Identification and estimation of dynamic factors from large macroeconomic panels

*Jörg Breitung and Uta Kretschmer**

University of Bonn
and
Deutsche Bundesbank

February 2004

Abstract

Stock and Watson (2002a, 2002b) have suggested a principal component approach to forecasting macroeconomic variables using dynamic factors extracted from a large number of predictors. As has been demonstrated, the relative performance of this new forecasting technique may be dramatically better than forecasts based on traditional forecasting techniques. An important drawback of the Stock-Watson methodology is, however, that it is not able to distinguish the common dynamic factors from their lags. Therefore, it is difficult to attach any economic meaning to the "reduced form" factors. In our paper a two-step approach is suggested that allows to identify the original "structural" factors from the principal components of the Stock-Watson procedure. An information criterion based on a canonical correlation analysis of the reduced form factors is suggested that allows to determine the number of structural factors and the lag order of the associated lag polynomial. The new methodology is applied to estimate the structural factors from a large macroeconomic panel for the Euro area.

*Address: Institute of Econometrics, Adenauerallee 24-42, 53113 Bonn, Germany. The research for this paper was carried out within research project "Unit roots and cointegration in panel data" financed by the German Research Association (DFG). We like to thank the participants of the seminars at the University of Dortmund and Humboldt University Berlin for their helpful comments and suggestions.

1 Introduction

In a series of paper Stock and Watson (1998, 2002a, 2002b) considered a dynamic factor model for forecasting macroeconomic variables like output growth and inflation. Recent work demonstrates that this approach is a very promising competitor relative to alternative techniques based on ARIMA models or small structural economic models (e.g. Stock and Watson (1999), Angelini, Henry and Mestre (2001), Stock and Watson (2002b), Marcellino, Stock and Watson (2003), Bruneau, de Brandt and Flageollet (2003), Brisson, Campbell and Galbraith (2003)).

An important drawback of the static principal component approach is that it cannot disentangle the original factors from their lags. Moreover, it is assumed that the factors enter the forecasting equation with the same number of lags as in the factor model. Since the principal components are linear combinations of the original factors as well as their lags, it is not possible to attach an economic meaning to the estimated factors. To overcome this difficulty Giannone, Reichlin and Sala (2002) and Forni, Lippi and Reichlin (2003) suggested an identification procedure that is based on a principal component analysis of the residual covariance matrix of a vector autoregression (VAR) estimated from the common factors of the Stock-Watson procedure.

In this paper an alternative identification procedure for the original dynamic factors is proposed. Our approach avoids the estimation of a VAR based on the Stock-Watson factors but employs a canonical correlation analysis (CCA) of the "reduced form factors". The notion behind our identification procedure is that lagged factors are perfectly predictable from the past of the process so that the lags of the factors are associated with unit eigenvalues of the CCA.

An important problem is to determine the number of original dynamic (or "structural") factors from the set of reduced form factors. To this end we suggest a class of specification criteria that allows the consistent estimation of the number of structural factors. Our selection criteria are constructed analogously to the information criteria used to determine the lag order of an ARIMA model (e.g. Akaike (1973), Schwarz (1978)).

The rest of the paper is organized as follows. In Section 2 we first consider the special case that there is only a single lag of the dynamic factors. The general case and some additional details are considered in Section 3. To investigate the small sample properties of the suggested estimation procedure a couple of Monte

Carlo simulations were performed. The results presented in Section 4 suggest that the CCA analysis performs better than the principal component analysis of the residual covariance matrix of a fitted VAR. Furthermore, it is demonstrated that the proposed selection criteria are quite useful for determining the number of structural factors. An empirical example based on 192 monthly time series from 9 European countries is considered in Section 5. The specification criterion of Bai and Ng (2002) suggest 3 reduced form factors estimated by the Stock-Watson procedure. Furthermore, applying our specification criteria we found 2 structural factors. Accordingly, one of the structural factors enters with a lag. We find that the first structural factor is highly correlated with short term interest rates in European countries, whereas the second structural factor is highly correlated with change of the Producer Price Index (PPI) in various countries. Finally, in Section 6 we offer some conclusions and make some suggestions for future work.

2 A two-step estimation procedure

In this section we will focus on a simple case of the dynamic factor model. The main issues involved are best understood by looking at a simple case. Generalizations to more general models and some additional technical details are considered in the following section.

Consider the dynamic factor model

$$\begin{aligned} x_t &= A_0 f_t + A_1 f_{t-1} + u_t \ , \ t = 1, \ldots, T, \\ &= [A_0 \vdots A_1] \begin{bmatrix} f_t \\ f_{t-1} \end{bmatrix} + u_t \equiv A F_t + u_t \end{aligned} \tag{1}$$

where x_t is a $N \times 1$ vector of time series, u_t is a $N \times 1$ vector of idiosyncratic errors and f_t is a $k \times 1$ vector of independent common factors with vector autoregressive representation

$$f_t = \Gamma f_{t-1} + \varepsilon_t \ ,$$

where Γ and $\Sigma_\varepsilon = E(\varepsilon_t \varepsilon_t')$ are $k \times k$ matrices. Under suitable assumptions, which will be considered in the following section, Stock and Watson (2002a) showed that $r = 2k$ eigenvalues of the covariance matrix $\Sigma_x = E(x_t x_t')$ are $O(N)$, whereas the remaining eigenvalues are $O(1)$. Using the spectral decomposition of the sample covariance matrix $\widehat{\Sigma}_x = \sum_{i=1}^{N} \lambda_i v_i v_i'$, where $v_1, \ldots, v_N$ are the orthonormal eigenvectors corresponding to the ordered eigenvalues $\lambda_1 \leq \cdots \leq \lambda_N$, the common component $\xi_t = A_0 f_t + A_1 f_{t-1}$ can be estimated using the first r principal

components

$$\hat{\xi}_t = V_r \widehat{F}_t \ ,$$

where $V_r = [v_1, \ldots, v_r]$ and $\widehat{F}_t = V_r' x_t$. The principal components (PC) estimator is based on the linear combinations $\widehat{F}_t$ that minimize the least-squares criterion function

$$S_{NT} = \frac{1}{NT} \sum_{i=1}^{T} (x_t - B\widetilde{F}_t)'(x_t - B\widetilde{F}_t) \tag{2}$$

with respect to B and $\widetilde{F}_t$ (cf. Stock and Watson 2002a).

To identify the "dynamic factors" (or structural factors) f_t from the "static" (reduced form) factors F_t Giannone et al. (2002) and Forni et al. (2003) suggest an additional principal component analysis based on the residual variance covariance matrix of the VAR

$$F_t = CF_{t-1} + \eta_t \ , \tag{3}$$

where

$$C = \begin{bmatrix} \Gamma & 0 \\ I_k & 0 \end{bmatrix}$$

and $\eta_t = [\varepsilon_t', 0]'$. As shown by Bai and Ng (2002) $\widehat{F}_t$ is a consistent estimator for HF_t where H is a regular rotation matrix. Therefore, if F_t is replaced by $\widehat{F}_t$ the VAR becomes

$$\widehat{F}_t = C^* \widehat{F}_{t-1} + e_t^* \ .$$

In this representation $C^* = HCH^{-1}$ and, for $N \to \infty$, the asymptotic covariance matrix of e_t^* is $\Sigma_e^* = H\Sigma_\eta H'$, where $\Sigma_\eta = E(\eta_t \eta_t')$. Obviously, Σ_e^* is of rank k because $rk(\Sigma_\eta) = k$. Therefore, Giannone et al. (2002) suggest to estimate the structural factors by applying a principal component analysis to the residual covariance matrix $\widehat{\Sigma}_e^*$.

An alternative approach is to exploit the singularity of the matrix C^* and estimate the structural factors by a reduced rank regression of $\widehat{F}_t$ on $\widehat{F}_{t-1}$. Let

$$S_{ij} = T^{-1} \sum_{t=2}^{T} \widehat{F}_{t-i} \widehat{F}_{t-j}', \quad i, j \in \{0, 1\} \ .$$

The canonical correlation analysis (CCA) is based on the generalized eigenvalue problem

$$|\lambda S_{00} - S_{01} S_{11}^{-1} S_{10}| = 0$$
$$\text{or} \qquad |\lambda S_{11} - S_{10} S_{00}^{-1} S_{01}| = 0.$$

Note that the j'th eigenvalue can be seen as an (uncentered) R^2 from a regression of the linear combination of $w_j'\widehat{F}_t$ on $\widehat{F}_{t-1}$, where w_j is the eigenvector associated with the j'th eigenvalue. Since k linear combinations of $\widehat{F}_t$ given by $\tilde{\varepsilon}_t = [I, -\Gamma]H^{-1}\widehat{F}_t$ are (asymptotically) unpredictable, k eigenvalues of (4) converge to zero. On the other hand the linear combination $\tilde{f}_{t-1} = [0, I_k]H^{-1}\widehat{F}_t$ is perfectly predictable given $\widehat{F}_{t-1}$ as N and T tend to infinity. Thus, k eigenvalues of (4) converge to unity in probability.

