{\max}(\mathcal{H}_r)}{L_{\max}(\mathcal{H}_p)}.$$

The asymptotic distribution of this statistic is a functional of Brownian motion, which generalizes the so-called Dickey-Fuller test, see Dickey and Fuller (1981), for testing a unit root in a univariate autoregressive model. The asymptotic distribution does not depend on parameters, but depends on the type of deterministic terms and different tables are provided by simulations, because the distributions are analytically quite intractable, see Johansen (1996, Ch. 15). It should be noted that the asymptotic distribution is not a χ^2 distribution as one often finds when applying likelihood methods.

After the rank is determined, and for the data below we find $r = 1$, we often normalize on one of the variables to avoid the indeterminacy in the choice of coefficients. When that is done, one can find the asymptotic distribution of the remaining parameter estimators in order to be able to test hypotheses on these, using either likelihood ratio tests statistics or $t-$test statistics. Thus the only nonstandard test is the test for rank, and all subsequent likelihood ratio tests in the model are asymptotically distributed as $\chi^2(f)$, where f is the number of restrictions being tested.

4.5 Asymptotic distribution of the coefficients of the cointegrating relation

Unlike usual regression, as described in Section 3, the estimators of the parameters in the cointegrating relation are not asymptotically Gaussian. Nevertheless one can estimate scale factors, $\hat{\tau}_i$, so that

$$t_{\beta_i = \beta_{i0}} = \hat{\tau}_i^{-1}(\hat{\beta}_i - \beta_{i0}) \xrightarrow{d} N(0,1). \tag{33}$$

Thus one can use these t-ratios for testing hypotheses on individual coefficients, for instance that they are zero. In general one can also test any linear (or nonlinear) hypothesis on the cointegrating parameters using a likelihood ratio test, which is asymptotically distributed as $\chi^2(f)$, where f is the number of restrictions tested.

A simple example of maximum likelihood estimation is given in model (22), where the scale factor can be chosen as $\hat{\tau} = \hat{\sigma}_{11.2}^{1/2}(\sum_{t=1}^{T} x_t^2)^{-1/2}$, and the limit is Gaussian because $B_{1|2}$ is independent of B_2 in (23).

4.6 Regression analysis of cointegrating relations

The cointegration coefficients can also be estimated by regression, provided we know the value of r, but inference is difficult in the sense that running a regression of X_{1t}

on $X_{2t}, \ldots, X_{pt}$ will give consistent estimators of the cointegrating coefficients, but the corresponding $t-$ ratios will not converge to the normal distribution, and one cannot find scale factors so that (33) holds. This was illustrated in Example 3 above, where the equations for $\xi = 1$, become

$$\begin{aligned} Y_t &= \beta X_t + \varepsilon_{1t}, \\ \Delta X_t &= \varepsilon_{2t}. \end{aligned}$$

This is an example of two cointegrated series, where the usual t-test leads to a strange limit distribution if $\sigma_{12} \neq 0$, see (21). The problem of how to modify the regression approach by finding a nonparametric estimator of the so-called long-run variance, $C\Omega C'$, was solved by Phillips (1991).

If, however, X_t contains a trend, then the analysis is different because a regression will in fact give valid inference because

$$\sqrt{\sum_{t=1}^{T} X_t^2}(\hat{\beta} - \beta) = \frac{\sum_{t=1}^{T} \varepsilon_{1t} X_t}{\sqrt{\sum_{t=1}^{T} X_t^2}} = \frac{\sum_{t=1}^{T} \varepsilon_{1t}(\sum_{i=1}^{t} \varepsilon_{2i} + \mu t)}{\sqrt{\sum_{t=1}^{T}(\sum_{i=1}^{t} \varepsilon_{2i} + \mu t)^2}} \approx \frac{\mu}{|\mu|} \frac{\sum_{t=1}^{T} \varepsilon_{1t} t}{\sqrt{\sum_{t=1}^{T} t^2}}$$

which converges to $N(0, 1)$. The reason for the change of result is that the trend dominates the random walk asymptotically in this case.

5 An example of a cointegration analysis of sea level and temperature 1881–1995

The data for this analysis consists of annual temperature anomalies from 1881 to 1995 taken from Hansen *et al.* (2001).

