ISSN 1745-8587

BCAM 2402

Maximum likelihood estimation of dynamic factor models using general cross sectional covariance

Tom Spencer
Birkbeck,University of London

May 2024

Birkbeck Centre for Applied Macroeconomics

▪ Birkbeck, University of London ▪ Malet Street ▪ London ▪ WC1E 7HX ▪

Maximum likelihood estimation of dynamic factor models using general cross sectional covariance

Tom Spencer*

May 31, 2024

Abstract

The existing literature on large dynamic factor models invariably assumes that the cross sectional covariance matrix is diagonal. This is due to the curse of dimensionality which means that many parameters need to be estimated for large data sets. This paper introduces a novel maximum likelihood approach which relaxes this diagonal assumption. All of the parameters are concentrated out so the parameters are jointly estimated with the factors. Importantly, the cross sectional covariance matrix is concentrated out so does not need to be explicitly estimated. The approach uses a neat simplification of the log-likelihood which makes estimation for large dimensional data feasible. Implementation of the general covariance approach is by numerical optimisation of the concentrated log-likelihood with respect to the factors. A diagonal version of the general covariance approach is also introduced, mainly for comparative reasons. Out of sample tests using Monte Carlo simulations shows the new general approach performs well, with smaller prediction errors overall compared to a range of existing diagonal approaches. Understandably, the general approach does particularly well for high cross sectional covariance. This is most apparent for low numbers of factors. This paper opens up the literature to new ways of estimating dynamic factor models and improvements in inference and forecasting for big data.

1 Introduction

With more data ever available, interest in modelling big data sets has grown accordingly and computing power has made this increasingly feasible. One branch of the big data literature is that of factor models, which has a broad set of uses, for example psychology, genetics or market research. The latent factors account for unobserved latent variables in the data which represent trends or traits. These factors are aimed at being maximally correlated with the observable variables. Principal components is by far the most common approach employed to estimate such models due to its simplicity. Factor models are generic models which do not require a structure to the data to be specified as in a typical panel data model although the cross sectional errors in panel data models are often modelled by factors.

In their usual static form factor models only model contemporaneous cross sectional relationships and so they are restrictive and not very useful for forecasting. The time series extension is dynamic factor models (DFM) where the factors themselves are dynamic, i.e autocorrelated. This means DFMs are able to capture dynamic interactions and are inherently useful for forecasting. In economics, DFMs are used for example in business cycle analysis, housing and inflation modelling and in finance some applications are CAPM and

*Birkbeck, University of London (tomspencer149@gmail.com)

yield curve modelling. The present paper adds to the large wealth of literature on factor models and their dynamic extension.

Factor models aim to capture a large portion of the correlation and variance within the data using the latent factors. The cross sectional errors of the individual series are usually assumed to be independent. This is a convenient assumption but is not realistic in all practical situations. For example in applications which have a sectoral structure, one may expect the error terms in a each sector to be correlated. Estimating the latent factors under this diagonality assumption is not ideal, since the estimated factors would try to naturally account for as much of the correlation within the data as possible, and some of this may in reality be coming from the cross sectional errors. Hence allowing the cross sectional errors to be correlated should give more appropriate factors, which in turn should make inference more robust and forecasting more accurate.

This paper introduces a novel approach which relaxes this diagonal assumption by concentrating out the parameters so the cross sectional covariance matrix is concentrated out and does not need to be explicitly estimated. The parameters are jointly estimated with the factors and they can be recovered once the new factors have been estimated using relevant formulae. The approach uses a simplification of the log-likelihood which makes estimation feasible for large cross sectional size. This simplification is common in the vector autoregression literature but is rare in the factor models literature.

The approach is implemented by numerical optimisation of the concentrated log-likelihood with respect to the factors, so does not require Kalman filter or Expectation Maximisation algorithms. As a result the approach is actually very simple and can be implemented with very little code. The results look promising for the direct optimisation approach with overall improved forecasting power in terms of root mean squared error of predictions compared to principal components. This is demonstrated using out of out of sample tests on simulated data sets.