Let w_j denote the j'th eigenvector associated with the ordered eigenvalue $\hat{\mu}_j \in \{\hat{\mu}_1 \geq \cdots \geq \hat{\mu}_r\}$ of (4). Furthermore, let $W_k = [w_1, \ldots, w_k]$ denote the $r \times k$ matrix of eigenvectors corresponding to the k largest eigenvalues. Then, as N and T tend to infinity $W_k'\widehat{F}_{t-1}$ span the same space as f_{t-1} and, therefore, $\widehat{f}_t = W_k'\widehat{F}_t$ is an estimator for f_t.

Note that the eigenvalue problem (4) implies the normalization $W_k'S_{11}'W_k = I_k$. Since S_{11} converges to $\Sigma_z = E(\widehat{F}_t\widehat{F}_t')$, if the common factors are stationary, it follows that the estimated structural components $\widehat{f}_t$ are asymptotically uncorrelated.

3 Extracting the common dynamic factors in the general model

In this section the simple model with $m = 1$ and $p = 1$ is generalized to allow for an arbitrary number of lags. Furthermore, some technical details of the identification of the dynamic factors are considered.

The general model is

$$\begin{aligned} x_t &= A_0 f_t + \cdots + A_m f_{t-m} + u_t \qquad (4) \\ &\equiv AF_t + u_t \ , \qquad (5) \end{aligned}$$

where the vector of common dynamic factors possesses a VAR(p) representation

$$f_t = \Gamma_1 f_{t-1} + \cdots + \Gamma_p f_{t-p} + \varepsilon_t \ ,$$

$\Gamma_1, \ldots, \Gamma_p$ and $\Sigma = E(\varepsilon_t\varepsilon_t')$ being $k \times k$ matrices.

Following Stock and Watson (2002a) we assume that the idiosyncratic errors u_t may be weakly correlated across series, whereas the common factors $f_t, \ldots, f_{t-p}$ give rise to a strong correlation among the series. These properties are formalized in the following

Assumption: *(a) Let* $\gamma_{N,t}(j) = N^{-1}E(u_t'u_{t+k})$. *Then,* $\lim_{N\to\infty} \sup_t \sum_{j=-\infty}^{\infty} |\gamma_{N,t}(j)| < \infty$. *(b) The eigenvalues of* $\Sigma_u = E(u_t u_t')$ *are* $o(N)$ *whereas the* $r = k(p+1)$ *largest eigenvalues of* $\Sigma_x = E(x_t x_t')$ *are* $O(N)$. *(c)* $\lim_{N\to\infty} \sup_{t,s} N^{-1} \sum_{i=1}^{N} \sum_{i=1}^{T} |cov(u_{is}u_{it}, u_{js}u_{jt}) < \infty$.

We first consider the problem of identifying the lag lengths m and p in (4). Let $C_j = E(Z_t Z_{t-j}')$ denote the j'th autocovariance matrix of the static (reduced form) factors, where $Z_t = V_r' x_t$ and V_r is the matrix of the first r eigenvectors of the covariance matrix $\Sigma_x = E(x_t x_t')$. Note that, as T tend to infinity, $\widehat{F}_t$ converges in probability to Z_t. Furthermore, if A in (5) has full column rank, then Z_t converges to HF_t as N and T tends to infinity, where H is a regular $r \times r$ matrix (cf. Bai and Ng 2002). To determine the lag length m the following property can be used.

Property 1: *The eigenvalues of a CCA between* Z_t *and* Z_{t-m^*} *based on the eigenvalue problem*

$$|\lambda C_0 - C_{m^*}' C_0^{-1} C_{m^*}| = 0 \tag{6}$$

are all less then one if $m^* > m$.

This property follows from the fact that for $m^* > m$ the vectors F_t and F_{t-m^*} do not share any lags and, therefore, there exists no linear combination of F_t that is perfectly predictable using F_{t-m^*}. This suggests that m can be determined by checking whether a CCA between F_t and F_{t-m^*} (resp. the estimated counterparts) results in eigenvalues sufficiently samller than unity.

To determine the lag length p we can again use a property based on the canonical correlation of Z_t and its lags. Let $Z_t^{(j)} = [Z_t', Z_{t-2}', \ldots, Z_{t-2j}']'$ and $\widetilde{C}_j = E(Z_t^{(j)} {Z_{t-1}^{(j)}}')$. Then, the following property turns out to be useful for the identification of p.

Property 2: *Assume that* A *is of full column rank and* $m \leq 1$. *The eigenvalues of a CCA between* $Z_t^{(\ell)}$ *and* $Z_{t-1}^{(\ell)}$ *based on the eigenvalue problem*

$$|\lambda \widetilde{C}_0 - \widetilde{C}_\ell' \widetilde{C}_0^{-1} \widetilde{C}_\ell| = 0 \tag{7}$$

are all zero or one if $2(\ell+1) > p(m+1)$.

This property results from the fact that in an autoregression of order p the vector $Z_t^{(\ell)}$ must include all relevant lags $F_t, \ldots, F_{t-p}$ to yield a linear combination $v_j' Z_t^{(\ell)}$

that is not predictable conditional on $Z_{t-1}^{(\ell)}$. If ℓ is too small then there exist no linear combination that equals the innovation of the process and, therefore, the corresponding eigenvalue is larger than zero but less then unity. Assume, for example, that $f_t = \gamma_1 f_{t-1} + \gamma_2 f_{t-2} + \epsilon_t$ and $m = 1$ so that $Z_t = Z_t^{(0)}$ includes only f_t and f_{t-1}. Then all linear combinations of f_t and f_{t-1} are predictable conditional on $Z_{t-1}^{(0)}$ but not perfectly so, as the innovation of the process enters the linear combination. According to property 2, the lag must be specified as $\ell = 1$ to ensure that all eigenvalues are either zero or one.

Note that Property 2 requires that $m \geq 1$ and, by assuming full column rank of A, all elements of f_t and f_{t-1} enters the common factor representation of y_t. This might appear as a quite restrictive assumption as it exclude cases, where some of the dynamic factors enter without lag. However, we can sidestep this problem by over-specifying the number of static factors r. Assume that f_t enters without lag so that $r = k$. Estimating $r = 2k$ factors renders some linear combinations of the factors $f_t^{(1)} * = \varepsilon_t$ and $f_t^{(2)} = \Gamma_1 f_{t-1} + \cdots + \Gamma_p f_{t-p}$. Therefore, given that A_0 has full rank, it is sufficient to assume that $f_t^{(2)}$ has a nonsingular covariance matrix which only rules out that some of the factors are white noise. As this is a testable assumption, this assumption does not seem to be very problematical in practice.

These properties suggest to determine the lag orders m and p by using a sequential procedure. First, the lag order m is determined based on Property 1, that is, $\hat{m}$ is the smallest number m^* such that all eigenvalues of (6) are either zero or one. Second, Property 2 is used to determine an upper bound for the lag length p, which results from the inequality $p < \hat{p}_{max} = 2\hat{\ell}/(m+1)$, where $\hat{\ell}$ is the smallest number such that all eigenvalues are either zero or one. The actual lag order p can be determined in a later stage by selecting the lag-order in a vector autoregression of the dynamic factors f_t.

To determine the number of dynamic factors the following property can be used.

Property 3: *Assume that A has full column rank. If $\ell > p(m+1)/2$, then k eigenvalues of (7) are equal to zero, whereas all other eigenvalues are larger than zero.*

This property suggests to determine the number of dynamic factors by considering the number of zero eigenvalues of (7). To this end, the likelihood-ratio statistic

of Tiao and Tsay (1989) can be used. The test statistic is based on the smallest k eigenvalues of (7)

$$LR(k) = -T \sum_{i=r-k+1}^{r} \log(1 - \hat{\mu}_i) \ ,$$

where μ_i^0 is the i'th ordered eigenvalue of (4). Under the null hypothesis of k zero eigenvalues the LR statistic is asymptotically χ^2 distributed. If $r - k$ lags of f_{jt} enter the vector F_t, then $r - k$ eigenvalues are known to be equal to one and the problem reduces to the calculation of the remaining k eigenvalues. Therefore, the $LR(k)$ statistic has an asymptotic χ^2 distribution with k^2 degrees of freedom. In the following theorem it is stated that the asymptotic properties of the $LR(k)$ statistic are not affected by replacing the true factors F_t by the their PCA counterparts $\widehat{F}_t$.

Theorem 1: *Let $\widehat{F}_t$ denote the vector of static factors. Assume that Assumption 1 holds, the dimension of f_t is k and A has full column rank. Let $\hat{\mu}_1 \geq \cdots \geq \hat{\mu}_r$ denote the eigenvalues of the problem*

$$|\hat{\mu}\widehat{S}_{11} - \widehat{S}_{10}\widehat{S}_{00}^{-1}\widehat{S}_{10}'| = 0 \ , \tag{8}$$

where

$$\widehat{S}_{ij} = T^{-1} \sum_{t=2}^{T} \widehat{F}_{t-i}\widehat{F}_{t-j}', \quad i, j \in \{0, 1\}.$$

If $min\{N, T\} \to \infty$, then the statistic $LR(k)$ is asymptotically χ^2 distributed with k^2 degrees of freedom.