Inspection of the data in Figure 4 shows that both variables are clearly nonstationary, but it is difficult to see if they are stationary around a linear trend or if there is a random walk component in the data. The differences, however, look like stationary processes. In order to investigate this we analyse the data using model (31) ($p = r = 2$) and test model (29) for $r = 0, 1$.

We use the notation $X_t = (temperature_t,\ sea\ level_t) = (T_t, h_t)$, $t =$ 1881 to 1995 and fit the model

$$\Delta X_t = \Pi X_{t-1} + \Gamma \Delta X_{t-1} + \mu + \varepsilon_t.$$

We use the unrestricted drift term μ, which creates a linear trend in the processes, because each of the variables show a clear trending behavior. The choice of linearity in trend is of course just a simple description of some of the variables left out. The adequacy of the model is checked by residual analysis in Figures 5 and 6 where it is seen that there is only little autocorrelation in the residuals and no seriously large normalized residuals.

The primary hypothesis of interest is the test for $\Pi = \alpha\beta'$, which, if accepted, would establish cointegration. In Table 1 we summarize the analysis of the rank. The

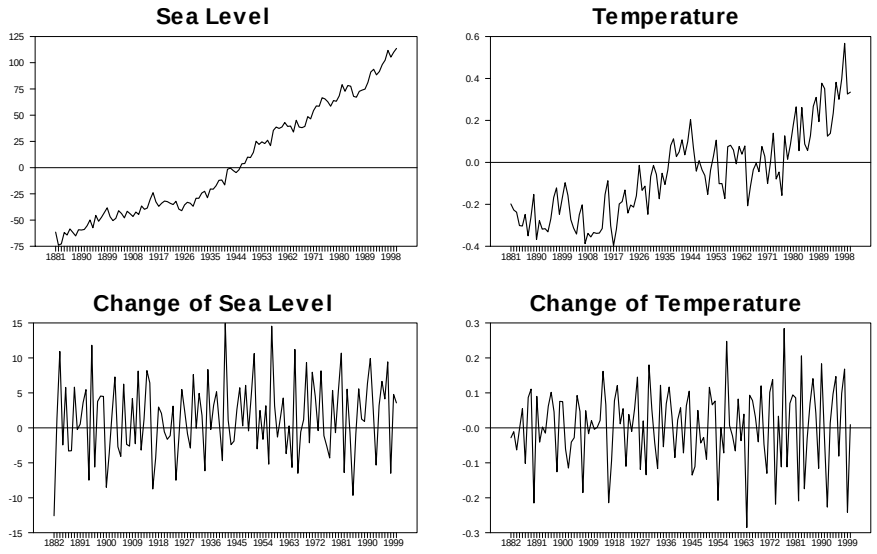

Figure 4: Plot of sea level and temperature in levels and differences. Note the clear nonstationarity in the levels, which could be due to a stochastic trend or possibly a deterministic trend. The differences, however, behave like stationary processes.

Rank determination of $\Pi = \alpha\beta'$ without forcing variables				
$p-r$	r	$-2\log LR(\mathcal{H}_r\|\mathcal{H}_p)$	95%quantile	p-value
2	0	20.76	15.41	0.005
1	1	0.36	3.84	0.540
The reciprocal roots of $\phi(z)=0$ are 1.005, 0.625, -0.187 ± 0.040i				

Table 1: We find that the hypothesis $r = 0$ is rejected and $r = 1$ accepted

findings are that $r = 0$ should be rejected (p-value 0.005), but $r = 1$ is not rejected (p-value 0.54) so that the analysis indicates that the two variables are nonstationary but cointegrate. We have also given the reciprocal roots, and the largest is actually 1.002, so very close to a unit root. We find the estimates of the cointegrating relation

$$U_t = T_t - \underset{(-7.37)}{0.0031} h_t,$$

with $t_{\beta=0}$ in parenthesis, which allows one to evaluate the significance of the coefficients.

