CENTRE FOR ECONOMETRIC ANALYSIS
CEA@Cass

http://www.cass.city.ac.uk/cea/index.html

Cass Business School
Faculty of Finance
106 Bunhill Row
London EC1Y 8TZ

Common Stochastic Trends and Aggregation in Heterogenous Panels

Stepana Lazarova, Lorenzo Trapani and Giovanni Urga

CEA@Cass Working Paper Series

WP–CEA–01-2004

Common stochastic trends and aggregation in heterogeneous panels

Stepana Lazarova
London School of Economics

Lorenzo Trapani
Cass Business School and Bergamo University

Giovanni Urga*
Cass Business School

This version: January 2004

*Corresponding author: Cass Business School, Faculty of Finance, 106 Bunhill Row, London EC1Y 8TZ Tel.+/44/(0)20/70408698; Fax.+/44/(0)20/70408885; e-mail: g.urga@city.ac.uk. http://www.cass.city.ac.uk/faculty/g.urga. We wish to thank Stephen Hall for inspiring discussions and comments. We are grateful to participants at the 9th Conference on Panel Data (Geneva, 22-23 June, 2000), SED 2000 (San Jose', Costa Rica, 29 June-2 July 2, 2000) and various colleagues at Bergamo University (Italy), London Business School, Cass Business School for useful comments. We have enormously benefited from comments and suggestions from Cheng Hsiao. The usual disclaimer applies. We acknowledge ESRC (Project N. R022251032) for financial support. Lorenzo Trapani wishes to thank Marie Curie Training Site for financial support (grant N. HPMT-CT-2001-00330).

Common stochastic trends and aggregation in heterogeneous panels

Abstract

In nonstationary heterogeneous panels where each unit cointegrates, a large number of conditions needs to be satisfied in order for cointegration to be preserved in the aggregate relationship. In reality, the conditions most likely will not hold. This paper takes a closer look at what happens when the conditions are violated. Of particular interest is the question whether a mild violation can still lead to an aggregate estimator that summarises the macro relationship reasonably well. We derive the asymptotic measure of the degree of non cointegration of the aggregated estimate and we provide an estimation and testing procedure. A Monte Carlo exercise evaluates the small sample properties of the estimator.

J.E.L. Classification Numbers: C12, C13, C23
Keywords: Aggregation, Cointegration, Heterogeneous Panel, Monte Carlo Simulation.

1 INTRODUCTION

The issue of aggregation has been of considerable interest in the econometric literature. Many macroeconomic theories are based on the behaviour of individual agents, households or firms, but often only aggregate data are available to test them. The question then arises of how well the aggregate relationship approximates the properties of the individual components. Conversely, the data may be given at a disaggregated level. The micro relationships can then be summarised in many ways, the simple aggregation of the components being just one way to pursue.

When the variables in the economic system are integrated, an important observation is that the cointegration on the micro level does not automatically mean cointegration on the macro level (see e.g. Pesaran and Smith (1995)). If the cointegration does not carry through the aggregation process, the macro estimates are not consistent, making the information provided by macro summary meaningless. It was long thought that satisfying the representative agent assumption is the only way to preserve cointegration. Gonzalo (1993) shows that this conjecture is not correct and that the agents need not be homogeneous. When the micro cointegrating vectors are heterogeneous across units, the cointegration can still be preserved if there is enough cointegration among the variables in the economic system. Sufficient amount of cointegration is present when the series in the system are driven by sufficiently low number of common stochastic trends.

Several authors elaborate on this point. General conditions for cointegration amongst the aggregates to occur are presented in Granger (1993). Gonzalo (1993) derives some specific sufficient conditions under which cointegration at the micro level implies cointegration at the macro level. To address the previously held belief about *sine qua non* homogeneity of agents, he shows in which circumstances the equality of micro cointegrating vectors is necessary and sufficient for preserving cointegration at macro level. He also considers the opposite direction, rarely discussed in the aggregation literature, giving conditions under which cointegration at macro level implies cointegration at the micro level. Ghose (1995) presents conditions under which a subset of regressors in the model can be aggregated without damaging the consistency of the estimate of the parameters of interest.

The conditions for preserving cointegration in the process of aggregation are arguably very strong. Observing cointegration in aggregate macro variables implies validity of a considerable and unlikely amount of constraints on the micro system. In addition, the number of constraints grows in quadratic proportion with the number of economic units in the system. It is therefore

questionable whether these conditions are of any empirical relevance.

Granger (1993) considers a case in which only few common stochastic trends are shared across virtually all of the original series of the model. The remaining trends are shared by only small groups of the series. In such a case, the coefficients of the shared common trends in the aggregate regression are larger than the coefficients for the idiosyncratic common trends by an order of magnitude. Removing the large trends from the aggregate regression leaves only "small" $I(1)$ elements in the residuals that may not be found by standard tests applied to relatively small samples. In this case, the system is sufficiently well described by a small number of dominant components and the aggregate relationship "approximately cointegrates".

In this paper we intend to bring further insights on the aggregation conditions for nonstationary heterogenous panels. Our standpoint is that in real economic systems the tight set of aggregation conditions is indeed unlikely to be satisfied. We believe, however, that the aggregate relationship does not become entirely meaningless when the conditions get "mildly violated", in that though the panel equation might not satisfy the formal condition for cointegration, the aggregate data may only have "small" non-stationary components and this makes the (strictly speaking spurious) macro relationship observationally equivalent to a cointegration equation. We argue that when the aggregate relationship "approximately cointegrates" in the sense stated above, then it should be treated as if it were actually a cointegrating relationship. Indeed this is quite natural in the light of Granger's (1993) argument. Thus, we derive a measure of the degree of non-cointegration of the aggregate estimate and we propose an estimation and testing procedure.

The remainder of the paper is organised as follows. The theoretical framework is presented in Section 2, where we set up a model of a heterogenous panel. We first present the asymptotic properties of the aggregate estimate. To characterise the effect of the presence of excess stochastic trends we concentrate on a simple case of a panel driven by two common stochastic trends when only one trend ought to be present. Further, we discern the factors determining the behaviour of the aggregate estimate. In Section 3, we propose a procedure that allows one to test for the degree of non-cointegration, we derive an estimation framework and a testing procedure. Finally, a numerical example illustrates how to implement the procedure. Moreover, as the results given in Section 2 for the aggregate estimation are asymptotic, in Section 4 we present a Monte-Carlo examination of a small-sample distribution of the aggregate estimate. Section 5 concludes.

2 THE THEORETICAL FRAMEWORK

Let us consider a simple system with one explanatory variable:[1]:

$$y_{it} = \beta_i x_{it} + u_{it}, \quad t = 1, 2, \ldots, T.$$

The explanatory variables are I(1) processes that share k common trends:

$$x_{it} = \alpha_{i1} z_{1t} + \alpha_{i2} z_{2t} + \ldots + \alpha_{ik} z_{kt} + v_{it}, \quad i = 1, 2, \ldots, n,$$

where

$$z_{jt} = z_{jt-1} + \epsilon_{jt}.$$

In matrix form the system can be rewritten as

$$y_t = Bx_t + u_t, \tag{1}$$
$$x_t = Az_t + v_t, \tag{2}$$
$$z_t = z_{t-1} + \epsilon_t. \tag{3}$$

with $y_t = [y_{1t}, ..., y_{nt}]'$ and x_t, z_t defined similarly, $B \equiv diag\{\beta_1, ...\beta_n\}$, $A \equiv [\alpha_1, ...\alpha_n]'$ and $\alpha_i \equiv [\alpha_{i1}, ..., \alpha_{ik}]$, and with u_t, v_t, ϵ_t vectors of disturbances. The coefficients in the model are fixed. The matrix A is assumed to be of rank k so that x_t is indeed driven by no less than k stochastic trends. The trends vector starts at $z_0 = 0$. We define the vector $\varepsilon_t = (u_t', v_t', \epsilon_t')'$ and the vector of partial sums $S_t = \sum_{i=1}^{t} \varepsilon_i$ and set $S_0 = 0$. We assume the vector ε_t to be an invertible $MA(\infty)$ process satisfying $\varepsilon_t = \Phi(L)\eta_t = \sum_{j=0}^{\infty} \Phi_j \eta_{t-j}$, with η_t uncorrelated and with unit variance. Moreover, the sequence of innovations satisfies the following assumptions:

Assumption 1:

1a) $E(\varepsilon_t) = 0$ for all t;

1b) $\sup_{i,t} E|\varepsilon_{it}|^{\zeta+\delta} < \infty$ for some $\zeta > 2$ and $\delta > 0$;

1c) $\Sigma = \lim_{T\to\infty} T^{-1} E(S_T S_T')$ exists[2];

1d) $\{\varepsilon_t\}_1^{\infty}$ is strong mixing with mixing numbers α_m satisfying $\sum_1^{\infty} \alpha_m^{1-2/\zeta} < \infty$;

[1] We use this very simple case to easy our presentation but the model can be generalised to a full multivariate case of p regressors (x_{hit} with *h=1,...,p*) with (2) valid for all p-variables driven by the k stochastic trends.

[2] Notice that this assumption usually also requires Σ to be positive definite - see Phillips (1986) for technical details. However, this is not required for model (1)-(3) and doesn't make Proposition 1 invalid, as proved in Appendix.

1e) the components of ϵ_t are independent and the trends z_t have unit long-run variance so that the lower diagonal $k \times k$ block of the matrix Σ is an identity matrix, $\text{var}_{LR} z = \lim_{T\to\infty} T^{-1} E(z_t z_t') = I_k$.

Assumptions 1a)-1d) are needed for the central limit theorem for the functional spaces to be valid. The assumption of orthonormality 1e) makes the trends z_{it} neutral in the model so that the behavior of the system is fully described by the coefficients β_i and A.