It is also possible to construct model selection criteria for a consistent choice of the number of structural factors k (cf. Bai and Ng 2003). The general form of the model selection criterion is given by

$$IC(k^*, r) = LR(k^*) + (r - k^*)^2 c(T), \tag{9}$$

where the LR statistic is used to measure the fit relative to a specification without restriction. Note that the term $(r - k^*)^2$ represents the number of parameters in the conditional model $F_t|F_{t-1}$. For the penalty function $c(T)$ we consider

$$\begin{aligned} &\text{AIC:} \qquad c(T) = 2 \\ &\text{SIC:} \qquad c(T) = \log(T). \end{aligned}$$

It is well known that minimizing the Schwartz criterion (SIC) yields a weakly consistent estimator for the model order, whereas for the Akaike criterion (AIC) the probability to choose $\hat{k} > k$ does not converge to zero. As the following theorem states, this property carries over when computing the model selection criteria by using the estimated factors $\widehat{F}_t$.

Theorem 2: *Let $\hat{\mu}_1 \leq \cdots \leq \hat{\mu}_r$ denote the eigenvalues of (12). Under the assumptions of Theorem 1, minimizing $IC(k^*, r)$ with respect to k^* yields a consistent estimator of k if $N \to \infty$, $c(T) \to \infty$ and $c(T)/T \to 0$.*

4 Small sample properties

To investigate the performance of the proposed methodology to estimate the (number of) structural factors some Monte Carlo experiments were conducted. The data is simulated by the model

$$x_t = A_0 f_t + A_1 f_{t-1} + u_t \ , \tag{10}$$

where the components of the k-dimensional vector f_t are independent AR(1) processes with $f_{it} = \gamma f_{i,t-1} + \varepsilon_{it}$ $(i = 1, \ldots, k)$ with $\varepsilon_{it} \sim N(0,1)$. The elements of the $N \times k$ matrices are independent uniform random variables from the interval (0,1). The idiosyncratic errors are generated as $u_t \sim i.i.N(0, I_N)$. We also tried out alternative models based on various parameter values but the general conclusions remain the same.All results are based on 10,000 replications of the model.

First, we compare the performance of our identification procedure and the approach suggested by Forni et al. (2003). To this end we compute the R^2 from a regression of the true structural factor (if $k = 1$) on the estimated structural factors. In the case $k = 2$ we follow Boivin and Ng (2003) and compute the following measure:

$$R^2_* = \frac{tr[(\sum f_t \hat{f}_t')(\sum \hat{f}_t \hat{f}_t')^{-1}(\sum \hat{f}_t f_t')]}{tr[\sum f_t f_t']} \ .$$

Table 1 presents the means of the R^2 or R^2_* measures for various sample sizes N and T. It turns out that in most cases the identification procedure based on a CCA of the reduced form factors performs better then the identification procedure based on a PCA. Furthermore, the results suggest that both procedures appear

to consistently estimate the true factors as the performance measures converge to unity for $N, T \to 0$.

To investigate the ability of the information criterion to find the correct number of structural factors k, we simulate data according to model (10) with $k = 1$ and $k = 2$. The number of structural factors is determined by using the information criterion with $c(T) = 2$ [AIC] and $c(T) = \log(T)$ [SIC]. To concentrate on the properties of the proposed information criteria, we assume the number of reduced form factors (r) as known. In practice, the information criteria suggested by Bai and Ng (2002) can be used to obtain a consistent estimate of r.

The results for various values of N and T can be found in Table 2. If $K = 1$ and $p = 1$ it follows that $r = 2$. Accordingly, the information criterium has to be computed for the two possibilities: $k = 1$ (the true number) and $k = 2$ (with no lags). It turns out that both criteria are very successful in finding out the true number of structural factors, even in rather small sample sizes. The LR statistic applied to test the hypothesis $k = 1$ accepts the correct model in roughly 95 percent of the cases. This results suggests that the actual size of the LR test is close to the nominal size of 0.05.

If $k = 2$ the criteria have problems in determining the underlying number of structural factors. In this case we have $r = 4$ and, therefore, the number of possible combinations of k and p are substantially larger. As a result, the information criteria are less successful in this case. From results reported in Table 2 it turns out that both selection criteria have difficulties to determine the correct number of structural factors. In particular, the SIC performs poorly if T is small. As T becomes more substantial, the SIC criterium improves and it is apparent that it is able to find out the correct number of structural factors with a probability converging to one as T tends to infinity.

The AIC criterion tends to perform better for a small number of time periods. However, as expected from Theorem 1, the AIC criterion is not consistent since the probability of over-specification of the structural factors does not disappear as $T \to \infty$.

5 Empirical Application

In this section we apply our proposed methodology for extracting structural factors to a macroeconomic data set of the Euro area. The data consists of the main macroeconomic categories on country level and of monetary series on Euro-wide

aggregate level. In particular we compare the estimated "structural factors" to the "reduced form factors" resulting from the principal component analysis of Stock and Watson (2002a).

5.1 The Data

Our data comprise national as well as some Euro area aggregate monthly macroeconomic time series from 1984:02 to 2002:12. To cover this time span we have to confine ourselves to nine of the twelve Euro area countries. We had to drop Portugal and Ireland for which most data were available from the beginning of 1986 or only on a quarterly basis, respectively. Greece entered the Euro area in 2001, making the available time span too short to be useful for our analysis. For each country we include approximately 25 variables representing the main macroeconomic categories: real output, employment, real retail, manufacturing, consumption, housing, inventories, exchange rates, interest rates, prices and money. A detailed list of all variables is given in Appendix B. Nearly all national data are taken from the OECD Main Economic Indicators database provided by Datastream.

This data set was prepared for our analysis in several respects. Firstly, seasonally unadjusted data was revised using the Tramo/Seat methodology (cf. Maravall and Gomez (2001)). Secondly, all nonnegative series excluding interest rates were transformed into their logarithms. According to the results of unit root tests we take first differences of the variables. These transformations were also applied to prices and wages although it is not clear whether these series are all either I(1) or I(2) variables. Since we prefer to apply the same transformations across countries for each group of variables, such as prices or wages, we decided to assume that prices and wages are $I(1)$ for all countries. For the same reason we retained interest rates in levels.

In addition all series were standardized to have sample mean zero and unit sample variance. Furthermore, we eliminated large outliers and structural breaks due to the reunification of Germany in 1990. Finally, a visual inspection of the data was also conducted and series being suspect of major redefinitions, problems in the data collection and other inconsistencies were discarded from the data set.

Data on the Euro area aggregate level are provided on the ECB website and contain information on money aggregates (M1, M2 and M3). The data were preprocessed in the same way like the country level data. Our final data set comprises 192 monthly time series being available for the full time period from

1984:02 to 2002:12.

5.2 Empirical Results

The estimation of the dynamic factor model given in equation (4) involves determining the number of common factors. Following Stock and Watson (2002a) the dynamic factor model with p lags and k factors can be represented as a static factor model with $r = k(p+1)$ "reduced form" factors. The number of factors can be estimated using the information criteria proposed by Bai and Ng (2002). Minimizing the value of the least squares objective function of the factor model plus some penalty function depending on N and T yields a consistent estimate of the number of static factors. Their selection criterion considered here is specified as[1]

$$IC_{p2}(r^*) = \log[\widehat{S}_{NT}(r^*)] + r^* \left(\frac{N+T}{NT}\right) \log\left[\min(N,T)\right]. \tag{11}$$

and $S_{NT}(r^*)$ constructed as (2). We compute $IC_{p2}(r^*)$ for $r^* = 1, 2, \ldots, 20$. The results for the $r^* = 1, \ldots, 6$ are given in the second column of Table 3. The criteria is minimized at $r^* = 3$ indicating that three reduced form factors are appropriate for the Euro area data set and that they can be estimated consistently using the principal component approach of Stock and Watson (2002a).

However, the estimated number of static factors gives no information on the dynamic structure of the factor model. In our case, we have three different possible combinations of a factor model resulting in three static factors: (i) there are in fact three static factors, that is we have a pure static factor model; (ii) there is one dynamic factor with three lags or (iii) there is a dynamic factor model involving one dynamic factor having one lag and one static factor. Figure 1 shows the three estimated static factors, where the order of the factors corresponds to the size of the eigenvalues. Visual inspection suggests that the first factor looks different from the other two factors. There does not seem to be a dynamic relationship to the other two factors, whereas a relation among the second and third factor might exist. However, drawing conclusions about the dynamic structure of the factors and consequently determining the number of structural factors based only on graphical representation is difficult. Therefore, we apply the second step of our proposed estimation procedure.

The structural factors are estimated using the canonical correlation analysis

[1]Due to its precision and robustness in presence of serially and cross-correlated idiosyncratic error terms (see Bai and Ng 2002), we have decided to use this specific criteria.

(CCA) of the three static factors estimated by the principal component method. The number of common dynamic factors can be determined by the $IC(k^*, r)$ model selection criterion suggested in Section 3. The smallest eigenvalue of the CCA between $\widehat{F}_t$ and $\widehat{F}_{t-1}$ is 0.199. The LR statistic is 50.40 clearly rejecting the hypothesis that this eigenvalue is equal to zero. This result suggests that the lag order of the autoregressive representation of the dynamic factors is larger than the number of lags that enter the factor representation (5). We therefore augment the vector of factors by further lags such that $Z_t^{(1)} = [Z_t', Z_{t-2}']'$ (see Property 2). The resulting test statistics and model selection criteria are presented in Table 3. Whereas the AIC criterion suggests a single dynamic factor, the SIC criterion is minimized at $k = 2$. For the latter number the LR statistic is close to the critical value of 0.05. Summing up, there is some evidence that there are one or two dynamic factors.