We then estimate the remaining parameters of the model

$$\begin{aligned}
\Delta h_t &= \underset{(0.86)}{4.15} U_{t-1} - \underset{(-3.11)}{0.2805} \Delta h_{t-1} + \underset{(0.60)}{3.04} \Delta T_{t-1} + \underset{(3.55)}{2.22} \qquad (\hat{\sigma}_h = 4.939) \\
\Delta T_t &= \underset{(-4.26)}{-0.40} U_{t-1} - \underset{(-1.40)}{0.0024} \Delta h_{t-1} - \underset{(-0.54)}{0.053} \Delta T_{t-1} - \underset{(-1.91)}{0.023} \qquad (\hat{\sigma}_T = 0.095)
\end{aligned}$$

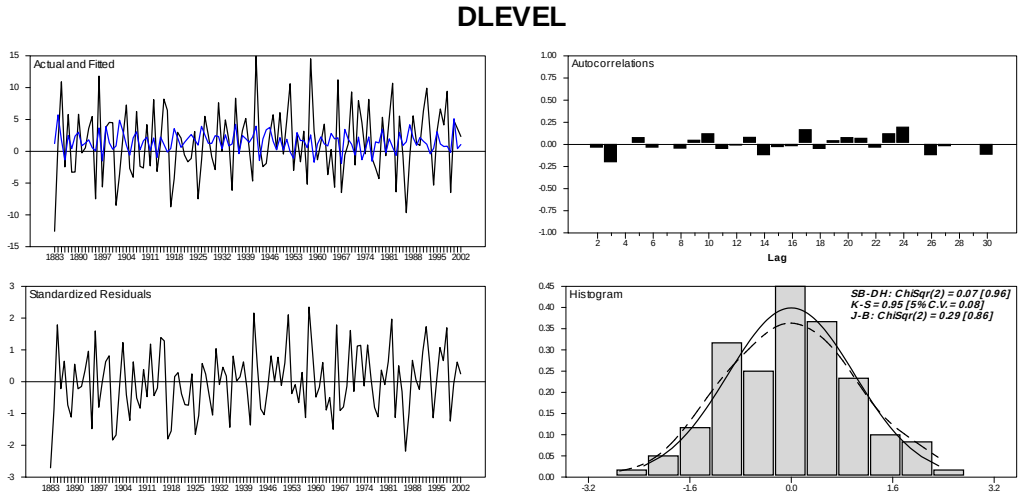

Figure 5: Plot of actual and fitted values of Δh_t, the normalized residuals, their autocorrelation function, and histogram.

Note that h_t almost satisfies the equation

$$\Delta h_t = \underset{(-3.11)}{-0.2805}\Delta h_{t-1} + \underset{(3.55)}{2.22},$$

because the coefficients to the cointegrating relation U_{t-1} is 4.15 ($t = 0.86$) and to the lagged changes ΔT_{t-1} is 3.04 ($t = 0.60$) are insignificant. We can test three overidentifying restrictions by eliminating U_{t-1} in the first equation and ΔT_{t-1} in both, and find

$$\begin{aligned}
\Delta h_t &= \underset{(-2.86)}{-0.25}\,\Delta h_{t-1} + \underset{(3.83)}{1.87}, \\
\Delta T_t &= \underset{(-5.64)}{-0.45}(T_{t-1} - \underset{(-7.37)}{0.0031}h_{t-1}) - \underset{(-1.47)}{0.0025}\Delta h_{t-1} - \underset{(-2.30)}{0.027}, \\
-2\log LR &= 2.55 \approx \chi^2(3), \quad p - value = 0.47.
\end{aligned}$$

The results of this model show that temperature reacts to a disequilibrium between T_t and h_t, as measured by the disequilibrium error $U_{t-1} = T_{t-1} - 0.0031h_{t-1}$, whereas h_t seems to move without being related to temperature.

As a further check of the results we plot the two estimated eigenvectors from the estimation algorithm in Figure 7, and see that the analysis has found two linear combinations where one could be stationary and the other not.