When we aggregate the regressors across the units, we obtain

$$\overline{x}_t = a_1 z_{1t} + a_2 z_{2t} + \ldots + a_k z_{kt} + \overline{v}_t,$$

where $\overline{x}_t \equiv \sum_{i=1}^n x_{it}$, $a_i \equiv \sum_{j=1}^n \alpha_{ji}$ and $\overline{v}_t \equiv \sum_{i=1}^n v_{it}$. We assume there is at least one i for which $a_i \neq 0$, so that $\overline{x}_t$ is $I(1)$. For the dependent variable we have

$$y_{it} = \beta_i \alpha_{i1} z_{1t} + \beta_i \alpha_{i2} z_{2t} + \ldots + \beta_i \alpha_{ik} z_{kt} + \beta_i v_{it} + \overline{u}_{it}, \quad i = 1, \ldots, n,$$

so the aggregate variable is of the following form:

$$\overline{y}_t = b_1 z_{1t} + b_2 z_{2t} + \ldots + b_k z_{kt} + \overline{u}_t,$$

where $\overline{y}_t \equiv \sum_{i=1}^n y_{it}$, $b_i \equiv \sum_{j=1}^n \beta_j \alpha_{ji}$ and $\overline{u}_t \equiv \sum_{i=1}^n \beta_i v_{it} + \sum_{i=1}^n u_{it}$. We again assume there is at least one i for which $b_i \neq 0$, so that $\overline{y}_t$ contains a unit root.

We consider the least-squares regression equation

$$\overline{y}_t = \widehat{\beta} \overline{x}_t + \widehat{e}_t,$$

where $\widehat{\beta}$ is the conventional least-squares regression coefficient. When $\overline{y}_t$ and $\overline{x}_t$ are cointegrated, the estimator $\widehat{\beta}$ is superconsistent for finite n and large T and converges in probability to the true value of the aggregation coefficient. In case the aggregate series are not cointegrated, the regression is spurious and $\widehat{\beta}$ converges in distribution to a certain non-degenerate random variable. The following proposition characterises the limiting distribution of the $\widehat{\beta}$ estimator for large T and finite n.

Proposition 1 *If y_t and x_t are generated by (1)-(3) where the innovation sequence $\{\varepsilon_t\}_1^\infty$ satisfies Assumption 1, then in the OLS regression of $\overline{y}_t$ on $\overline{x}_t$*

$$\widehat{\beta} \Rightarrow S = \frac{\sum_{i=1}^k \sum_{j=1}^k b_i a_j W_{ij}}{\sum_{i=1}^k \sum_{j=1}^k a_i a_j W_{ij}}, \tag{4}$$

where $W_{ij} = \int_0^1 W_i(r)W_j(r)dr$, W_i *and* W_j *are independent standard Wiener processes and where "*$\Rightarrow$*" denotes weak convergence of the associated probability measures as* $T \to \infty$.[3]

The case of a large number of units leads to a different interpretation of the OLS estimator $\hat{\beta}$. To show this, we may assume the regression coefficients α_{ij} to be *i.i.d.* random variables across i with mean $\bar{\alpha}_j$ and the β_is to be *i.i.d.* random variables across i as well, with mean $\bar{\beta}$. Notice that $\bar{\beta}$ is the average of the individual cointegrating relations. We also assume the β_is and α_{ij}s to be incorrelated. By the weak law of large numbers

$$\frac{1}{n}\sum_{i=1}^{n}\alpha_{ij} \xrightarrow{p} \bar{\alpha}_j$$

$$\frac{1}{n}\sum_{i=1}^{n}\beta_i \xrightarrow{p} \bar{\beta}$$

and also, by uncorrelatedness

$$\frac{1}{n}\sum_{i=1}^{n}\alpha_{ij}\beta_i \xrightarrow{p} \bar{\alpha}_j\bar{\beta}$$

where "$\xrightarrow{p}$" denotes covergence in probability; all results hold for any j. Then, as $T \to \infty$ and then $n \to \infty$ according to Phillips and Moon's (1999) definition of sequential limit we obtain

$$\begin{aligned}
\hat{\beta} \;\Rightarrow\; & \left[\sum_{j=1}^{k}\sum_{l=1}^{k}\left(\frac{1}{n}\sum_{i=1}^{n}\alpha_{ij}\right)\left(\frac{1}{n}\sum_{i=1}^{n}\alpha_{il}\right)W_{jl}\right]^{-1} \\
& \times\left[\sum_{j=1}^{k}\sum_{l=1}^{k}\left(\frac{1}{n}\sum_{i=1}^{n}\alpha_{ij}\beta_i\right)\left(\frac{1}{n}\sum_{i=1}^{n}\alpha_{il}\right)W_{jl}\right] \\
\to\; & \left(\sum_{j=1}^{k}\sum_{l=1}^{k}\bar{\alpha}_j\bar{\alpha}_l W_{jl}\right)^{-1}\bar{\beta}\left(\sum_{j=1}^{k}\sum_{l=1}^{k}\bar{\alpha}_j\bar{\alpha}_l W_{jl}\right) = \bar{\beta} \qquad (5)
\end{aligned}$$

According to (5), for large n, the OLS estimate $\hat{\beta}$ picks up the average relation between $\bar{y}_t$ and $\bar{x}_t$, regardless of the existence of the cointegrating relation between the aggregated variables. Our analysis, instead, focuses on the case of finite n to derive the degree of non-cointegration of the aggregate estimate[4].

[3]The proof of this limiting representation can be found in Park and Phillips (1988, 1989).

[4]The analysis of common stochastic trends in the presence of large n is natural within a principal component framework as in Bai (2002) who derives an inferencial theory for generalized dynamic factor models proposed by Forni *et al* (2000).

2.1 Perfect cointegration versus spurious regression

Using Proposition 1, the properties of the estimator $\widehat{\beta}$ in the presence of one or more common stochastic trends can be inferred. If the x_is are driven by a single stochastic trend, $k = 1$, the limiting distribution of $\widehat{\beta}$ is

$$S = \frac{b_1 a_1 W_{11}}{a_1 a_1 W_{11}} = \frac{b_1}{a_1} = \frac{\sum_{j=1}^{n} \beta_j \alpha_{j1}}{\sum_{j=1}^{n} \alpha_{j1}}. \tag{6}$$

In the presence of a single trend, therefore, the aggregate relationship cointegrates. The OLS estimator $\widehat{\beta}$ converges to a weighted average of β_j coefficients where weights are given by the α_j coefficients. This finding is consistent with the analysis of Gonzalo (1993). Hall, Lazarova and Urga (1999) highlight this case when providing a counterexample to the general statement of Pesaran and Smith (1995) that the aggregate relationship does not cointegrate even if the individual unit do cointegrate.

When there is more than one common stochastic trend present among the right-hand side variables, the aggregate regression is generally spurious and the estimator $\widehat{\beta}$ converges in distribution to a non-degenerate random variable. Cointegration occurs only if there exists a constant c such that

$$b_i a_j = c a_i a_j \quad \text{for every } i,\, j,$$

that is

$$a_j (b_i - c a_i) = 0 \quad \text{for every } i,\, j.$$

This condition is satisfied if and only if either

$$a_j = \sum_{i=1}^{n} \alpha_{ij} = 0 \quad \text{for every } j$$

or

$$b_i - c a_i = \sum_{j=1}^{n} (\beta_j - c) \alpha_{ji} = 0 \quad \text{for every } i.$$

In the first case the aggregate series $\overline{x}_t$ will not have a unit root. We exclude this situation by assumption. In the other case, either the coefficients β must be homogeneous or the vector $(\beta_1, \beta_2, \ldots, \beta_n) - c(1, 1, \ldots, 1)$ needs to be orthogonal to each of the columns of the matrix A, $\alpha_i = (\alpha_{1i}, \alpha_{2i}, \ldots, \alpha_{ni})$, $i = 1, 2, \ldots, k$. The latter case can occur only if the vector β lies in the space spanned by the columns of matrix A together with the vector $(1, 1, \ldots, 1)$, which is always true if $k = 1$ but grows ever less likely as k increases. The homogeneity of cointegrating vectors will naturally always lead to cointegrating aggregate relationship for any number of common stochastic trends.

If there is more than one trend driving the regressors and none of the above conditions is satisfied, the aggregate relationship does not cointegrate. This qualifies the regression as spurious. However, the spread of the limiting variable depends continuously on the parameters of the underlying processes, so for some values of parameters the distribution can be nearly degenerate. In such a border case, the estimator $\widehat{\beta}$, though not consistent, may not be entirely worthless. It is therefore of some interest to analyse on which factors the variance of the limiting distribution depends.

In order to illustrate this point, we consider the simplest but already effective case when there are only two common stochastic trends and the conditions for cointegration are not satisfied[5]. It is possible that if there is a single common stochastic trend and we "contaminate" the sample with an "unimportant" second trend, the limiting distribution of $\widehat{\beta}$ will not depart far from the case of perfect cointegration. Therefore we need to establish what are the properties that classify a trend as important.

If at one extreme the second column of matrix A was a zero vector, then the aggregate regression would cointegrate. From the expression (6) we can see that the limiting distribution of the $\widehat{\beta}$ estimator would degenerate to a constant $S_1 = \frac{b_1}{a_1}$. At the other extreme, if the first column of matrix A was a zero vector, then the limiting distribution of $\widehat{\beta}$ would be a constant $S_2 = \frac{b_2}{a_2}$. For the cases in between, when matrix A has rank 2, the limiting estimator S is a random variable whose distance from S_1 is

$$\begin{aligned} S - S_1 &= \frac{b_1a_1W_{11} + b_1a_2W_{12} + b_2a_1W_{21} + b_2a_2W_{22}}{a_1a_1W_{11} + a_1a_2W_{12} + a_2a_1W_{21} + a_2a_2W_{22}} - \frac{b_1}{a_1} = \\ &= \frac{(b_2a_1 - b_1a_2)W_{12} + \left(b_2a_2 - \frac{b_1}{a_1}a_2^2\right)W_{22}}{a_1^2W_{11} + 2a_1a_2W_{12} + a_2^2W_{22}} = \\ &= \frac{\left(\frac{b_2}{a_2} - \frac{b_1}{a_1}\right)W_{12} + \frac{a_2}{a_1}\left(\frac{b_2}{a_2} - \frac{b_1}{a_1}\right)W_{22}}{\frac{a_1}{a_2}W_{11} + 2W_{12} + \frac{a_2}{a_1}W_{22}} = \\ &= \left(\frac{b_2}{a_2} - \frac{b_1}{a_1}\right)\frac{W_{12} + kW_{22}}{\frac{1}{k}W_{11} + 2W_{12} + kW_{22}}, \end{aligned}$$

where $k = \frac{a_2}{a_1}$, so that

$$S - S_1 = (S_2 - S_1)\, f(k), \tag{7}$$

where

$$f(k) = \frac{W_{12} + kW_{22}}{\frac{1}{k}W_{11} + 2W_{12} + kW_{22}}. \tag{8}$$

[5]The generalisation to $k > 2$ will make the presentation extremely cumbersome and it will only add more to the degree of non-cointegration affecting the decomposition of S.