The (maximal) two dynamic factors that are identified by the canonical correlation analysis are depicted in Figure 2. Comparing the structural factors with the reduced form factors estimated by the principal component approach of Stock and Watson shows that first structural factor coincides with the first reduced form factor. This is also apparent from the estimated linear combination which is for the first structural parameter obtained as $w_1 = [0.998, -0.020, 0.057]'$. The second structural factor involves the eigenvector $w_2' = [-0.038, 0.524, 0.851]$ and, therefore, can be seen as a linear combination of the second and third factor of the Stock-Watson analysis.

Having found two structural factors in the data set, it is interesting to find a reasonable interpretation for these factors. To this end we have computed the correlation coefficients between the structural factors and all variables in the data set. The 10 variables with the highest correlation to the two structural factors are listed in the Appendix (see Table B.1). The correlations suggest that the first structural factor represent the short run interest rates in the Euro area. In Figure 3 this factor is compared to the 3-month Euribor of France. The second structural factor is highly correlated with inflation as measured by PPI of various countries. For example, as can be seen from Figure 4, the second factor evolves similar to the PPI inflation in France. It is obvious that the there is a strong correlation between these series.

6 Conclusion

In this paper an identification procedure is suggested that allows to estimate the number of dynamic factors (or "structural factors") from the static factors (or "reduced form factors") obtained from a principal component analysis (PCA) as in Stock and Watson (2002a, 2002b). As the reduced form factors are linear combinations of the original and lagged structural factors it does not make sense to attach any economic meaning to the reduced form factors. Therefore, it is important to disentangle the original structural factors from the reduced form factors.

In contrast to the procedure suggested by Giannone et al. (2002) and Forni et al. (2003), which is based on a principal component analysis of the residual covariance matrix from a fitted VAR, our procedure is based on a canonical correlation analysis (CCA) of the reduced form factors. An important advantage of this approach is that it is straightforward to construct model selection criteria similar to the ones used in traditional time series analysis (e.g. Akaike and Schwarz criteria). Moreover, our simulation experiments show that the CCA procedure outperforms the PCA procedure in relevant sample sizes.

We have applied the CCA procedure to a set of 192 monthly time series of 9 European countries. Applying the criteria of Bai and Ng (2002) we found three reduced form factors, whereas our criteria suggest two structural factors, i.e., one of the reduced form factors enters without a lag. In order to attach an economic meaning to the structural factors we consider the correlation to the original variables. The results suggest that the first structural factor represents short term interest rates, whereas the second factor is related to PPI inflation of various countries.

We expect that the identification of the underlying structural factors is also useful for forecasting and structural modelling. For example, the dynamic factors may enter the PCA and the forecasting equation with a different lag order. In this case the reduced form factors may imply a number lags that are unnecessary for forecasting. Identifying the structural factors allows a separate determination of the lag length in the forecast equation and therefore improves the flexibility of the forecast.

Appendix A: Proofs

Proof of Theorem 1

We first proof a Lemma on the distribution of the eigenvalues of the CCA if the vectors y_t and x_t have some identical elements.

Lemma A.1. *Let $y_t = [z_t', y_{2t}']'$ and $x_t = [z_t', x_{2t}']'$, where z_t (x_{2t} and y_{2t}) is $n \times 1$ ($m \times 1$). Assume that there exist m linear combinations $[w_1'y_t, \ldots, w_m'y_t]' \equiv W'y_t$ with $E(W'y_t|x_t) = 0$ and $Var(W'y_t|x_t) = \Sigma_m$. Furthermore $E(x_t) = 0$ and $E(x_t x_t') = \Sigma_x$. Then, as $T \to \infty$, the LR statistic is asymptotically distributed as*

$$LR(k) = -T \sum_{i=n+1}^{n+m} \log(1 - \lambda_i) \xrightarrow{d} \chi^2(m^2),$$

where λ_i denote the eigenvalues (in descending order) of

$$|\lambda \widehat{S}_{yy} - \widehat{S}_{yx} \widehat{S}_{xx}^{-1} \widehat{S}_{xy}'| = 0 \ , \tag{12}$$

and

$$\widehat{S}_{ab} = T^{-1} \sum_{t=2}^{T} \widehat{a}_t \widehat{b}_t', \quad a, b \in \{x, y\}.$$

Proof: Let $\tilde{x}_{2t}$ ($\tilde{y}_{2t}$) denote the projection residuals of x_{2t} (y_{2t}) on z_t, and

$$\tilde{x}_t = \begin{bmatrix} z_t \\ \tilde{x}_{2t} \end{bmatrix}, \quad \tilde{y}_t = \begin{bmatrix} z_t \\ \tilde{y}_{2t} \end{bmatrix}.$$

Accordingly we define

$$\begin{aligned} S_{\tilde{x}\tilde{x}} &= \frac{1}{T} \sum_{t=1}^{T} \tilde{x}_t \tilde{x}_t' = \begin{bmatrix} S_{zz} & 0 \\ 0 & S_{\tilde{x}_2\tilde{x}_2} \end{bmatrix} \\ S_{\tilde{y}\tilde{y}} &= \frac{1}{T} \sum_{t=1}^{T} \tilde{x}_t \tilde{x}_t' = \begin{bmatrix} S_{zz} & 0 \\ 0 & S_{\tilde{y}_2\tilde{y}_2} \end{bmatrix} \\ S_{\tilde{x}\tilde{y}} &= \frac{1}{T} \sum_{t=1}^{T} \tilde{x}_t \tilde{x}_t' = \begin{bmatrix} S_{zz} & 0 \\ 0 & S_{\tilde{x}_2\tilde{y}_2} \end{bmatrix} \end{aligned}$$

Since z_t enters both x_t and y_t, the first n eigenvalues are unity and the corresponding eigenvectors are $[v_1, \ldots, v_n] = V = [I, 0]'$ and the remaining eigenvectors can

be represented as $W_T = [w_{T1}, \ldots, w_{Tm}]' = [0, B_T']'$, where $B_T = [b_{T1}, \ldots, b_{Tm}]$. The eigenvalues $\lambda_{n+1}, \ldots, \lambda_{n+m}$ can be written as

$$\begin{aligned} \lambda_{n+j} &= w_{Tj}' S_{\tilde{y}\tilde{y}}^{-1/2} S_{\tilde{x}\tilde{y}}' S_{\tilde{x}\tilde{x}}^{-1} S_{\tilde{x}\tilde{y}} S_{\tilde{y}\tilde{y}}^{-1/2} w_{Tj} \\ &= b_{Tj}' S_{\tilde{y}_2\tilde{y}_2}^{-1/2} S_{\tilde{x}_2\tilde{y}_2}' S_{\tilde{x}_2\tilde{x}_2}^{-1} S_{\tilde{x}_2\tilde{y}_2} S_{\tilde{y}_2\tilde{y}_2}^{-1/2} b_{Tj}. \end{aligned}$$

Using $b_{Tj} \xrightarrow{p} b_j$, $S_{\tilde{y}_2\tilde{y}_2} \xrightarrow{p} \Sigma_{\tilde{y}_2\tilde{y}_2}$, $S_{\tilde{x}_2\tilde{x}_2} \xrightarrow{p} \Sigma_{\tilde{x}_2\tilde{x}_2}$ and $\sqrt{T} vec(\Sigma_{\tilde{y}_2\tilde{y}_2}^{-1/2} S_{\tilde{x}_2\tilde{y}_2}' \Sigma_{\tilde{x}_2\tilde{x}_2}^{-1/2}) \xrightarrow{d} N(0, I_{m^2})$ it follows that $T \sum_{n+1}^{n+m} \lambda_j$ has an asymptotic χ^2 limiting distribution with m^2 degrees of freedom. Finally, $LR(m) = T \sum_{n+1}^{n+m} \lambda_j + o_p(1)$ and, therefore $LR(m)$ is asymptotically $\chi^2(m^2)$ distributed. ■

As shown by Bai (2003, Theorem 1) the estimated factors can asymptotically be represented as

$$\widehat{F}_t = H' F_t + \frac{1}{\sqrt{N}} \xi_{Nt} + o_p(N^{-1/2}) \tag{13}$$

where

$$\xi_{Nt} = V^{-1} Q \frac{1}{\sqrt{N}} \sum_{i=1}^{N} a_i u_{it} \ ,$$

and V is a diagonal matrix with the probability limit of the first r eigenvalues of $(1/NT)\widehat{\Sigma}_x$ on the leading diagonal, $Q = \lim_{N\to\infty} \lim_{T\to\infty} E(T^{-1} F_t' H F_t)$ and a_i' is the i'th row of the matrix A in (5). For $N \to \infty$ it follows that

$$\begin{aligned} \widehat{S}_{00} &= T^{-1} \sum_{T=1}^{T} H' F_t F_t' H + o_p(1) \\ \widehat{S}_{11} &= T^{-1} \sum_{T=1}^{T} H' F_{t-1} F_{t-1}' H + o_p(1) \\ \widehat{S}_{10} &= T^{-1} \sum_{T=1}^{T} H' F_{t-1} F_t' H + o_p(1) \ . \end{aligned}$$

Furthermore, due to the normalization of the principal component estimator we have $\widetilde{S}_{00} = I_r + O(T^{-1})$ and $\widetilde{S}_{11} = I_r + O(T^{-1})$. Accordingly, the eigenvalues can asymptotically be represented as

$$\lambda_j = \tilde{v}_j' \widehat{S}_{10}' \widehat{S}_{10} \tilde{v}_j$$

where v_j is the probability limit of the eigenvector corresponding to the j'th eigenvalue. Since $m \cdot k$ components of F_t are linear combinations of the lags $f_{t-1}, \ldots, f_{t-m}$, these components are common to F_t and F_{t-1}. From Lemma A.1

it follows that in this case the k smallest eigenvalues are χ^2 distributed with k^2 degrees of freedom, whereas the remaining eigenvalues converge to unity as $N \to \infty$.