As a final check of the results we plot temperature against sea level, see Figure 8, and there it is apparent that there is something wrong with the model so far analysed. It is as if there is one relation before 1940 and another one after 1960, corresponding to the leveling off of temperature between 1940 and 1960. We conclude that the model

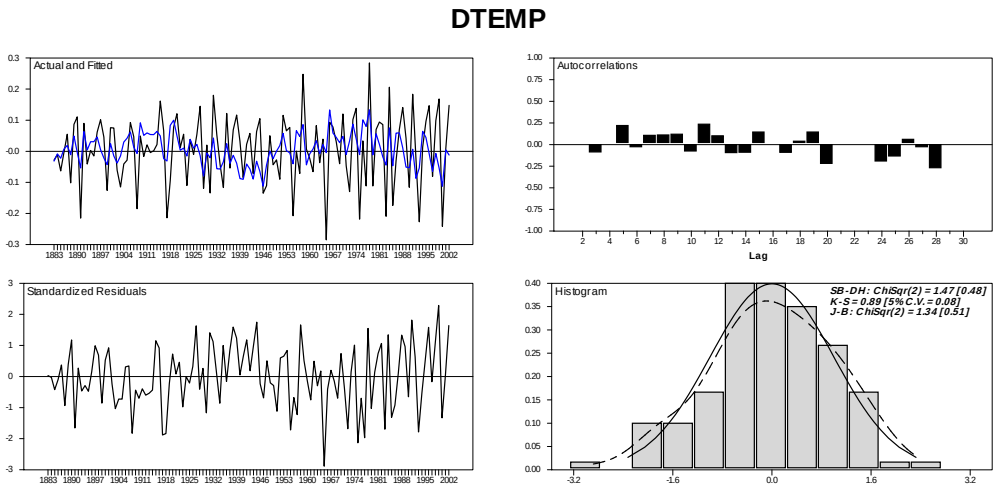

Figure 6: Plot of actual and fitted values of ΔT_t, the normalized residuals, their autocorrelation function and histogram.

is too simple and that it is a better idea, and perhaps even more interesting, to include in the analysis also the radiative forcing variables.

We conclude this first analysis by summarizing the findings.

- A bivariate autoregressive model with 2 lags fits the data quite well.
- There is one cointegrating stationary relation $T_t = 0.0031 h_t$.
- Sea level is not adjusting but temperature is adjusting to the disequilibrium error.
- The plot of temperature versus sea level indicates that the model does not describe the variation satisfactorily.

5.1 The analysis of temperature and sea level including forcing variables

We next improve the analysis by including variables measuring radiative forcing, such as greenhouse gases, ozone, aerosols, volcanic activity, and solar irradiation to see if they can explain better the variation in temperature and sea level. The data is taken from Myhre, Myhre, and Stordal (2001).

We define the variables

$$\begin{aligned} X_{1t} &= (temperature, sea\ level)_t = (T_t, h_t), \\ X_{2t} &= (wmgg(CO_2, NH_4), aerosol(sulphate, soot), sun, vol(dust), ozone)_t \end{aligned}$$

The forcing variables are given in $Watts/m^2$ and are therefore positive for $wmgg$ and negative for $aerosol$.

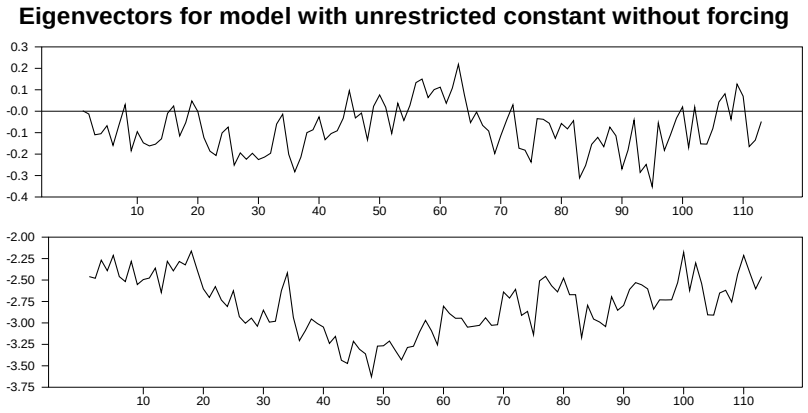

Figure 7: The plot shows the first eigenvector as stationary and the second as nonstationary, which confirms the tests in Table 1