The random variable $f(k)$ does not have a standard distribution, nevertheless some useful facts can be easily verified. First, as $k \to 0$ or $k \to \pm\infty$, it collapses to a degenerated random variable with mass of probability at points 0 or 1, respectively. These cases correspond to the situation where $a_1 = 0$ or $a_2 = 0$, when one of the trends does not play any role in the aggregate relationship and when the regression cointegrates. Further,

$$f(k) \stackrel{D}{=} 1 - f(\tfrac{1}{k}),$$
$$f(k) \stackrel{D}{=} f(-k),$$

where "$\stackrel{D}{=}$" signifies equality in distribution. The distribution of $f(k)$ for some values of k is illustrated in Fig. 1. The value of mean and variance for the distribution of $f(k)$ is depicted in Fig. 2.

[Insert somewhere here Figures 1 and 2]

From the simulations it can be inferred that

1. $Ef(k)$ always lies between 0 and 1, $0 \leq Ef(k) \leq 1$ for all k,
2. $Var\ f(k)$ attains its maximum at $k = \pm 1$,
3. $Var\ f(1) = 0.222$,
4. $Ef(1) = 0.5$.

Equation (7) can be written as

$$S = (1 - f(k)) \cdot S_1 + f(k) \cdot S_2$$

and the distribution of S can be interpreted as a linear combination of the two extreme estimates S_1 and S_2. As $0 \leq Ef(k) \leq 1$ for all k, the mean of distribution of S,

$$ES = (1 - Ef(k)) \cdot S_1 + Ef(k) \cdot S_2,$$

is always a convex combination of the estimates S_1 and S_2.

To obtain a further understanding of what determines the form of the limiting distribution S, we express the term $S_2 - S_1$ in equation (7) as

$$\begin{aligned} S_2 - S_1 &= \frac{b_2}{a_2} - \frac{b_1}{a_1} = \frac{\sum_{j=1}^n \beta_j \alpha_{j2}}{\sum_{j=1}^n \alpha_{j2}} - \frac{\sum_{j=1}^n \beta_j \alpha_{j1}}{\sum_{j=1}^n \alpha_{j1}} = \\ &= \sum_{j=1}^n \beta_j \left(\frac{\alpha_{j2}}{\sum_{j=1}^n \alpha_{j2}} - \frac{\alpha_{j1}}{\sum_{j=1}^n \alpha_{j1}} \right) = \sum_{j=1}^n \beta_j h_j = \tau' h = \\ &= \|\tau\| \, \|h\| \cos(\tau, h) \end{aligned}$$

where

$$h_j = \left(\frac{\alpha_{j2}}{\sum_{j=1}^{n} \alpha_{j2}} - \frac{\alpha_{j1}}{\sum_{j=1}^{n} \alpha_{j1}} \right),$$

and $\tau \equiv (\beta_1, \ldots, \beta_n)'$, $h = (h_1, \ldots, h_n)'$, where $\|\cdot\|$ is the Euclidean vector norm, and $\cos(\tau, h)$ is the cosine of the angle between vectors τ and h. Thus we have

$$S = \frac{b_1}{a_1} + \|\tau\| \, \|h\| \cos(\tau, h) f(k). \tag{9}$$

Given that $\frac{b_1}{a_1} = S_1$ is a fixed constant, the variance of S depends on four factors:

1. $\|\tau\|$: The standard deviation of the distribution of S increases linearly with the size of $(\beta_1, ..., \beta_n)$ coefficients. This observation is intuitive.

2. $\|h\|$: The graphical interpretation of the vector h and its norm for a two-dimensional case (corresponding to two units case, $n = 2$) is given in Fig. 3. Vectors α_1 and α_2 divided by the sum of their components will lie on the manifold $y_1 + y_2 + \ldots + y_n = 1$, which in the two-dimensional case is a line connecting the points (1,0) and (0,1). The difference of the two normalised vectors lies in the hyperplane perpendicular to the $n \times 1$ vector $i_n \equiv (1, 1, \ldots, 1)'$. In our case h lies on the line perpendicular to the axis of the first quadrant. We can see that the length of the vector h is small when the vectors α_1 and α_2 lie in the a direction close to the direction of the vector e and in close directions to each other. If the vector α_i lies close to the vector i_n, it means that units react nearly homogeneously on a given shock. If the angle between the vectors α_1 and α_2 is small, the responses of each unit to the first and the second trend are in similar proportion. Thus we can see $\|h\|$ as a measure of heterogeneity of the reaction of units to the trends, both across trends and across units.

3. $\cos(\tau, h)$: The angle between τ and h does not have a direct economic interpretation. However, the vector h is perpendicular to the vector i_n. Thus if τ is close to the vector i_n, $\cos(\tau, h)$ is small. Therefore, if the coefficients τ are homogeneous to a large degree, the spread of S is relatively small.

4. $k = \frac{a_2}{a_1}$: As $\sum_{j=1}^{n} \alpha_{ji} = a_i$, vectors α_i lie on the manifold $y_1 + y_2 + \ldots + y_n = a_i$, i.e. on a hyperplane cutting each coordinate axis in distance of a_i from the origin. The two-dimensional case is again illustrated in Fig. 3. The bigger is the sum of the components of α_i, the further is the

relevant hyperplane from the origin. We can regard the numbers a_1 and a_2 as a characteristics of the size of the impact of the individual trends. For given lengths of the vector α_i, the impact of the i-th trend is the bigger the more homogeneous is the response of units to that trend. The ratio $\frac{a_2}{a_1}$ can therefore be regarded as a measure of the relative importance of the trends for the aggregate regressor. In the extreme case of α_i being entirely heterogeneous, i.e. being perpendicular to the vector i_n, $a_i = 0$ and the i-th trend has no impact on the aggregate variable $\overline{x}$. As the variance of the random variable $f(k)$ is largest for $k = \pm 1$ and smallest for $k \to \pm\infty$ and $k = 0$, the variance of S is smaller when the relative importance of one of the trends is smaller, and is biggest when the two trends are of equal importance.

[Insert somewhere here Figure 3.]

The above interpretation of the factors determining variance of the distribution of S is intuitively appealing. An important observation is that the values of $\|h\|$ and k are not invariant under a change of stochastic trends z_1 and z_2. If a different pair of orthonormal stochastic trends is chosen as a basis, $\|\tau\|$ and size of $\cos(\tau, h)$ remain unchanged while the two measures characterizing the reaction of the units to the trends, $\|h\|$ and k, change in mutually opposite directions. Thus the latter two factors are not independent.

The analysis of the decomposition (9) gives us the answer to the question of what makes a trend important. It can be seen that the impact of a presence of a superfluous stochastic trend on a system depends both on the relative size and heterogeneity of response of units to this trend, as compared to the other trend. Therefore we can view a trend as unimportant when the size of the response to this trend relative to the other trend is small or if the response is rather homogeneous, or both. Though we conduct our analysis only for the simple case of two common stochastic trends, it gives us an insight into what happens in the general case of more than one excess trend. It follows that even in the presence of multiple common stochastic trends the system can still approximately cointegrate under favourable circumstances.

The observation about the ambiguity of the two parameters $\|h\|$ and k can however be viewed as a potential drawback to the interpretation of the decomposition (9). The interdependence of the two factors means that none of them individually can fully quantify the heterogeneity of response. In other words, knowing the value of one of them is worthless without knowing the value of the other. This setback can be overcome by using the fact that the decomposition (9) is valid for any two orthonormal stochastic trends.

The trends $z = (z_1, z_2)'$ in the DGP are orthonormal in the long-run sense, $Var_{LR}\ z = \lim_{T\to\infty} T^{-1}E(z_T z_T') = I_k$. If we multiply z by any 2×2 matrix M such that $MM' = I_k$ then

$$Var_{LR}\ Mz = Var_{LR}\ \widetilde{z} = M \cdot Var_{LR}\ z \cdot M' = MM' = I_k,$$

and the transformed trends will be also orthonormal. Then

$$x_t = A'z_t + v_t = A'M'Mz_t + v_t = \widetilde{A}'\widetilde{z}_t + v_t,$$

so the matrix of coefficients of x in the transformed basis is

$$\widetilde{A}' = A'M'.$$

The decomposition of S is valid for any two orthonormal stochastic trends, therefore we can write

$$S = \frac{\widetilde{b}_1}{\widetilde{a}_1} + \|\tau\| \left\|\widetilde{h}\right\| \cos(\tau, \widetilde{h}) f(\widetilde{k}),$$

where the meaning of the variables with tilde corresponds to the their meaning in the original basis.

One of the matrices satisfying the condition $MM' = I_k$ is the rotation matrix

$$M_0 = \begin{pmatrix} \cos\phi & -\sin\phi \\ \sin\phi & \cos\phi \end{pmatrix}, \tag{10}$$

where ϕ is any real number. If we select an angle ϕ such that

$$\tan\phi = \frac{a_1 - a_2}{a_1 + a_2} = \frac{1-k}{1+k},$$

the transformed ratio will be

$$\frac{\widetilde{a}_2}{\widetilde{a}_1} = \frac{a_1 \sin\phi + a_2 \cos\phi}{a_1 \cos\phi - a_2 \sin\phi} = \frac{\tan\phi + k}{1 - k\tan\phi} = \frac{\frac{1-k}{1+k} + k}{1 - k\frac{1-k}{1+k}} = 1$$

We can write the decomposition of S with respect to the trends transformed by M_0 as

$$S = \frac{\widetilde{b}_1}{\widetilde{a}_1} + \|\tau\| \, \|h_0\| \cos(\tau, h_0) f(1), \tag{11}$$

where h_0 is the corresponding transform of the original vector h.

The decomposition (11) solves the ambiguity of the decomposition (9). As the random variable $f(k)$ attains maximum variance at $k = 1$, the norm

of vector h_0 is minimal among all possible transforms of the original vector h. Thus the value of $\|h_0\|$ alone fully quantifies the heterogeneity of responses of independent variables to the stochastic trends driving the economy. Another advantage of the form (11) is that to know the distribution of S for any parameter setting, one needs to know the distribution of $f(k)$ at a single point only, $k = 1$.