Proof of Theorem 2

We first give an important result on the asymptotic behavior of the eigenvalues of (**??**). To simplify the notation the we assume $p = 1$ so that $\widehat{F}_t^+ = \widehat{F}_t$. The results remain the same if $\widehat{F}_t$ is replaced by $\widehat{F}_t^+$.

Lemma A.2: *As $N \to \infty$ it holds that $\hat{\mu}_j$ is $O_p(1)$ for $j = 1, \ldots, r-k$, whereas $\hat{\mu}_j$ is $O_p(T^{-1})$ for $j = r-k+1, \ldots, r$.*

PROOF: The j'th eigenvalue can be represented as

$$\begin{aligned}
\hat{\mu}_j &= \frac{\left(T^{-1}\sum_{t=2}^{T} v_j'\widehat{F}_t\widehat{F}_{t-1}'\right)\left(T^{-1}\sum_{t=2}^{T}\widehat{F}_{t-1}\widehat{F}_{t-1}'\right)^{-1}\left(T^{-1}\sum_{t=2}^{T}\widehat{F}_{t-1}\widehat{F}_{t-1}'v_j\right)}{\left(T^{-1}\sum_{t=2}^{T} v_j'\widehat{F}_t\widehat{F}_t'v_j\right)} \\
&= \left(T^{-1}\sum_{t=2}^{T} v_j'\widehat{F}_t\widehat{F}_{t-1}'\right)\left(T^{-1}\sum_{t=2}^{T}\widehat{F}_{t-1}\widehat{F}_t'v_j\right) + O_p(T^{-1}) \\
&= \begin{cases} \mu_j^0 + O_p(T^{-1/2}) + O_p(N^{-1}) & \text{for } \mu > 0 \\ O_p(T^{-1}) + O_p(N^{-1}) & \text{for } \mu_j^0 = 0 \end{cases}
\end{aligned}$$

Since $\mu_j^0 = 1$ for $j = 1, \ldots, r-k$ and $\mu_j^0 = 0$ for $j = r-k+1, \ldots, r$ it follows that $\hat{\mu}_j$ is $O_p(1)$ for $j = 1, \ldots, r-k$ and $O_p(T^{-1/2})$ for $j = 1, \ldots, r-k$. Therefore, from Lemma A.2 it follows for $\delta = 1, 2, \ldots$ that

$$\frac{1}{T}[IC(k+\delta, r) - IC(k, r)] = -\sum_{i=r-k-\delta}^{r-k-1} \log(1-\hat{\mu}_j) - \delta c(T)/T \ .$$

The first term of the the r.h.s. of this equation is positive and $O_p(1)$. Thus, if $c(T)/T \to 0$ as $T \to \infty$, then

$$\lim_{T\to\infty} P[IC(k+\delta, r) - IC(k, r) < 0] = 0$$

for all values of N.

Next, consider

$$\begin{aligned} IC(k-\delta,r) - IC(k,r) &= T \sum_{i=r-k+1}^{r-k+\delta} \log(1-\hat{\mu}_j) + r\delta c(T) \\ &\simeq -T \sum_{i=r-k+1}^{r-k+\delta} \hat{\mu}_j + r\delta c(T) \ . \end{aligned}$$

From Lemma A.2 it follows that as $N \to \infty$, the first term is negative and $O_p(1)$. Thus, if $c(T) \to +\infty$, then

$$\lim_{T\to\infty} P[IC(k-\delta,r) - IC(k,r) < 0] = 0.$$

It follows that under the conditions given in the theorem the criterion $IC(k^*,r)$ yields a consistent estimator of the number of structural factors.

Appendix B: The data set

BELGIUM

BG PPI - CHEMICALSNADJ
BG COMPOSITE L. INDICATOR: NEW CAR REGISTRATIONS SADJ
BG CONSTRUCTION - BUILDING PERMITS ISSUED, RESIDENTIAL VOLA
BG CONSTRUCTION - BUILDINGS STARTED NADJ
BG CONSTRUCTION - DWELLINGS STARTED VOLA
BG CONSTRUCTION - PERMITS ISSUED NADJ
BG CPI - ALL ITEMS NON-FOOD NON-ENERGY NADJ
BG CPI - ENERGY NADJ
BG CPI - FOOD EXCL. RESTAURANTS NADJ
BG CPI - RENT NADJ
BG EXPORTS FOB CURA
BG IMPORTS CIF CURA
BG INDUSTRIAL PRODUCTION - CONSTRUCTION VOLA
BG INDUSTRIAL PRODUCTION - CONSUMER GOODS, DURABLES VOLA
BG INDUSTRIAL PRODUCTION - CONSUMER GOODS, NON-DURABLES VOLA
BG INDUSTRIAL PRODUCTION - CRUDE STEEL NADJ
BG INDUSTRIAL PRODUCTION - MANUFACTURING VOLA
BG NET TRADE BALANCE CURA
BG PPI - ALL ITEMSNADJ
BG PPI - CONSUMER GOODS NADJ
BG PPI - FOOD BEVERAGES & TOBACCO NADJ
BG PPI - INTERMEDIATE GOODS NADJ
BG PPI - MANUFACTURED GOODS NADJ
BG PPI - PETROLEUMPRODUCTS NADJ
BG REGISTERED UNEMPLOYMENT (PERCENT OF TOTAL LABOUR FORCE) SADJ
BG RETAIL SALES VOLA
BG YIELD OF GOVERNMTNT BONDS (5 YEAR)
BG LEND. RATES, MORTGAGE LOANS TO HOUSHOLDS
BG LEND. RATES, SHORT-TERM LOANS TO ENTERPRISES, 6 MONTHS
BG LEND. RATES, SHORT-TERM LOANS TO ENTERPRISES, BANK ADVANCES
BG LEND. RATES, MEDIUM & LONG-TERM LOANS TO ENTERPRISES
BG TIME DEPOSITS, MATURITY LENGHT: 3 MONTHS

GERMANY

BD CALL MONEY RATE
BD COMPOSITE L. IND.: 6-MONTHS RATE OF CHANGE AT ANNUAL RAT
BD CONSTRUCTION - PERMITS ISSUED CURN
BD CPI - ENERGY NADJ
BD CPI - FOOD (EXCL. REST) NADJ
BD CPI - NON-FOOD, NON-ENERGY NADJ
BD CPI - RENT NADJ
BD CPI NADJ
BD EXPORTS FOB CURA
BD IMPORTS CIF CURA
BD INDUSTRIAL PRODUCTION - CRUDE STEEL NADJ
BD INDUSTRIAL PRODUCTION - INDUSTRY EXCLUDING CONSTRUCTION VOLA
BD PASSENGER CARS REGISTERED VOLA
BD PPI - ALL ITEMS NADJ
BD PPI - MANUFACTURING INDUSTRY NADJ
BD REAL EFFECTIVE EXCHANGE RATES
BD RETAIL TURNOVER VOLA
BD LEND. RATES, SHORT-TERM LOANS TO ENTERPRISES, 6 MONTHS

SPAIN

ES CPI- ALL ITEMS NON- FOOD NON-ENERGY NADJ
ES CPI NADJ

SPAIN (continued)

ES EXPORTS FOB CURA
ES IMPORTS CIF CURA
ES INDUSTRIAL PRODUCTION SADJ
ES MONEY SUPPLY: M3 - SPANISHCONTRIBUTION TO EURO M3 CURN
ES PASSENGER CARS REGISTERED VOLA
ES PPI - CONSUMER GOODS NADJ
ES PPI - INTERMEDIATE GOODS NADJ
ES PPI - INVESTMENT GOODS NADJ
ES PPI - MANUFACTURING ALL ITEMS NADJ
ES REAL EFFECTIVE EXCHANGE RATE INDEX - CPI BASED SADJ
ES GOVT BOND YIELD, LONGTERM
ES TREASURY BILL RATE
ES LEND. RATES, MORTGAGE LOANS TO HOUSHOLDS; OVER 3 YEARS
ES LEND. RATES, CONSUMER LOANS, OVER ONE YEAR
ES LEND. RATES, SHORT-TERM LOANS TO ENTERPRISES, VARIABLE RATES
ES LEND. RATES, MEDIUM & LONG TERM LOANS TO ENTERPRISES, 1 to 3 YEARS
ES TIME DEPOSITS, MATURITY LENGHTS: OVER 1 AND UP TO 3 YEARS