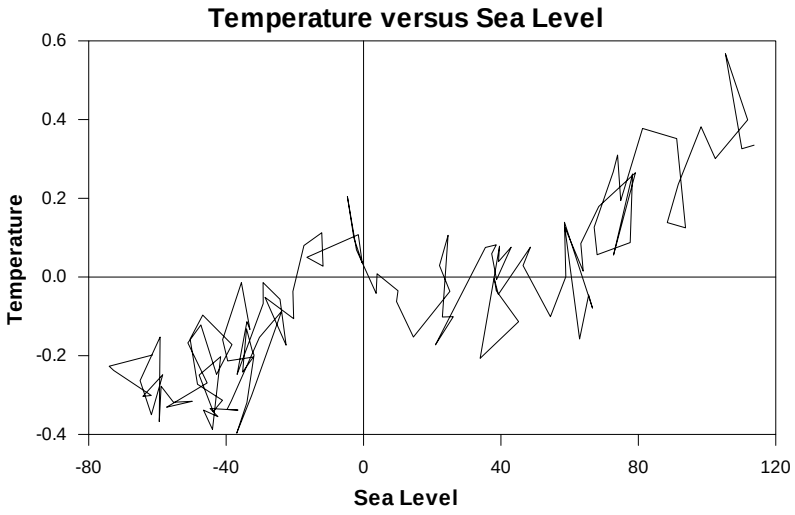

Figure 8: The plot shows that there is something that does not work well in the model. There seems to be a linear relation before 1940 and another after 1960

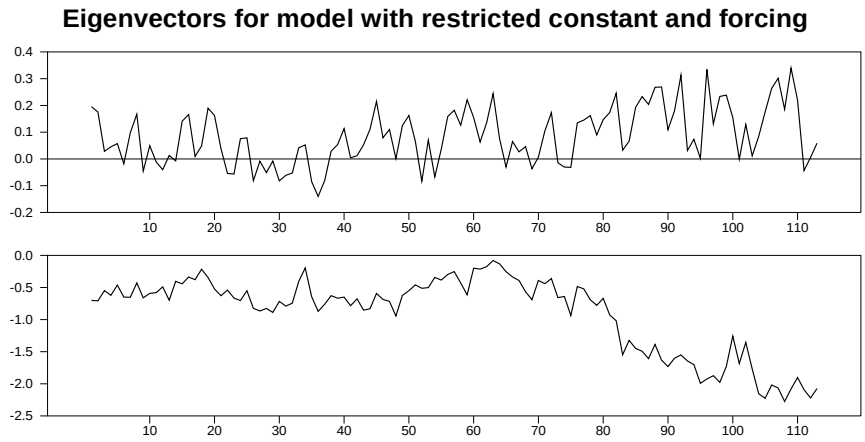

Figure 9: The plot confirms that the first eigenvector is stationary and the second is nonstationary.

We want to analyse X_{1t} conditional on the forcing variables and therefore specify a conditional or partial model, by the equations for ΔX_{1t} conditional on the past of X_{1t} and X_{2t} and the forcing variables ΔX_{2t}. We decompose the parameters as $\mu = (\mu_1', \mu_2')'$ and similarly for $\Gamma = (\Gamma_1', \Gamma_2')'$. Then the conditional model is, see (17) and (18), is

$$\Delta X_{1t} = \alpha_1(\beta_1' X_{1t-1} + \beta_2' X_{2t-1}) + \omega \Delta X_{2t} + \Gamma_1^* \Delta X_{t-1} + \mu_1^* + \varepsilon_{1t}^*,$$

where $\omega = \Omega_{12}\Omega_{22}^{-1}$ and $\Gamma_1^* = \Gamma_1 - \omega\Gamma_2$, $\mu_1^* = \mu_1 - \omega\mu_2$, and $\varepsilon_{1t}^* = \varepsilon_{1t} - \omega\varepsilon_{2t}$ is independent of ε_{2t}, see Johansen (1996, Ch. 8), and equation (22).