Using the decomposition (11) we can assess the degree of departure of the system from the case of the perfect aggregation. If we choose the spread of the distribution of S as a measure of this departure, we have

$$Var\ S = \|\tau\|^2 \|h_0\|^2 \cos^2(\tau, h_0)\ Var\ f(1). \tag{12}$$

The variance of the distribution is not invariant to scaling of the variables in the aggregate regression. Therefore we normalise the variance by the squared norm of $\|\tau\|$. Moreover, we normalise the variance by the value of var $f(1)$ since this value is constant for any parameter setting. Finally, to obtain a linear measure of spread, we take the square root of the normalised variance. The proposed measure of the non-cointegration is then of the following form:

$$D(S) = \frac{1}{\|\tau\|} \left(\frac{Var\ S}{Var\ f(1)} \right)^{\frac{1}{2}} = \frac{\tau' h_0}{\|\tau\|} = \|h_0\| \cos(\tau, h_0). \tag{13}$$

The value of $D(S)$ ranges from minus to plus infinity and is zero under the case of perfect cointegration, so that the higher departure from cointegration, the higher $|D(S)|$. The departure from the cointegration depends linearly on the measure of heterogeneity of response and on the angle between vectors τ and h_0 [6]. To violate the conditions for cointegration, both parameters must be different from zero. Therefore in the neighbourhood of zero it is the product of the two values that matters. With the linear change of the two parameters, the measure $D(S)$ changes quadratically in the product. In the neighbourhood of zero the change in $D(S)$ is relatively flat. This heuristic observation provides a further support to our view that even if the aggregation conditions are mildly violated, the aggregate regression is still useful in characterising the macro relationship.

[6] Notice that equation (13) would actually lead to a slightly different definition of $D(S)$, as it would be $D(S) = \|h_0\| |\cos(\tau, h_0)|$. To have a differentiable statistics, anyway, we removed the absolute value, obtaining (13).

3 TESTING FOR PERFECT COINTEGRATION VERSUS SPURIOUS REGRESSION

In this section, we propose a procedure that allows to test for cointegration using the $D(S)$ statistics. We will use the covariance structure of (1)-(3) to derive an estimation framework and propose a testing strategy.

The relevant set of hypotheses to be tested is:

$$\begin{cases} H_0 : D(S) = 0 \\ H_1 : D(S) \neq 0 \end{cases} \tag{14}$$

3.1 Covariance structure

The long run covariance matrix for (1)-(3) is assumed to be

$$\lim_{T\to\infty} \frac{1}{T} E\left(S_T S_T'\right) = \Sigma = 2\pi H_\varepsilon\left(0\right) = \Phi\left(1\right)\left[\Phi\left(1\right)\right]' = \quad + \Lambda + \Lambda'$$

where

1. $H_\varepsilon\left(\omega\right)$ is ε_t's spectral density matrix at density ω;
2. given that $\Phi\left(L\right) = \sum_{j=0}^{\infty} \Phi_j L^j$, under Assumption 1e) we may write

$$\Phi\left(1\right) = \begin{bmatrix} \Phi_{11} & 0 & 0 \\ 0 & \Phi_{22} & 0 \\ 0 & 0 & I_k \end{bmatrix} \tag{15}$$

   where $\Phi_{11} = \Phi_{11}\left(1\right)$ and $H_\varepsilon(0) = \Phi(1)[\Phi(1)]'$ [7];

3. $\equiv \sum_{j=0}^{+\infty} \Phi_j \Phi_j'$ and $\Lambda \equiv \sum_{j=2}^{+\infty} E[\eta_1 \eta_j'] = \sum_{j=2}^{+\infty} \sum_{i=0}^{+\infty} \Phi_{j+1} \Phi_i'$ [8].

The following theorem, with "$\Rightarrow$" denoting weak convergence of the associated probability measures as $T \to \infty$, holds[9]:

Proposition 2 *Let W^* be a $2n+k$-dimensional standard Brownian motion, partitioned as $\left(W^*\right)' = [W'^*_y, W'^*_x, W'^*_z]'$ where the three vectors are of order n, n and k respectively. Under Assumption 1 and using (15) we have:*

1. $T^{-\frac{1}{2}} S_T \Rightarrow \Phi(1) W^*(1)$;

[7] It is possible to introduce non diagonal elements in (15), though the algebra will become more complicated. The results are available upon request.

[8] Details are in Hamilton (1995).

[9] See Park and Phillips (1988).

2. $T^{-1}\sum_{t=1}^{T} S_{t-1}\eta_t' \Rightarrow \Phi(1)\int W^*\left(dW^*\right)'\Phi'(1) + \Lambda$;

3. $T^{-2}\sum_{t=1}^{T} S_{t-1}S_{t-1}' \Rightarrow \Phi(1)\int W^*\left(W^*\right)'\Phi'(1)$.

This theorem considers the convergence of the $I(1)$ process S_T to a linear combination of normally distributed variables (central limit theorem), and of products of the process with itself and with its increments according to a standard framework.

For the sake of notation, in what follows we will refer to $\Phi(1)W^*$ as W, with $W' = [W_y', W_x', W_z']'$, and the W_is have the same meaning as the W_i^*s, $i \in \{y, x, z\}$. Notice that under Assumption 1e) $W_z \equiv W_z^*$. Furthermore, we will also employ the notation $W' = [W_1', W_2', W_3']'$, where the correspondence between numbers and variables is straightforward.

3.2 Estimation

The parameters to be estimated are the β_is and the α_{ij}s, together with the covariance structure. Henceforth, for a generic parameter d and its estimate $\hat{d}$, the limit distribution of the random variable $T\left(\hat{d} - d\right)$ as $T \to \infty$ will be referred to as L_d. We will consider the following estimators:

- the β_is will be estimated via OLS equation by equation. The i-th parameter estimator is

$$\hat{\beta}_i \equiv \frac{\sum_{t=1}^{T} y_{it}x_{it}}{\sum_{t=1}^{T} x_{it}^2}. \tag{16}$$

  Under Assumptions 1a)-1d), $\hat{\beta}_i$ is consistent with limit distribution

$$T(\hat{\beta}_i - \beta_i) \Rightarrow \frac{\alpha_i'\int W_z dW_1^{*\prime}\Phi_{11,i}' + \sum_{j=0}^{\infty}\Phi_{2j,i}\Phi_{1j,i}'}{\alpha_i'\int W_z W_z'\alpha_i} \tag{17}$$

  where $\Phi_{11,i}$ and $\Phi_{22,i}$ are the i-th rows respectively of blocks Φ_{11} and Φ_{22} in matrix $\Phi(1)$. Also, Φ_{kj} is block Φ_{kk} in Φ_j;

- the α_{ij}s will be estimated jointly via OLS as in Park and Phillips (1988). Stacking the vectors in matrices, we get the OLS estimator for A:

$$\hat{A} = (X'Z)(Z'Z)^{-1}. \tag{18}$$

  Assumptions 1a)-1d) ensure that $\hat{A}$ is consistent, and its limit distribution is

$$T(\hat{A} - A) \Rightarrow \left[\Phi_{22} \int dW_2^* W_z'\right] \left[\int W_z W_z'\right]^{-1}. \tag{19}$$

Equations (17) and (19) are the limit distributions we utilize for estimation purposes. We chose to estimate the $\hat{\beta}_i$s equation by equation since estimating matrix B in (1) would have led to a degenerate Brownian motion in the estimate limit distribution.[10]

Notice that there are some unknowns in the limit distribution of both the $\hat{\beta}_i$s and the $\hat{\alpha}_{ij}$s, as they both depend on the covariances and, as far as the β_is are concerned, also on the α_{ij}s. This can be solved by employing a consistent estimator of the true distribution. Let $f = f(\alpha, \Phi)$ be the limit distribution of $T[\hat{\beta}_i - \beta_i]$. Since f is a continuous function of α_{ij} and Φ, an estimator of f, say $\hat{f} = \hat{f}(\hat{\alpha}_{ij}, \hat{\Phi})$, by the Slutsky's theorem will be consistent ($plim \quad \hat{f} = f$) if and only if both $\hat{\alpha}_{ij}$ and $\hat{\Phi}$ are consistent. So, $\hat{\beta}_i$ will be consistent if we can replace α_{ij} and Φ with two consistent estimators. Now, Park and Phillips (1988) show that $\hat{A}$ is a consistent estimator of A, and thus we replace α_{ij}s with the $\hat{\alpha}_{ij}$s. A consistent covariance matrix can be find also using Park and Phillips' (1988) results.

3.3 Testing

In this subsection, we will refer to the parameters employing the compact notation $\theta \equiv [\tau', vec'(A)]'$, with θ a $3n \times 1$ vector and $\hat{\theta}$ its OLS estimator described above. To emphasize the dependence of $D(S)$ on the parameters θ, we will write $D(S) = \psi(\theta)$. The transformation $\psi : C \subset R^{3n} \to R$, where $C = R^{3n} \backslash P$ (P being the manifold described by equation $\|\tau\| = 0$, i.e. by equations $\tau_i = 0$ for all i), is analytical in C.

The main results for testing are in the following theorem[11]:

Proposition 3 *Let* $\widehat{D(S)} = \psi\left(\hat{\theta}\right)$. *Then, under the null hypothesis as in (14)*

$$T\widehat{D(S)} \Rightarrow [J\psi(\theta)]' L_\theta, \tag{20}$$

and, by consistency,

$$\left[J\psi\left(\hat{\theta}\right)\right]' L_\theta \xrightarrow{p} [J\psi(\theta)]' L_\theta,$$

[10] See proof of Proposition 1 in Appendix.