FINLAND

FN BOP: CURRENT BALANCE CURN
FN BOP: FINANCIAL BALANCE INCL. RESERVES CURN
FN BOP: NET ERRORS AND OMISSIONS CURN
FN CPI - ENERGY NADJ
FN CPI - FOOD NADJ
FN CPI - NON FOOD NON ENERGY NADJ
FN EMPLOYMENT - INDUSTRY VOLN
FN EXPORTS FOB CURA
FN IMPORTS CIF CURA
FN INDUSTRIAL PRODUCTION - CONSUMER GOODS VOLA
FN INDUSTRIAL PRODUCTION - CRUDE STEEL NADJ
FN INDUSTRIAL PRODUCTION - INTERMEDIATE GOODS VOLA
FN INDUSTRIAL PRODUCTION - INVESTMENT GOODS VOLA
FN INDUSTRIAL PRODUCTION - MANUFACTURING VOLA
FN INDUSTRIAL PRODUCTION - WOOD FELLINGS SADJ
FN OFFICIAL RESERVES EXCLUDING GOLD CURN
FN PPI - CONSUMER GOODS NADJ
FN PPI - INTERMEDIATE GOODSNADJ
FN PPI - INVESTMENT GOODS NADJ
FN PPI - MANUFACTURING NADJ
FN LEND. RATES, MORTGAGE LOANS TO HOUSHOLDS, HOUSING LOANS
FN LEND. RATES, CONSUMER LOANS, CREDIT TO HOUSHOLDS"
FN LEND. RATES, MEDIUM & LONG TERM LOANS TO ENTERPRISES

FRANCE

FR COMPOSITE LEADING INDICATOR: NEW CAR REGISTRATIONS VOLA
FR COMPOSITE LEADING INDICATOR: SHARE PRICES SBF 250 VOLA
FR CPI - ENERGY NADJ
FR CPI - FOOD NADJ
FR CPI - RENT NADJ
FR CPI - SERVICES EXCLUDING RENT NADJ
FR EXPORTS FOB CURA
FR IMPORTS FOB CURA
FR INDUSTRIAL PRODUCTION - CONSUMER GOODS VOLA
FR INDUSTRIAL PRODUCTION - ENERGY VOLA
FR INDUSTRIAL PRODUCTION - INVESTMENT GOODS VOLA
FR INDUSTRIAL PRODUCTION - MANUFACTURING VOLA
FR PPI - AGRICULTURAL GOODS SADJ
FR PPI - CHEMICALS NADJ
FR PPI - MANUFACTURED PRODUCTS NADJ
FR PPI - METAL PRODUCTS NADJ

FRANCE (continued)

FR REGISTERED UNEMPLOYED VOLA
FR RETAIL SALES VOLA
FR UNIT LABOUR COST- TEXTILE INDUSTRIES NADJ
FR COMPOSITE LEADING INDICATOR: 3 MONTH INTERBANK RATE(PIBOR)
FR COMPOSITE LEADING INDICATOR: BOND YIELD GUARANTEED BY GOVT.
FR COMPOSITE LEADING INDICATOR: BOND YIELD GUARANTEED BY GOVT.
FR TIME DEPOSITS, EURIBOR 3 MONTHS, PRIOR TO 1999 PIBOR 3 MONTHS

ITALY

IT CPI - ALL ITEMS NON-FOOD NON-ENERGY NADJ
IT CPI - ALL ITEMS NON-FOOD NON-ENERGY NADJ
IT CPI - ENERGY NADJ
IT CPI - FOOD NADJ
IT CPI - SERVICES LESS HOUSING NADJ
IT CPI NADJ
IT EXPORTS FOB CURA
IT GROSS BOND ISSUES - BANKING SECTOR CURN
IT HOURLY RATES - INDUSTRY NADJ
IT IMPORTS CIF CURA
IT INDUSTRIAL PRODUCTION - CONSUMER GOODS VOLA
IT INDUSTRIAL PRODUCTION - CRUDE STEEL NADJ
IT INDUSTRIAL PRODUCTION - INTERMEDIATE GOODS VOLA
IT INDUSTRIAL PRODUCTION - INVESTMENT GOODS VOLA
IT INDUSTRIAL PRODUCTION - PASSENGER CARS NADJ
IT MANUFACTURING - NEW ORDERS VOLN
IT OFFICIAL RESERVES EXCLUDING GOLD CURN
IT PPI NADJ

LUXEMBOURG

LX CPI- ALL ITEMS NON-FOOD NON-ENERGY NADJ
LX CPI - FOOD NADJ
LX CPI - FOOD NADJ
LX CPI ENERGY NADJ
LX EMPLOYMENT - INDUSTRY VOLN
LX EMPLOYMENT - IRON & STEEL VOLN
LX HOURS OF WORK - INDUSTRY, MONTHLY VOLN
LX INDUSTRIAL PRODUCTION - CONSTRUCTION VOLA
LX INDUSTRIAL PRODUCTION - CRUDE STEEL NADJ
LX INDUSTRIAL PRODUCTION - INDUSTRY EXCLUDING CONSTRUCTION VOLA
LX INDUSTRIAL PRODUCTION - MANUFACTURING VOLA
LX MONTHLY EARNINGS - INDUSTRY VOLN
LX PPI - INDUSTRIAL PRODUCTS NADJ
LX REGISTERED UNEMPLOYED VOLA

NETHERLANDS

NL CONSTRUCTION - PERMITS ISSUED CURN
NL CPI - ENERGY NADJ
NL CPI - FOOD NADJ
NL CPI - NON FOOD-NON ENERGY NADJ
NL CPI NADJ
NL CPI- RENT NADJ
NL EXPORTS FOB CURA
NL HOURLY WAGE RATE MANUFACTURING VOLN
NL INDUSTRIAL PRODUCTION - CRUDE STEEL NADJ
NL INDUSTRIAL PRODUCTION - MANUFACTURING VOLA
NL INDUSTRIAL PRODUCTION - NATURAL GAS NADJ
NL PPI - CONSUMER GOODS OUTPUT NADJ
NL PPI - CRUDE PETROLEUM OUTPUT NADJ
NL PPI - PETROLEUMPRODUCTS NADJ

NETHERLANDS (continued)

NL PPI - INTERMEDIATE GOODS OUTPUT NADJ
NL PPI - INVESTMENT GOODS OUTPUT NADJ
NL PPI - MANUFACTURED GOODS NADJ
NL PPI - OUTPUT NADJ
NL PPI - TOTAL INPUT NADJ
NL RETAIL SALES VOLA
NL COMPOSITE LEADING INDICATOR: YIELD ON LONG-TERM GOVT.BONDS
NL COMPOSITE LEADING INDICATOR: YIELD ON LONG-TERM GOVT.BONDS
NL LEND. RATES, MORTGAGE LOANS TO HOUSHOLDS, FIXED FOR 5 YEARS
NL LEND. RATES, SHORT-TERM LOANS TO ENTERPRISES, BANK BASE RATE
NL TIME DEPOSITS, MATURITY LENGHTS, 2 YEARS
NL TIME DEPOSITS, MATURITY LENGHTS, 4 YEARS

AUSTRIA

AT CPI - ALL ITEMS NON-FOOD NON-ENERGY NADJ
AT WPI - FOOD NADJ
AT 1-YEAR PUBLIC SECTOR BONDS
AT CPI - ENERGY NADJ
AT CPI - FOOD INCL. RESTAURANTS NADJ
AT CPI - RENT NADJ
AT EXPORTS FOB CURA
AT HOURLY WAGE RATES - ALL INDUSTRIES NADJ
AT IMPORTS CIF CURA
AT INDUSTRIAL PRODUCTION - CRUDE STEEL NADJ
AT INDUSTRIAL PRODUCTION - MANUFACTURING VOLA
AT OFFICIAL DISCOUNT RATE
AT OFFICIAL RESERVES EXCLUDING GOLD CURN
AT PPI: MANUFACTURED PRODUCTS NADJ
AT RETAIL SALES VOLA
AT SHARE PRICES - VSE WBI INDEX NADJ

EU

M1 ECB
M2 ECB
M3 ECB

Table B.2: Variables with highest correlation to each structural factor

correlation coefficient	name of series
first structural factor	
0.9741	FR TIME DEPOSITS, EURIBOR 3 MONTHS, PRIOR TO 1999 PIBOR 3 MONTHS
0.9630	NL TIME DEPOSITS, MATURITY LENGTH, 4 YEARS
0.9611	AT OFFICIAL DISCOUNT RATE
0.9596	BG LEND. RATES, SHORT-TERM LOANS TO ENTERPRISES, 6 MONTHS
0.9574	NL LEND. RATES, MORTGAGE LOANS TO HOUSEHOLDS, FIXED FOR 5 YEARS
0.9527	ES TIME DEPOSITS, MATURITY LENGTHS: OVER 1 AND UP TO 3 YEARS
0.9461	BG LEND. RATES, MORTGAGE LOANS TO HOUSHOLDS
0.9455	NL TIME DEPOSITS, MATURITY LENGTH, 2 YEARS
0.9431	BG TIME DEPOSITS, MATURITY LENGTH, 3 MONTHS
0.9416	ES LEND. RATES, MORTGAGE LOANS TO HOUSEHOLDS, OVER 3 YEARS
second structural factor	
0.7127	FR PPI - MANUFACTURED PRODUCTS NADJ
0.6275	BD LEND. RATES, SHORT-TERM LOANS TO ENTERPRISES, 6 MONTHS
0.6076	FR PPI - CHEMICALS NADJ
0.5868	ES PPI - MANUFACTURING ALL ITEMS NADJ
0.5569	BD COMPOSITE L. IND.: 6-MONTHS RATE OF CHANGE AT ANNUAL RATE
0.5255	NL CPI- RENT NADJ
0.5236	FN PPI - MANUFACTURING NADJ
0.5225	IT PPI NADJ
0.5060	ES PPI - INTERMEDIATE GOODS NADJ
0.5005	FR PPI - METAL PRODUCTS NADJ

References

Akaike, H. (1973), Maximum Likelihood Identification of Gaussian Autoregressive-Moving Average Models, Biometrika, 60, 255–265.