The equation for ΔX_{2t} given the past of X_{1t} and X_{2t} is assumed to be

$$\Delta X_{2t} = \Gamma_2 \Delta X_{t-1} + \mu_2 + \varepsilon_{2t}.$$

Thus we assume that there is a cointegrating relation, $\beta_1' X_{1t-1} + \beta_2' X_{2t-1}$, between temperature, sea level, and the forcing variables, and that ΔX_{2t} does not react to the disequilibrium error from this cointegrating relation.

Fitting the unrestricted conditional vector autoregressive model to these data gives much the same fit as the previous model, and the plots corresponding to Figures 5 and 6 have been left out.

5.2 Cointegration analysis of the conditional model

With the forcing variables in the model, the nonstationarity of temperature and sea level can be determined partly by the forcing variables, and partly by the dynamics of the model. If it were completely explained by the forcing variables, then the rank of $\alpha\beta_1'$ would be two and we would find two cointegrating relations.

Rank determination of $\Pi = \alpha\beta'$ with forcing				
$p-r$	r	$-2\log LR(\mathcal{H}_r\|\mathcal{H}_p)$	95%quantile	p-value
2	0	51.26	45.74	0.014
1	1	20.93	23.52	0.100
The reciprocal roots of $\phi(z) = 0$ are 0.731, 0.482, -0.143 ± 0.050i				

Table 2: We find that the hypothesis $r = 0$ is rejected and $r = 1$ accepted

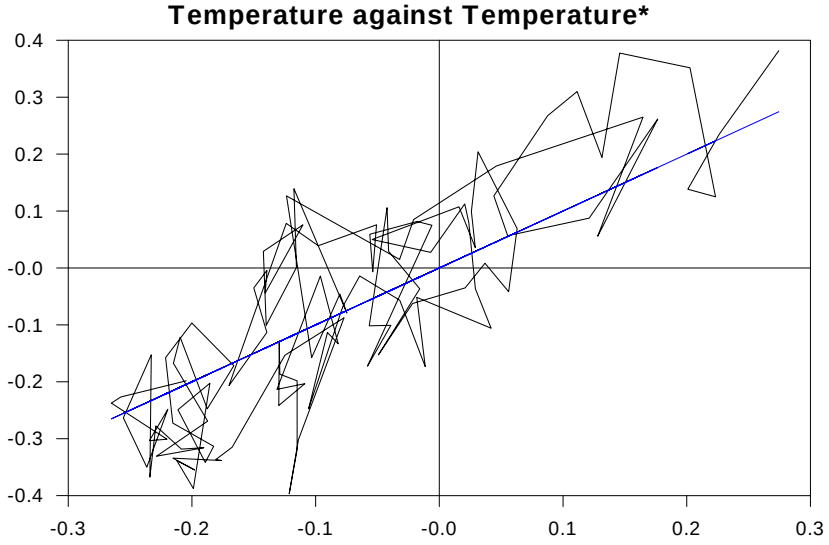

Figure 10: Plot of T_t and the right hand side, T_t^*, of the cointegrating relation solved for T_t. The plot indicates that the two variables move together for the whole period.

Similarly the forcing variables could explain the deterministic trend in temperature and sea level, and we therefore restrict the drift term to be proportional to α, so that the equations do not generate a linear trend.

The rank analysis is given in Table 2, together with the reciprocal roots of $\phi(z) = 0$. The hypothesis of no cointegration is rejected with p-value of 0.01 and rank $r = 1$ is accepted with a $p-$value of 0.10. Note that the largest inverse root in this model is estimated to be 0.731, so the evidence for nonstationarity created by the dynamics is not so clear-cut in the model which conditions on the forcing variables. If we reject $r = 1$, then the rank is two and all nonstationarity in T_t and h_t would be due to the forcing variables, which causes most of the nonstationarity in temperature and sea level.