[11] Proof and the explicit expressions for the derivatives are in Appendix 1.

with

$$\left[J\psi\left(\hat{\theta}\right)\right] \equiv \left.\frac{\partial\psi}{\partial\theta}\right|_{\theta=\hat{\theta}} = \left.\begin{bmatrix} \frac{\partial}{\partial\tau} \\ \frac{\partial}{\partial\alpha_2} \\ \frac{\partial}{\partial\alpha_1} \end{bmatrix}\psi\right|_{\theta=\hat{\theta}}$$

and under the null H_0

$$\frac{\partial\psi}{\partial\beta_{\mathrm{i}}} = \frac{\mathrm{h_{0i}}||\tau|| - 2\beta_{\mathrm{i}}\tau'\mathrm{h_0}||\tau||^{-1/2}}{||\tau||^2} = \frac{\mathrm{h_{0i}}}{||\tau||} \tag{21}$$

$$\frac{\partial\psi}{\partial\alpha_{1\mathrm{i}}} = \frac{1}{||\tau||}\sum_{j=1}^{n}[\beta_j(\frac{\partial\mathrm{h_{0j}}}{\partial\alpha_{1\mathrm{i}}})] \tag{22}$$

$$\frac{\partial\psi}{\partial\alpha_{2\mathrm{i}}} = \frac{1}{||\tau||}\sum_{j=1}^{n}[\beta_j(\frac{\partial\mathrm{h_{0j}}}{\partial\alpha_{2\mathrm{i}}})] \tag{23}$$

In equation (20) $\left[J\psi\left(\hat{\theta}\right)\right]' L_\theta$ converges in probability to $[J\psi(\theta)]' L_\theta$ as we prove in Appendix, and it is the approximation one should employ to test whether cointegration is preserved in the aggregate relationship.

3.4 A numerical example

A numerical example can illustrate how to implement the measure of departure from cointegration $D(S)$ in (13) and its testing framework as in (20). Let us suppose a panel consists of data on two variables for five units, and, for the sake of computational simplicity, that $\eta_t \sim N(0, I_{2n+k})$, where obviously $2n + k = 12$. Suppose further that appropriate tests lead to the conclusion that all the series contain a unit root, that individual units cointegrate with coefficients

$$\widehat{\tau} = \begin{pmatrix} 0.046 & 0.037 & 0.254 & 0.53 & 0.807 \end{pmatrix}' \tag{24}$$

and that the independent series are driven by two common stochastic trends. Given this conclusion, the two largest principal components among the independent variables are estimated. These two components are then scaled by the square root of their estimated long-run variance. The scaling makes the sample principle components orthonormal.

In the next step regression of independent variables on the normalized principal components is performed. Let the estimated coefficients be

$$\widehat{A} = \begin{pmatrix} 0.98 & 0.76 & 0.60 & 0.61 & 0.60 \\ 0.49 & 0.38 & 0.30 & 0.43 & 0.38 \end{pmatrix}'. \tag{25}$$

From the coefficient matrices (24) and (25) we have $\|\hat{\tau}\| = 1$ and $\widehat{h} = \begin{pmatrix} -0.029 & -0.022 & -0.018 & 0.045 & 0.023 \end{pmatrix}'$, so that $\cos(\widehat{\tau}, \widehat{h}) = 0.55$. To find $\widehat{h}_0$, we have to multiply matrix $\widehat{A}$ by matrix $\widehat{M_0}$ from Equation (10). The sums of the elements in columns of $\widehat{A}$ are $\hat{a}_1 = 3.55$ and $\hat{a}_2 = 1.98$, so that $\tan\hat{\phi} = \frac{\hat{a}_1 - \hat{a}_2}{\hat{a}_1 + \hat{a}_2} = 0.28$. Then

$$\widehat{M_0} = \begin{pmatrix} 0.96 & -0.27 \\ 0.27 & 0.96 \end{pmatrix}$$

and

$$\widetilde{\widehat{A}} = \widehat{M_0}\widehat{A} = \begin{pmatrix} 0.80 & 0.63 & 0.50 & 0.47 & 0.47 \\ 0.74 & 0.57 & 0.45 & 0.58 & 0.53 \end{pmatrix}'.$$

Using the last outcome, $\widehat{h}_0 = \begin{pmatrix} -0.070 & -0.054 & -0.043 & 0.110 & 0.056 \end{pmatrix}'$, so that $\left\|\widehat{h}_0\right\| = 0.158$ and $\cos(\widehat{\tau}, \widehat{h}_0) = 0.55$. Gathering the results, we see that the departure from cointegration is

$$\widehat{D(S)} = 0.088.$$

In order to test whether $H_0 : D(S) = 0$ is accepted or rejected, we need to simulate critical values using Equation (20). We simulated two possible samples, with lengths $T = 30$ and $T = 100$, viewable respectively as representative of the cases of annual and quarterly data arising in empirical applications. The simulations gave the following results[12]:

T	*p-value*
30	0.1764
100	0.0174

Table 1: P-values for the null hypothesis of cointegration.

Such p-values would lead to accept the null (presence of cointegration) when $T = 30$ and to reject it for $T = 100$.

4 SOME MONTE CARLO EVIDENCE

The results about the distribution of $\hat{\beta}$ were obtained under the large sample hypothesis in the previous section, and were therefore only asymptotically

[12]The simulations were performed with GAUSS and the routine is available upon request. The number of replications we chose for each experiment was 50000.

valid. In this section we would like to see how well the asymptotic distribution characterises the real small sample distribution.

In order to examine the small sample properties of the $\widehat{\beta}$ estimator we evaluate data generated by the system described by equations (1)-(3). As in the main body of the paper, we study only the case of two common stochastic trends in a setting in which only a single trend should be present if the aggregate relationship is to cointegrate. We consider sample sizes of $T = 30$ and $T = 100$, as in Section 2.3, for their being representative of the range of annual and quarterly data in empirical applications. In our experiments we choose the values of parameters $\|\tau\|$, $\|h\|$, $\cos(\tau, h)$ and k and then generate the parameter matrices A and τ randomly subject to the constraints on these four parameters. We carried out the simulations for a broad range of parameters. However, the ratio of small sample means and variances to their asymptotic counterparts was preserved throughout. Therefore, we restrict the above parameters to the value $\|\tau\| = \|h\| = \cos(\tau, h) = k = 1$. Similarly, we consider only the panels with two units, $N = 2$. The number of replications in all experiments is 50 000, as in the previous numerical example.

The innovations in the simulation are set to follow processes of the ARMA (1,1) form where the noise terms are independent standard normal. The set of values for both autoregressive parameter ρ and moving average parameter ϑ is $\{-0.9, -0.3, 0, 0.3, 0.9\}$ where the values ± 0.3 represent a moderate departure from non-autocorrelation and ± 0.9 represent a nearly nonstationary or nearly non-invertible processes. In correspondence with the underlying model, we generate the stochastic trends z by summing the ARMA errors and then we scale them by the square root of their long-run variance $\lambda = (1 + \vartheta)^2/(1 - \rho)^2$. To keep the variance of the innovations comparable across the experiments with different ARMA parameters, we normalize the stationary errors in the equations generating x and y variables by the square root of their variance $\sigma^2 = (\vartheta^2 + 2\vartheta\rho + 1)/(1 - \rho^2)$.

To identify the effect of the serial correlation in different parts of the system, we distinguish four cases. First, the trends are generated by ARMA processes and there is no noise in the process generating variables x and y. The results for the case $T = 30$ are reported in Table 2a. We can see that the mean of the asymptotic distribution is a good guide even in sample of this size. The variance increases monotonically with respect to both autoregressive and moving average parameters. When both ARMA parameters are negative and large, the small sample variance is up to about 3.5 times smaller than the asymptotic value. On the opposite side of the range the asymptotic value underpredicts the actual variance about three times. While in both extremes the discrepancy means that the asymptotic

value is not a precise guide, in the case of negative parameters it means that the spread of the distribution is actually much better than what we would conclude from the limiting distribution. On the other hand, for the samples of this size the variance will be at most three times bigger than the asymptotic variance. It is worth noticing that, as expected, the combinations of parameters where $\rho = -\vartheta$ correspond to the results for white noise, the only exception being the case when ρ is 0.9.

ϑ	-0.9		-0.3		0		0.3		0.9	
ρ	mean	var	mean	var	mean	var	mean	var	mean	var
-0.9	0.500	0.060	0.501	0.064	0.500	0.102	0.500	0.154	0.496	0.223
-0.3	0.499	0.031	0.499	0.145	0.501	0.195	0.504	0.223	0.500	0.241
0	0.500	0.052	0.501	0.179	0.497	0.224	0.500	0.249	0.496	0.256
0.3	0.500	0.092	0.497	0.223	0.500	0.260	0.500	0.272	0.502	0.287
0.9	0.503	0.612	0.501	0.637	0.494	0.684	0.500	0.674	0.504	0.685

Table 2a: Mean and variance of the simulated distribution of $\widehat{\beta}$ in small samples.

ϑ	-0.9		-0.3		0		0.3		0.9	
ρ	mean	var	mean	var	mean	var	mean	var	mean	var
-0.9	0.500	0.021	0.501	0.081	0.500	0.042	0.499	0.187	0.500	0.221
-0.3	0.499	0.028	0.502	0.185	0.498	0.213	0.499	0.225	0.497	0.228
0	0.500	0.051	0.502	0.204	0.499	0.223	0.500	0.229	0.497	0.232
0.3	0.500	0.088	0.499	0.223	0.502	0.234	0.500	0.237	0.500	0.240
0.9	0.498	0.387	0.500	0.319	0.505	0.396	0.499	0.391	0.505	0.391

Table 2b: Mean and variance of the simulated distribution of $\widehat{\beta}$ in small samples.

The experiments for the case of $T = 100$ are given in Table 2b. The results are qualitatively the same as for the previous case. Quantitatively, the small sample values are now closer to their limiting values so that the worst underprediction for the variance now amounts to less than two times.

The second set of experiments is carried out in the same setting with addition of white noise errors into the equations generating the variables x and y. The Table 3a reports the experiment for the case of $T = 30$. In this case, the mean of the short sample distribution tends to be smaller than its asymptotic counterpart. The difference increases with both ρ and ϑ. The variance of the sample distribution is now on average closer to the asymptotic value than in the previous case. Moreover, the small sample variance is now always smaller than the asymptotic value except for values of ρ close to 1.

That means that the real variance will be actually more favourable than its asymptotic prediction. In the case of $T = 100$ (see Table 3b), the pattern of the sample variances is preserved. The small sample values are now closer to the limiting values though the speed of convergence is perhaps not as fast as would be expected. The mean values on the other hand approach quickly the asymptotic value.