Angelini, E. J. Henry and R. Mestre (2001), Diffusion Index-based Inflation Forecasts for the Euro Area, ECB Working Paper No. 61.

Bai, J. and S. Ng (2002), Determining the Number of Factors in Approximate Factor Models, 70, 191–221.

Boivin, J. and S. Ng (2003), Are More Data Always Better for Factor Analysis?, NBER Working Paper Series No. 9829.

Brisson, M., B. Campbell and J.W. Galbraith (2003), Journal of Forecasting, 22, 515–531.

Bruneau, C. O de Brandt, A. Flageollet (2003), Forecasting Inflation in the Euro Area, mimeo.

Forni, M., M. Lippi and L. Reichlin (2003) Opening the black box: structural factor models versus structural VARs, http://www.dynfactors.org.

Giannone, D, L. Reichlin and L. Sala (2002), Tracking Greenspan: Systematic and unsystematic monetary policy revisited, manuscript, http://www.dynfactors.org.

Marcellino, M., J.H. Stock, M.W. Watson (2003), Macroeconomic Forecasting in the Euro Area: Country specific versus area-wide information, European Economic Review, 47, 1–18.

Schwarz, G. (1978), Estimating the Dimension of a Model, Annnals of Statisitcs, 6, 461–464.

Stock, J.H. and M.W. Watson (1998), Diffusion Indices, NBER Working Paper No. 6702.

Stock, J.H. and M.W. Watson (1999), Forecasting Inflation, Journal of Monetary Economics, 44, 293–335.

Stock, J.H. and M.W. Watson (2002a), Forecasting Using Principal Components from a Large Number of Predictors, Journal of the American Statistical Association, 97, 1167–79.

Stock, J.H. and M.W. Watson (2002b), Macroeconomic Forecasting Using Diffusion Indexes, Journal of Business and Economic Statistics, 20, 147–162.

Tiao, G.C. and R.S. Tsay (1989), Model Specification in Multivariate Time Series, Journal of the Royal Statistical Society, Series B, 51, 157–213.

Table 1: Simulation results

	T=50		T=100		T=150	
	CCA	PCA	CCA	PCA	CCA	PCA
	one common factor (k=1), measure: R^2					
	$\gamma = 0.4$					
N=10	0.8369	0.7763	0.8368	0.7846	0.8364	0.7868
	(0.0433)	(0.0728)	(0.0415)	(0.0630)	(0.0404)	(0.0589)
N=50	0.8502	0.7888	0.8489	0.7961	0.8489	0.7992
	(0.0263)	(0.0575)	(0.0236)	(0.0447)	(0.0233)	(0.0408)
N=100	0.8535	0.7922	0.8526	0.8001	0.8522	0.8026
	(0.0196)	(0.0511)	(0.0176)	(0.0387)	(0.0168)	(0.0335)
N=150	0.8544	0.7938	0.8534	0.8009	0.8532	0.8034
	(0.0173)	(0.0488)	(0.0151)	(0.0364)	(0.0142)	(0.0310)
N=200	0.8552	0.7943	0.8540	0.8015	0.8537	0.8041
	(0.0156)	(0.0475)	(0.0135)	(0.0355)	(0.0127)	(0.0297)
	$\gamma = 0.8$					
N=10	0.8913	0.8898	0.9017	0.9034	0.9051	0.9077
	(0.0413)	(0.0495)	(0.0331)	(0.0377)	(0.0291)	(0.0327)
N=50	0.9061	0.9069	0.9145	0.9193	0.9176	0.9233
	(0.0289)	(0.0381)	(0.0216)	(0.0260)	(0.0188)	(0.0221)
N=100	0.9097	0.9122	0.9178	0.9241	0.9210	0.9284
	(0.0264)	(0.0357)	(0.0198)	(0.0241)	(0.0164)	(0.0191)
N=150	0.9108	0.9138	0.9188	0.9256	0.9215	0.9293
	(0.0253)	(0.0347)	(0.0189)	(0.0231)	(0.0157)	(0.0184)
N=200	0.9109	0.9140	0.9190	0.9261	0.9220	0.9300
	(0.0249)	(0.0344)	(0.0187)	(0.0229)	(0.0153)	(0.0180)
	two common factors (k=2), measure: S_{F,F^0}					
	$\gamma_1 = \gamma_2 = 0.4$					
N=10	0.8072	0.7038	0.8089	0.7077	0.8085	0.7088
	(0.0609)	(0.1052)	(0.0559)	(0.1017)	(0.0565)	(0.1013)
N=50	0.8794	0.8073	0.8795	0.8123	0.8793	0.8131
	(0.0212)	(0.0433)	(0.0174)	(0.0346)	(0.0162)	(0.0314)
N=100	0.8876	0.8207	0.8867	0.8238	0.8866	0.8248
	0.0182	0.0378	0.0143	0.0287	0.0126	0.0248
N=150	0.8898	0.8239	0.8890	0.8272	0.8890	0.8287
	(0.0174)	(0.0370)	(0.0134)	(0.0272)	(0.0116)	(0.0229)
N=200	0.8905	0.8251	0.8900	0.8292	0.8899	0.8300
	(0.0169)	(0.0365)	(0.0125)	(0.0260)	(0.0109)	(0.0221)
	$\gamma_1 = \gamma_2 = 0.8$					
N=10	0.8039	0.7052	0.8119	0.7122	0.8136	0.7148
	(0.0714)	(0.1209)	(0.0640)	(0.1188)	(0.0643)	(0.1197)
N=50	0.9013	0.8428	0.9057	0.8497	0.9070	0.8522
	(0.02209	(0.0441)	(0.0178)	(0.0395)	(0.0162)	(0.0371)
N=100	0.9139	0.8600	0.9173	0.8663	0.9186	0.8683
	(0.0185)	(0.0343)	(0.0145)	(0.0283)	(0.0126)	(0.0264)
N=150	0.9174	0.8651	0.9210	0.8712	0.9223	0.8734
	(0.0178)	(0.0315)	(0.0137)	(0.0250)	(0.0116)	(0.0224)
N=200	0.9189	0.8673	0.9225	0.8739	0.9239	0.8755
	(0.0178)	(0.0301)	(0.0129)	(0.0228)	(0.0112)	(0.0206)

Table 2: Rates of success of model selection criteria

	$T = 50$			$T = 100$			$T = 150$		
	AIC	SIC	LR	AIC	SIC	LR	AIC	SIC	LR
one common factor ($k = 1$)									
$N = 10$	1.000	1.000	0.962	1.000	1.000	0.947	1.000	1.000	0.949
$N = 50$	1.000	1.000	0.957	1.000	1.000	0.943	1.000	1.000	0.954
$N = 100$	1.000	1.000	0.948	1.000	1.000	0.956	1.000	1.000	0.943
$N = 150$	1.000	1.000	0.955	1.000	1.000	0.952	1.000	1.000	0.948
$N = 200$	1.000	1.000	0.952	1.000	1.000	0.952	1.000	1.000	0.943
two common factors ($k = 2$)									
$N = 10$	0.840	0.554	0.968	0.924	0.777	0.946	0.957	0.906	0.941
$N = 50$	0.968	0.981	0.933	0.965	1.000	0.941	0.969	1.000	0.948
$N = 100$	0.965	0.995	0.928	0.974	1.000	0.946	0.983	1.000	0.955
$N = 150$	0.961	1.000	0.937	0.967	0.999	0.942	0.958	1.000	0.934
$N = 200$	0.973	1.000	0.935	0.970	1.000	0.955	0.982	1.000	0.954

Note: Entries report the frequencies (in percent) of choosing the correct model using $c(T) = 2$ for AIC and $c(T) = \log(T)$ for SIC. "LR" indicates the selection according to a LR statistic using a significance level of 0.05.