With $r = 1$, we find the cointegrating relation

$$\begin{aligned}\hat{\beta}' X_t &= T_t - \underset{(-2.02)}{0.0052} h_t - \underset{(-0.75)}{0.573} wmgg_t - \underset{(-1.81)}{1.405} aerosol_t \\ &\quad - \underset{(-0.59)}{0.1750} sun_t - \underset{(-0.11)}{0.0048} vol_t - \underset{(-0.17)}{1.1877} ozone_t + \underset{(0.05)}{0.011}\end{aligned} \tag{34}$$

and the equations

$$\begin{aligned}\Delta h_t &= \underset{(1.0958)}{6.6920}\hat{\beta}'X_{t-1} + \ldots, \qquad \hat{\sigma}_1 = 4.7836,\\ \Delta T_t &= \underset{(-5.180)}{-0.5437}\hat{\beta}'X_{t-1} + \ldots, \qquad \hat{\sigma}_2 = 0.0875,\end{aligned}$$

where we have left out the short term dynamics to save space, that is, the coefficients to ΔX_{2t} and ΔX_{t-1}.

In this data it appears that the contributions in the cointegrating relation (34) from the three variables solar irradiation, volcanic activity, and ozone are very small and we can test by a likelihood ratio that the coefficient are zero. Renewed estimation of this simpler model gives

$$\beta'X_t = T_t - \underset{(-3.66)}{0.0068}h_t - \underset{(-4.70)}{0.775}\, wmgg_t - \underset{(-3.62)}{1.532}aerosol_t - \underset{(-0.30)}{0.041} \tag{35}$$

and the equations

$$\begin{aligned}\Delta h_t &= \underset{(1.58)}{9.4720}\hat{\beta}'X_{t-1} + \ldots, \qquad \hat{\sigma}_1 = 4.690,\\ \Delta T_t &= \underset{(-4.84)}{-0.506}\hat{\beta}'X_{t-1} + \ldots, \qquad \hat{\sigma}_2 = 0.0820,\\ -2\log LR &= 0.221 \approx \chi^2(3)\ p - value = 0.97\end{aligned}$$

A graphical check on the cointegrating relation is found by plotting T_t against the right hand side of (35), see Figure 10, and now the movement around the identify line, looks much more like stationary deviations.

We find again that sea level does not react to the disequilibrium error, whereas temperature does adjust. Note also that the coefficients of the forcing variables are much more significant now, due to the collinearity of the variables, and that the constant term is insignificant.

A nice way of summarizing the findings, see Kaufmann, Kauppi, and Stock (2006), is in Figure 11 where we have plotted T_t, and the components of the cointegrating relation $wmgg^* = 0.775\ wmgg$, $aerosol_t^* = 1.532\ aerosol_t$, $level_t^* = 0.0068\ h_t$, and their sum

$$T_t^* = 0.0068\ h_t + 0.775\ wmgg_t + 1.532\ aerosol_t.$$

It is seen that the contribution from the forcing variables $wmgg$ and $aerosol$ are very large and with opposite sign, so that the measured temperature anomalies, is a delicate balance between the large contribution from heating and cooling.

6 Conclusion

We have contrasted two approaches. The regression or correlation based approach and the model based approach to statistical inference.

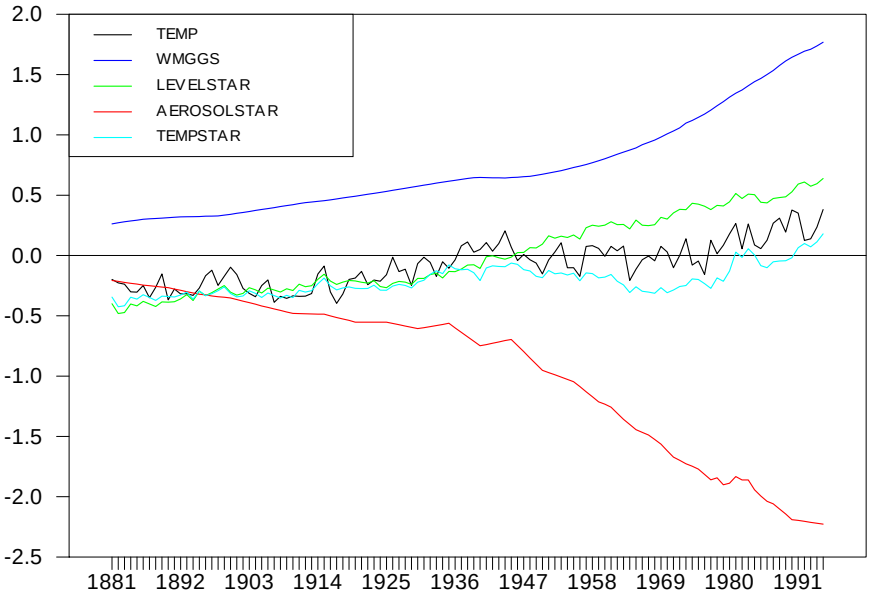

Figure 11: A plot of the effects of the different components in cointegration relation solved for temperature.