ϑ	-0.9		-0.3		0		0.3		0.9	
ρ	mean	var	mean	var	mean	var	mean	var	mean	var
-0.9	0.499	0.061	0.460	0.068	0.433	0.101	0.418	0.138	0.408	0.176
-0.3	0.494	0.032	0.425	0.130	0.409	0.162	0.403	0.175	0.400	0.184
0	0.490	0.054	0.415	0.156	0.404	0.178	0.398	0.185	0.398	0.189
0.3	0.490	0.090	0.406	0.179	0.399	0.191	0.396	0.197	0.396	0.199
0.9	0.443	0.444	0.328	0.235	0.317	0.233	0.314	0.236	0.314	0.233

Table 3a: Mean and variance of the simulated distribution of $\widehat{\beta}$ in small samples.

ϑ	-0.9		-0.3		0		0.3		0.9	
ρ	mean	var	mean	var	mean	var	mean	var	mean	var
-0.9	0.498	0.021	0.476	0.079	0.464	0.130	0.461	0.165	0.456	0.191
-0.3	0.494	0.029	0.462	0.164	0.458	0.182	0.455	0.189	0.457	0.196
0	0.492	0.052	0.455	0.181	0.456	0.190	0.454	0.194	0.455	0.196
0.3	0.488	0.087	0.457	0.193	0.454	0.197	0.452	0.199	0.455	0.203
0.9	0.469	0.341	0.439	0.279	0.438	0.273	0.442	0.274	0.438	0.272

Table 3b: Mean and variance of the simulated distribution of $\widehat{\beta}$ in small samples.

In the third set of experiments, the innovations generating the trends z are white noise while we now allow the errors in x and y to follow ARMA processes. The values of parameters ρ and ϑ in the Tables 4a and 4b now refer to the noise in the variables instead in the trends. In this setting, the asymptotic variance predicts the small sample variance remarkably well. The difference in any direction does not exceed 30 percent for the case of $T = 30$ and 15 percent for the case of $T = 100$. The mean of the sample distribution is overpredicted by the asymptotic value by up to thirty percent in the smaller sample but this undeprediction vanishes quickly as the sample size increases.

ϑ	-0.9		-0.3		0		0.3		0.9	
ρ	mean	var	mean	var	mean	var	mean	var	mean	var
-0.9	0.421	0.199	0.421	0.199	0.421	0.196	0.416	0.189	0.399	0.177
-0.3	0.406	0.172	0.411	0.171	0.405	0.173	0.408	0.180	0.408	0.183
0	0.404	0.170	0.407	0.175	0.405	0.178	0.408	0.182	0.408	0.187
0.3	0.405	0.169	0.408	0.174	0.412	0.187	0.410	0.195	0.411	0.199
0.9	0.398	0.183	0.430	0.262	0.439	0.287	0.435	0.292	0.440	0.287

Table 4a: Mean and variance of the simulated distribution of $\widehat{\beta}$ in small samples.

.

ϑ	-0.9		-0.3		0		0.3		0.9	
ρ	mean	var	mean	var	mean	var	mean	var	mean	var
-0.9	0.457	0.195	0.458	0.194	0.463	0.194	0.457	0.195	0.453	0.188
-0.3	0.459	0.189	0.459	0.190	0.459	0.190	0.461	0.191	0.458	0.192
0	0.461	0.189	0.456	0.191	0.456	0.189	0.456	0.191	0.458	0.193
0.3	0.457	0.190	0.454	0.193	0.457	0.193	0.455	0.193	0.458	0.196
0.9	4.454	0.191	0.464	0.216	0.462	0.218	0.467	0.223	0.462	0.219

Table 4b: Mean and variance of the simulated distribution of $\widehat{\beta}$ in small samples

Finally, in the last set of experiments we let all the innovations in the system to follow an ARMA process with identical parameter values. The mean of the small sample distribution behaves in similar way to the case in which the innovations in the x and y variables followed only a white noise process. The mean is again underpredicted for the smaller sample sizes but the value of the mean is much closer to the asymptotic value in larger samples. The variance, on the other hand, follows the pattern of the case where there was no noise in the variables x and y. The variance again slowly converges to the asymptotic values.

ϑ	-0.9		-0.3		0		0.3		0.9	
ρ	mean	var	mean	var	mean	var	mean	var	mean	var
-0.9	0.497	0.062	0.468	0.086	0.449	0.121	0.433	0.157	0.398	0.174
-0.3	0.493	0.033	0.426	0.129	0.414	0.161	0.409	0.178	0.407	0.191
0	0.490	0.053	0.414	0.152	0.409	0.178	0.404	0.192	0.403	0.206
0.3	0.486	0.092	0.409	0.179	0.409	0.201	0.406	0.214	0.404	0.229
0.9	0.444	0.456	0.351	0.455	0.355	0.524	0.357	0.549	0.356	0.575

Table 5a: Mean and variance of the simulated distribution of $\widehat{\beta}$ in small samples.

θ	-0.9		-0.3		0		0.3		0.9	
ρ	mean	var	mean	var	mean	var	mean	var	mean	var
-0.9	0.499	0.021	0.474	0.083	0.466	0.133	0.463	0.170	0.453	0.190
-0.3	0.493	0.029	0.461	0.164	0.461	0.184	0.456	0.193	0.458	0.197
0	0.493	0.052	0.457	0.178	0.458	0.193	0.455	0.194	0.455	0.199
0.3	0.489	0.086	0.457	0.191	0.458	0.200	0.457	0.205	0.457	0.207
0.9	0.468	0.347	0.443	0.317	0.450	0.338	0.445	0.336	0.443	0.345

Table 5b: Mean and variance of the simulated distribution of $\widehat{\beta}$ in small samples.

From the four sets of experiments it seems that the small sample mean is affected by the presence of noise in the processes generating the variables x and y. The degree of misprediction does not depend very much on the structure of the noise, it means that presence of autocorrelation does not make the results worst. The variance of the small sample distribution, on contrary, is mainly influenced by the presence of autocorrelation and moving average components in the innovation generating z. Finally, the mean and variance of the small sample distribution approach the asymptotic values as T increases.

The main issue that we wanted to address in our Monte Carlo experiments was whether the asymptotic distribution of the $\widehat{\tau}$ estimator was a valuable guide for the small samples case. The conclusion from the experiments is that at worst the variance in the small sample is only three times larger even for relatively large positive values of both autoregressive and moving average parameters. Furthermore, if the degree of autocorrelation is only moderate, the small sample variance is actually lower than the asymptotic value. This leads us to the conclusion that the knowledge of the limiting distribution of the $\widehat{\tau}$ estimator is useful in estimating the upper bound of the degree of non-cointegration in real data.

5 CONCLUSION

In nonstationary heterogeneous panels where each unit cointegrates, the aggregate relationship in general does not cointegrate unless a large number of conditions are satisfied. To satisfy aggregation conditions, the micro regressors need to share a single common stochastic trend. In reality this condition most likely will not be satisfied making the aggregate equation invalid. This paper takes a closer look at the implications for cointegration at macro level when there is a violation of those conditions. The question then is whether

this case can still lead to an aggregate estimator that summarises the macro relationship reasonably well.

We derive the asymptotic measure of the distance, $D(S)$, between the case of perfect cointegration and that with very heterogeneous response of the system variables. We prove that the departure from the cointegration depends linearly on two factors that measure the heterogeneity of the reaction of units to the trends, both across trends and across units. To violate the conditions for cointegration both factors must be different from zero. Further, we propose an estimation and testing procedure to evaluate the degree of non-cointegration in the aggregate estimate. Thus, our framework provides arguments of the usefulness of the notion of mild violation of conditions for preservation of cointegration in estimation and forecasting**.** Indeed, forecasting the aggregate relationship has dramatically different implications depending on the macro relationship being stationary or not. If the macro aggregate is not perfectly cointegrated, the impact of the I(1) component that makes the relationship spurious can still be negligible, and the aggregate dependent variable can still be predicted as if one were in a stationary environment. Qualitatively, within the forecasting framework, the measure $D(S)$ can be interpreted as an assessment of the impact of the nonstationary component on the forecast error. The testing procedure we propose may also be employed to understand whether this impact is significant, and therefore whether forecasting, and all other related issues, is to be performed within a stationary or nonstationary framework.

Finally, we conduct a series of Monte Carlo simulations in order to evaluate the small sample properties of those estimators, i.e. whether the asymptotic distribution of the $\widehat{\tau}$ estimator is a valuable guide for the small samples case. The findings from the simulations are encouraging. The mean and variance of the limiting distribution are good appoximations of the small sample case. Moreover, the small sample mean seems to be affected by the presence of noise in the processes generating the variables x and y, while the degree of misprediction does not depend very much on the structure of the noise. The variance of the small sample distribution, on the contrary, is mainly influenced by the presence of autocorrelation and moving average components in the innovation generating z. Finally, the mean and variance of the small sample distribution approach the asymptotic values as T increases from 30 to 100.

Phillips and Moon (1999) give a comprehensive analysis of cointegration issues in panel. Our paper, as a follow-up of their interesting results, gives support to the view that even if the aggregation conditions are mildly violated, the aggregate regression is still useful in characterising the macro relationship.

References

[1] Bai, J. (2002), "Estimating Cross-Section Common Stochastic Trends on Nonstationary Panel Data", unpublished manuscript, Department of Economics, Boston College.

[2] Forni,M., Hallin, M., Lippi, M. and L. Reichlin (2000), "The Generalized Dynamic-Factor Model: Identification and Estimation", *Review of Economics and Statistics*, 82, 540-554.

[3] Ghose, D. (1995), "Linear Aggregation in Cointegrated Systems", *Journal of Economic Dynamic and Control,* 19, 1011-1032.

[4] Gonzalo, J. (1993), "Cointegration and Aggregation", *Ricerche Economiche,* 47, 281-291.

[5] Granger, C. W. J. (1993), "Implications of Seeing Economic Variables Through an Aggregation Window", *Ricerche Economiche,* 47, 269-279.

[6] Greene, W. (1993), "Econometric Analysis", *Prentice Hall*, 3rd ed.

[7] Hall, S., Lazarova, S. and G. Urga (1999), "A Principal Components Analysis of Common Stochastic Trends in Heterogeneous Panel Data: Some Monte Carlo Evidence", *Oxford Bulletin of Economics and Statistics*, 61, 749-767.

[8] Hamilton, J.D. (1995), "Time Series Analysis", *Princeton University Press.*

[9] Park, J.Y. and Phillips, P.C.B. (1988), "Statistical Inference in Regressions with Integrated Processes: Part 1", *Econometric Theory*, 4, 468-497.