Table 3: Information criteria

# factors	IC_{p2}	AIC	SIC	LR(p-value)
1	–0.1512	0.0418*	0.1022	0.2213
2	–0.1860	0.0516	0.0667*	0.0455
3	–0.1961*	0.1909	0.1909	0.0000
4	–0.1867	—	—	—
5	–0.1765	—	—	—
6	–0.1635	—	—	—

Figure 1: **Estimated reduced form factors**

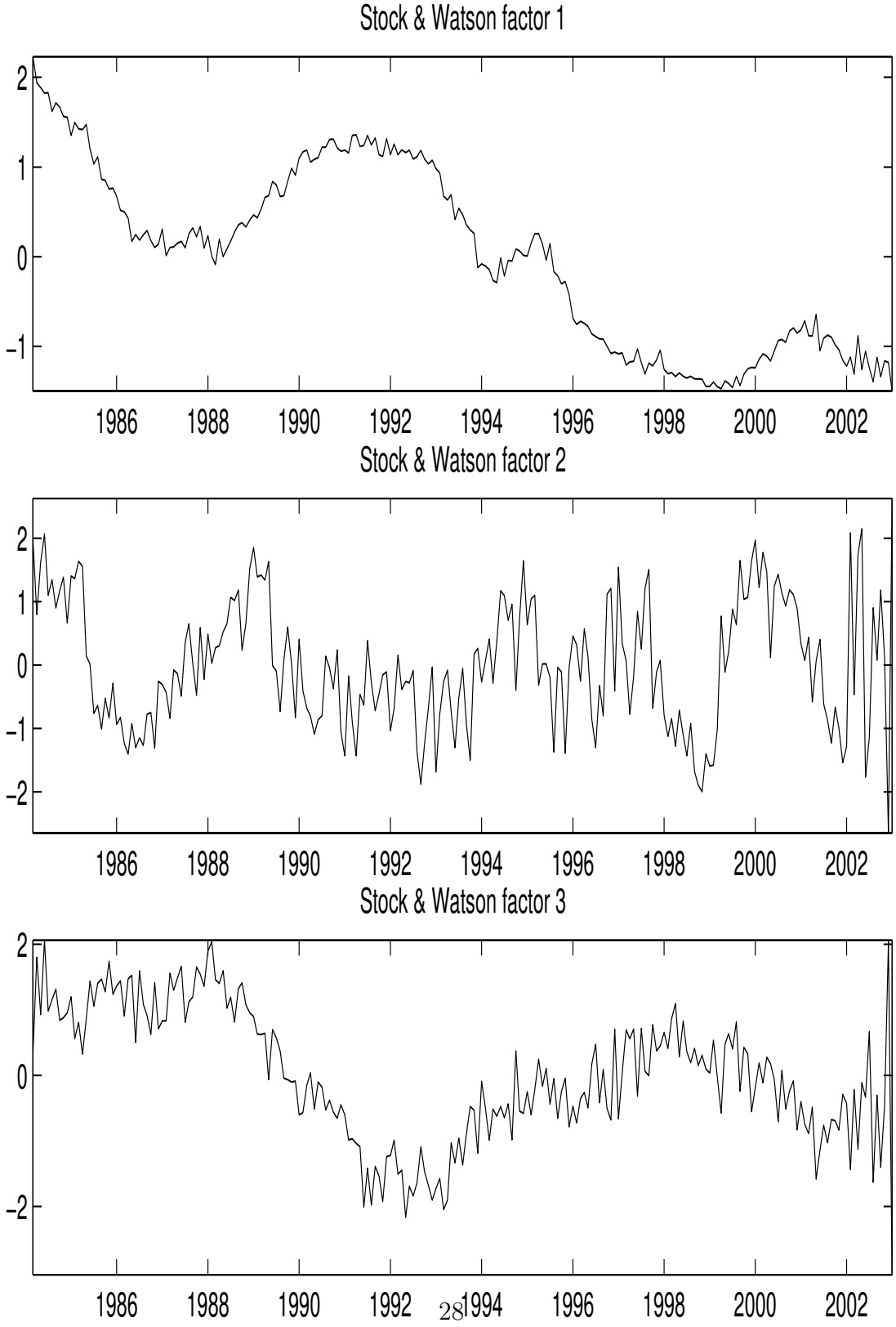

Figure 2: **Estimated structural factors**

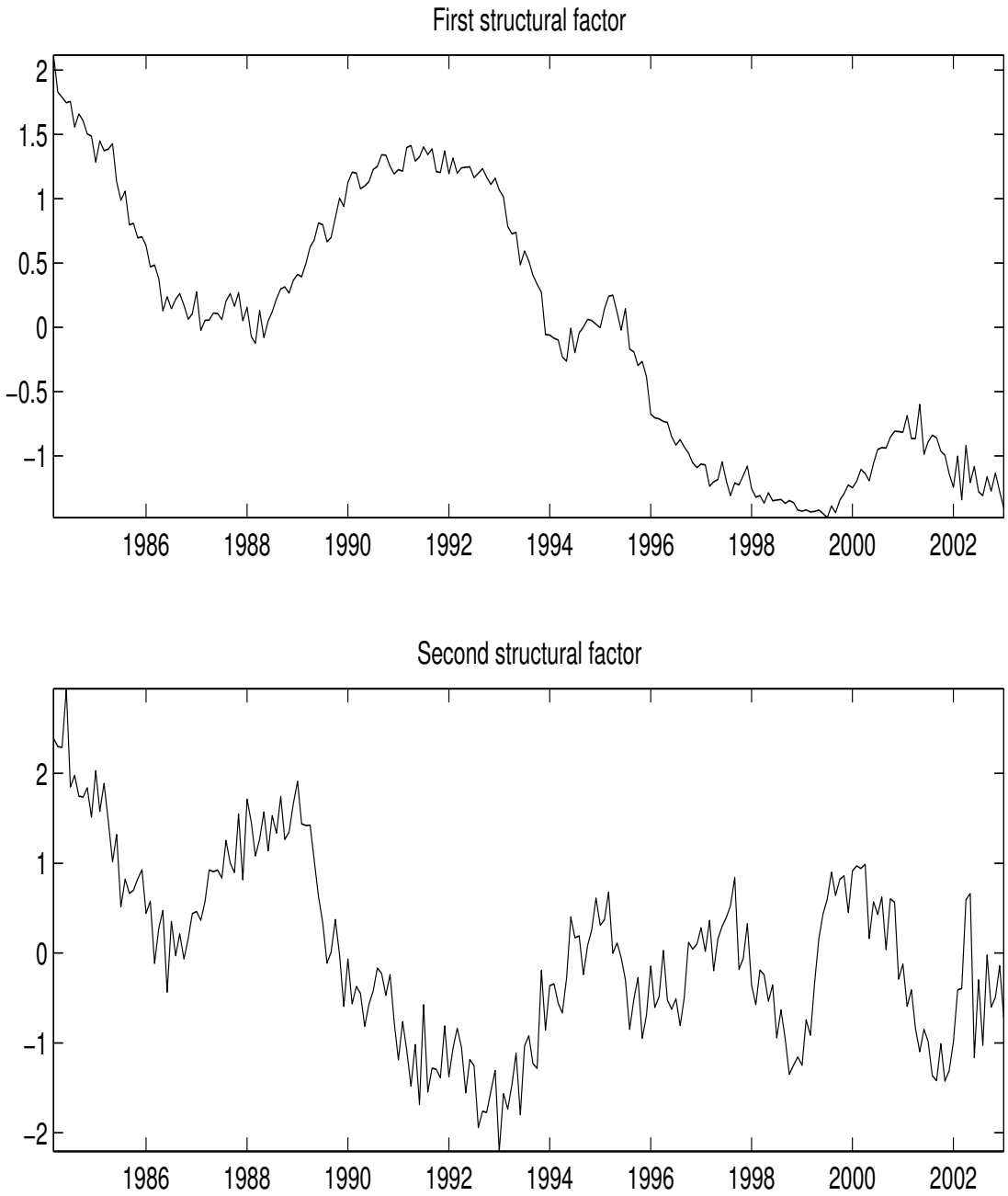

Figure 3: First structural factor and 3-month Euribor (France)

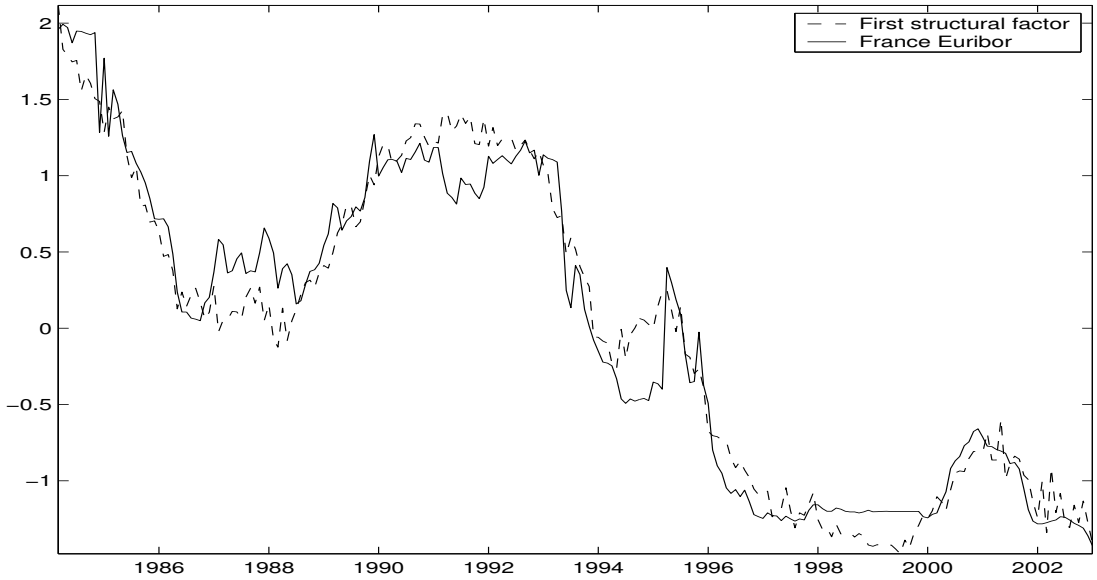

Figure 4: Second structural factor and PPI inflation (France)