It is argued that it is a good idea to distinguish between the *empirical* and the *theoretical* correlation and regression coefficients. We need a limit theorem to relate the empirical value to the theoretical value, and this limit theorem may not hold for nonstationary variables.

We illustrate by example that the empirical coefficients may therefore be spurious, in the sense that the conclusions drawn from them cannot be considered conclusions about the theoretical concepts.

The solution to the spurious correlation or regression problem in practice, is to model the data and check the model carefully before proceeding with the statistical analysis.

Model based analysis of the climate data is consistent with a long-run relation between temperature, sea level and forcing variables. The main effects seem to be from well mixed greenhouse gases and aerosols.

Temperature reacts to a disequilibrium in the long-run relation, but sea level does not. These results are consistent with the notion of the oceans as the main heat reservoir to which temperature reacts through the disequilibrium error.

7 References

Dennis, J., Johansen, S., and K. Juselius, (2005) *CATS for RATS: Manual to Cointegration Analysis of Time Series*, Estima, Illinois.

Dickey D.A., and W.A. Fuller, (1981) Likelihood ratio statistics for autoregressive time series with a unit root. *Econometrica* 49, 1057–1072.

Engle R.F., and C.W.J. Granger, (1987) Co-integration and error correction: Representation, estimation and testing. *Econometrica* 55, 251–276.

Granger, C.W.J. (1981) Some Properties of Time Series Data and Their Use in Econometric Model Specification. *Journal of Econometrics,* 121-130.

Granger, C.W.J., and P., Newbold, (1974) Spurious regressions in econometrics. *Journal of Econometrics* 2, 111–120.

Haavelmo, T., (1943) Statistical implications of a system of simultaneous equations. *Econometrica* 11, 1-12.

Hall, P., and C.C. Heyde (1980) *Martingale Limit Theory and its Application.* Academic Press, New York.

Hansen, J.E., R. Ruedy, Mki. Sato, M. Imhoff, W. Lawrence, D. Easterling, T. Peterson, and T. Karl, (2001) A closer look at United States and global surface temperature change. *J. Geophys. Res.*, 106, 23947-23963.

Johansen, S., (1988) Statistical analysis of cointegration vectors. *Journal of Economic Dynamics and Control* 12, 231–254.

Johansen, S. (1996) *Likelihood-based inference in cointegrated vector autoregressive models.* Oxford University Press, Oxford.

Johansen, S., (2006) Cointegration: a survey. In Mills T.C. and Patterson K. (eds) *Palgrave Handbook of Econometrics: Volume 1, Econometric Theory.* Palgrave Macmillan, Basingstoke 16, 17–34.

Juselius, K. (2006) *The cointegrated VAR model: Econometric methodology and macroeconomic applications.* Oxford University Press, Oxford.

Kaufmann, R.K., Kauppi, H., and J.H. Stock (2006) Emission, concentrations, & temperature: a time series analysis. *Climatic Change* 77, 249–278.

Myhre, G. Myhre, A. and F. Stordal (2001) Historical evolution of radiative forcing of climate. *Atmospheric Environment* 35**,** 2361–2373.

Phillips, P.C.B., (1986) Understanding Spurious Regressions in Econometrics. *Journal of Econometrics* 33, 311-340.

Phillips, P.C.B., (1991) Optimal inference in cointegrated systems. *Econometrica* 59, 283–306.

Sober, E., (2001) Venetian sea levels, British bread prices and the principle of the common cause. *British Journal for the Philosophy of Sciences* 52, 331–346.

von Storch, H., and F.W., Zwiers (2002) *Statistical Analysis in Climate Research.* Cambridge University Press.