[10] Park, J.Y. and Phillips, P.C.B. (1989), "Statistical Inference in Regressions with Integrated Processes: Part 2", *Econometric Theory*, 5, 95-131.

[11] Pesaran, M.H. and Smith, R. (1995), "Estimating Long-Run Relationships from Dynamic Heterogeneous Panels", *Journal of Econometrics*, Vol. 68, pp. 79-113.

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

[13] Phillips, P.C. B. and Moon, H.R. (1999), "Linear Regression Limit Theory for Nonstationary Panel Data", *Econometrica*, 67, 1057-1111.

Appendix

Proof of Proposition 1

In the non-degenerate case where matrix Σ is strictly non singular, proof of Proposition 1 can be found in Park and Phillips (1988). The structure of model (1)-(3), anyway, prevents the positive definteness requirement from holding. To understand this, one can consider (without loss of generality) the following modified version of the long run covariance structure of the model:

1. Σ is block-diagonal, i.e. no cross dependence is allowed among equations (1)-(3):
$$\Sigma = \begin{bmatrix} \Sigma_{yy} & 0 & 0 \\ 0 & \Sigma_{xx} & 0 \\ 0 & 0 & I_k \end{bmatrix}$$
2. the $n \times n$ submatrix Σ_{yy} is positive definite.

The x_is are cointegrated among themselves, being their joint representation an n-dimensional VAR with k stochastic trends. Hence, there exists a full rank $n \times (n-k)$ array γ s.t.

$$\Sigma_{xx}\gamma = 0 \tag{A.1}$$

according to a standard result in cointegration analysis - see Phillips (1986) or Hamilton (1995). Equation (A.1) states that matrix Σ_{xx} has rank k ,which rules out the positive definiteness assumption. Notwithstanding this, the following algebra shows that Proposition 1 holds. First

$$T\left(\hat{\beta}-\beta\right) = \frac{\frac{1}{T}\sum_{t=1}^{T}\left\{\left[\sum_{i=1}^{n}\left(\beta_i v_{it}+u_{it}\right)\right]\left(\sum_{i=1}^{n}x_{it}\right)\right\}}{\frac{1}{T^2}\sum_{t=1}^{T}\left(\sum_{i=1}^{n}x_{it}^2\right)^2}$$

With respect to the denominator, consider the following passages

$$\begin{aligned}
\frac{1}{T^2}\sum_{t=1}^{T}\left(\sum_{i=1}^{n}x_{it}^2\right)^2 &= \frac{1}{T^2}\sum_{t=1}^{T}\left(i_n' A z_t + v_t\right)^2 = \\
&= \frac{1}{T^2}\sum_{t=1}^{T}\left[\left(i_n' A z_t + v_t\right)\left(i_n' A z_t + v_t\right)'\right] = \\
&= \frac{1}{T^2}\sum_{t=1}^{T}\left[i_n' A z_t z_t' A' i_n + i_n' v_t v_t' i_n + 2 i_n' A z_t v_t' i_n\right] = \\
&= \frac{1}{T^2}\sum_{t=1}^{T}\left(i_n' A z_t z_t' A' i_n\right) \Rightarrow i_n' A \int W W' A' i_n
\end{aligned}$$

with the last equality holding asymptotically and the distribution limit leading to a non degenerate combination of functionals of Brownian motions. The same plain algebra would hold for the numerator, leading to the same conclusion.

Proof of Proposition 3

The result in the Theorem is a straightforward application of the so-called "Delta method" - see Greene (1993) for details. The proof will be sketched in order to underline some interesting features of the asymptotic properties of $\widehat{D(S)}$.

First, $\psi(\cdot)$ analiticity in its domain C allows for Taylor expansion in the neighborhood of any couple $(\theta^*, \theta_0) \in C$:

$$\psi(\theta^*) \approx \psi(\theta_0) + [J\psi(\theta_0)]'(\theta^* - \theta_0) + o(\|\theta^* - \theta_0\|)$$

and therefore

$$\psi\left(\hat{\theta}\right) - \psi(\theta) \approx [J\psi(\theta)]'\left(\hat{\theta} - \theta\right) + o(\left\|\hat{\theta} - \theta\right\|)$$

so that finally

$$T\left[\psi\left(\hat{\theta}\right) - \psi(\theta)\right] \approx T[J\psi(\theta)]'\left(\hat{\theta} - \theta\right) + To(\left\|\hat{\theta} - \theta\right\|) \qquad \text{(A.2)}$$

Now, the following considerations on (A.2) lead to a statistical definition of $\psi(\theta)$ and hence of $D(S)$:

1. the explicit form of $o(\left\|\hat{\theta} - \theta\right\|)$ is

$$o(\left\|\hat{\theta} - \theta\right\|) = \sum_{k=2}^{\infty} q(k;\theta) \left\|\hat{\theta} - \theta\right\|^k$$

As far as the probability limit is concerned, it holds that[13] $\left(\hat{\theta} - \theta\right) = O(T)$, meaning that for any $\delta > 0$, $T^{1-\delta}\left(\hat{\theta} - \theta\right) \to 0$ in probability. Hence, $\left\|\hat{\theta} - \theta\right\|^k = O(T^k)$. For this reason, the term $To(\left\|\hat{\theta} - \theta\right\|)$ in the RHS of equation (A.2) vanishes as T grows to infinity. Formally, $p\lim T\left\|\hat{\theta} - \theta\right\|^k = 0$ and therefore

$$d\lim\ T\left\|\hat{\theta} - \theta\right\|^k = 0$$

[13] It follows from the estimation theory results given above.

"$d\lim$" denoting the distribution limit. This means that

$$\begin{aligned} d\lim\ T\left[\psi\left(\hat{\theta}\right)-\psi\left(\theta\right)\right] &= d\lim\ T\left[J\psi\left(\theta\right)\right]'\left(\hat{\theta}-\theta\right)= \\ &= \left[J\psi\left(\theta\right)\right]' d\lim\ T\left(\hat{\theta}-\theta\right) \end{aligned}$$

2. a problem of equation (A.2) is that $[J\psi(\theta)]$ is not known. According to the Delta method, anyway, what one may do is approximate it via $\left[J\psi\left(\hat{\theta}\right)\right]$. If the Jacobian is a continuous function - which it is being the function analytic - then, according to Slutsky's theorem, one may write

$$\begin{aligned} d\lim\ T\left[\psi\left(\hat{\theta}\right)-\psi\left(\theta\right)\right] &= d\lim\ T\left[J\psi\left(\theta\right)\right]'\left(\hat{\theta}-\theta\right)= \\ &= \left[J\psi\left(\theta\right)\right]' d\lim\ T\left(\hat{\theta}-\theta\right)= \\ &= p\lim\ \left[J\psi\left(\hat{\theta}\right)\right]' d\lim\ T\left(\hat{\theta}-\theta\right) \end{aligned}$$

and hence, asymptotically, $T\left[\psi\left(\hat{\theta}\right)-\psi\left(\theta\right)\right]=T\left[J\psi\left(\hat{\theta}\right)\right]'\left(\hat{\theta}-\theta\right)$.

Under the null, $\psi(\theta)=0$, which eventually leads to (20).

As far as the derivatives involved in (21)-(23), let:

$$\tilde{a}_1 \equiv \sum_{k=1}^{n}\left[\alpha_{1k}\cos\phi-\alpha_{2k}\sin\phi\right]$$

$$\tilde{a}_2 \equiv \sum_{k=1}^{n}\left[\alpha_{1k}\sin\phi+\alpha_{2k}\cos\phi\right]$$

Clearly, by definition, $\tilde{a}_1=\tilde{a}_2\equiv\tilde{a}$. Consider then the following Lemma, whose proof is omitted as it follows from elementary calculus and trigonometrics:

Lemma *Let* $(\cos\phi)_{1i}\equiv\frac{\partial\cos\phi}{\partial\alpha_{1i}}$; $(\cos\phi)_{2i}\equiv\frac{\partial\cos\phi}{\partial\alpha_{2i}}$; $(\sin\phi)_{1i}\equiv\frac{\partial\sin\phi}{\partial\alpha_{1i}}$; $(\sin\phi)_{2i}\equiv\frac{\partial\sin\phi}{\partial\alpha_{1i}}$; $(a_1)_{1i}\equiv\frac{\partial a_1}{\partial\alpha_{1i}}$; $(a_1)_{2i}\equiv\frac{\partial a_1}{\partial\alpha_{2i}}$; $(a_2)_{1i}\equiv\frac{\partial a_2}{\partial\alpha_{1i}}$ and $(a_2)_{2i}\equiv\frac{\partial a_2}{\partial\alpha_{2i}}$. *Then the following results hold:*

$$(\sin\phi)_{1i}\equiv\frac{\partial\sin\phi}{\partial\alpha_{1i}}=2\cos\phi\frac{1}{1+\tan^2\phi}\frac{a_2}{\left(a_1+a_2\right)^2}=(\sin\phi)_1 \quad \text{for any } i$$

$$(\sin\phi)_{2i}=-(\sin\phi)_{1i}\frac{a_1}{a_2}=(\sin\phi)_2 \quad \text{for any } i$$

$$
\begin{aligned}
(\cos\phi)_{1i} &= -(\sin\phi)_{1i}\tan\phi = (\cos\phi)_1 \quad \text{for any } i \\
(\cos\phi)_{2i} &= -(\sin\phi)_{2i}\tan\phi = (\cos\phi)_2 \quad \text{for any } i \\
(\tilde{a}_1)_{1i} = (\tilde{a}_2)_{1i} &= \cos\phi - (\sin\phi)_1 [a_1\tan\phi + a_2] = (\tilde{a})_1 \quad \text{for any } i \\
(\tilde{a}_1)_{2i} = (\tilde{a}_2)_{2i} &= -\sin\phi - (\sin\phi)_2 [a_1\tan\phi + a_2] = (\tilde{a})_2 \quad \text{for any } i
\end{aligned}
$$

Hence, for

$$
d_{ij} = \begin{cases} 1 \text{ if } i = j \\ 0 \quad \text{else} \end{cases}
$$

it holds that

$$
\begin{aligned}
\frac{\partial h_{0j}}{\partial \alpha_{1i}} &= \frac{\{\alpha_{1j}[(\cos\phi)_1 - (\sin\phi)_1] - \alpha_{2j}[(\cos\phi)_1 + (\sin\phi)_1] + d_{ij}[\cos\phi - \sin\phi]\}}{\tilde{a}} + \\
&\quad - \frac{(\tilde{a})_1 [\alpha_{1j}(\cos\phi - \sin\phi) + \alpha_{2j}(\cos\phi + \sin\phi)]}{\tilde{a}^2}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial h_{0j}}{\partial \alpha_{2i}} &= \frac{\{\alpha_{1j}[(\cos\phi)_2 - (\sin\phi)_2] - \alpha_{2j}[(\cos\phi)_2 + (\sin\phi)_2] + d_{ij}[\cos\phi - \sin\phi]\}}{\tilde{a}} + \\
&\quad - \frac{(\tilde{a})_2 [\alpha_{1j}(\cos\phi - \sin\phi) + \alpha_{2j}(\cos\phi + \sin\phi)]}{\tilde{a}^2}
\end{aligned}
$$

Figure 1: Graphs of the distribution of the random variable *f(k)* for different values of k

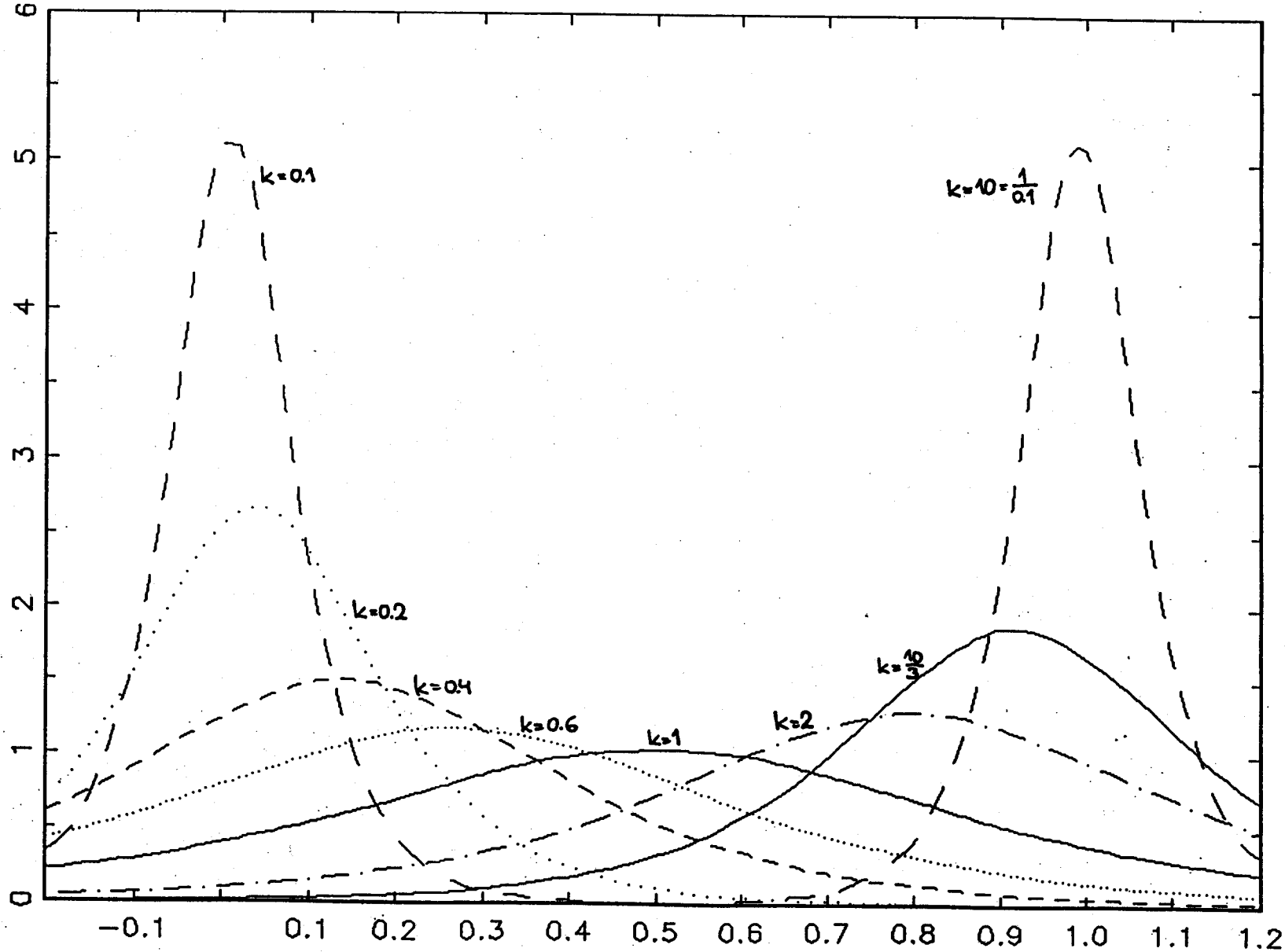
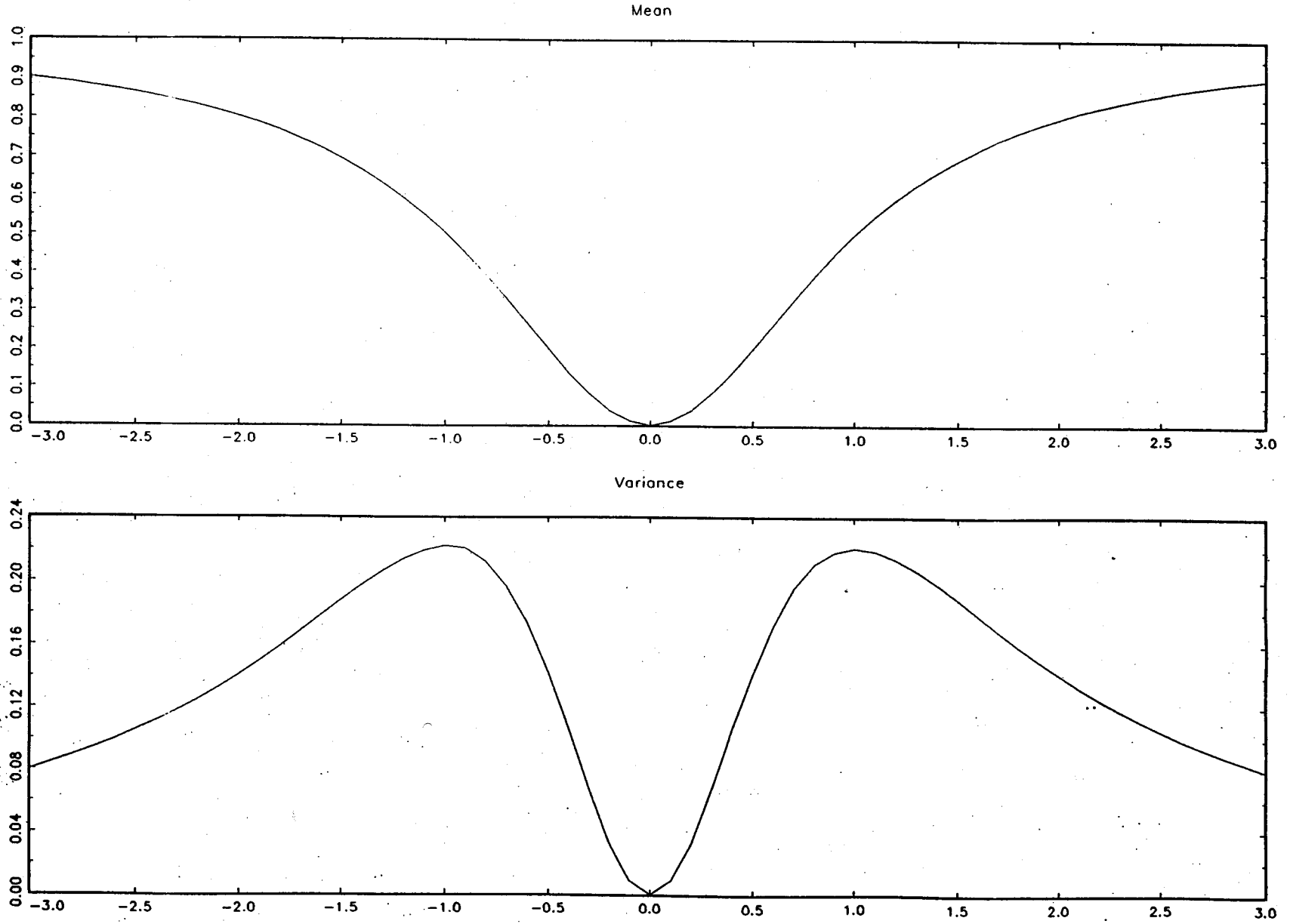


Figure 2: Mean and variance of the distribution of the random variable *f(k)* for different values of k



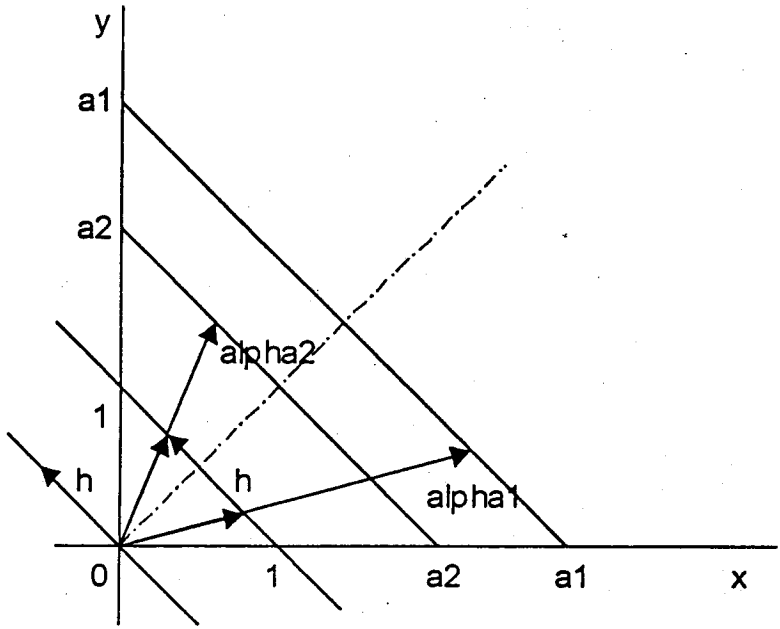


Figure 3: The graphical interpretation of the meaning of the vector h