The literature on DFMs largely assumes that the data and factors are I(0) and that the parameters are constant in time. These assumptions are maintained in the present paper for simplicity.

The remainder of the paper is set out as follows. Section 2 introduces notation and discusses existing popular approaches to estimating dynamic factor models. Section 3 explains the new general covariance approach and the diagonal equivalent, which is a restricted version of the general covariance approach. Section 4 describes the out of sample test set up and Section 5 outlines the results of the out of sample test and discusses them. Section 6 gives concluding remarks.

The present paper only considers the case where the time dimension, T, is larger than the cross sectional dimension, N. This is for convenience, and the approach can be easily extended to the case where $N > T$ as alluded to in Section 3.1. Also, testing on larger real world data sets, e.g. macroeconomic or financial data, is not covered in the present paper, but is anticipated to be covered in a later paper.

2 Background

2.1 Overview of DFMs

This section reviews the existing literature on DFMs, with particular regard to the diagonal cross sectional error covariance assumption. For simplicity, it is assumed that the factor dynamics are of autoregressive order 1 and that there are no lags of the factors in the cross sectional equation. The data generating process is assumed to be such that the cross section of N observed individuals x_t load onto the K unobserved factors f_t as follows:

$$x_t = \Lambda f_t + \eta_t, \tag{1}$$

for $t = 1, ..., T$ where x_t and η_t are $N \times 1$, Λ is $N \times K$ and f_t is $K \times 1$. η_t is assumed to be i.i.d., normally distributed $\eta_t \sim N(0, \Sigma)$ and independent of f_t. Estimation of Λ is feasible for large N from a practical point of view since only a few factors are used in the equation for each individual time series.

The factors are assumed to be dynamic and follow a vector autoregression (VAR):

$$f_t = \Phi f_{t-1} + \varepsilon_t, \tag{2}$$

where Φ is $K \times K$ and such that $\{f_t\}_{t=1,...,T}$ is stable, i.e. that the roots r of $|I_K - r\Phi| = 0$ lie outside the unit circle. ε_t is $K \times 1$, normally distributed as $\varepsilon_t \sim N(0, \Sigma_f)$ and uncorrelated with f_{t-1}. Σ and Σ_f are assumed to be positive definite. η_t and ε_t are usually assumed to have no serial correlation. η_t and ε_t are assumed to be independent as is standard. Eqs. (1) and (2) are referred to in this paper as the measurement and state equation respectively to align with state space modelling terminology.

Transposing and stacking Eqs. (1) and (2) gives:

$$X = F\Lambda' + \eta \tag{3}$$

and:

$$F = F^{-}\Phi' + \varepsilon, \tag{4}$$

where $X = (x_1, x_2 \dots x_T)'$and $(\eta_1, \eta_2 \dots \eta_T)'$ are $T \times N$, $\varepsilon = (\varepsilon_1, \varepsilon_2 \dots \varepsilon_T)'$, $F = (f_1, f_2 \dots f_T)'$ and F^{-} is the lagged F, i.e. $F^{-} = (f_0, f_1 \dots f_{T-1})'$. ε, F and F^{-} are all $T \times K$.

2.2 Estimation

Currently it seems the most accurate way of estimating large scale (i.e. large N) dynamic factor models is by quasi-maximum likelihood (QML) estimation such as Doz et al. (2012). The precursor to the QML approach is the two step approach of Doz et al. (2011). The two step approach uses initial estimates of the factors obtained by principal components (PCs). During the first step the parameters (including covariances) of Eqs. (1) and (2) are estimated by OLS, and then the second step uses the Kalman filter & smoother to obtain better estimates of the factors given the parameters. The QML approach extends this by iterating between calculating the factors given the parameters, then calculating the parameters given the factors, in an Expectation Maximisation (EM) algorithm. The EM algorithm in this context obtains better and better estimates of factors given parameters (E step) and parameters given factors (M step) until convergence.

The QML approach assumes that the covariance Σ of the idiosyncratic term η_t is diagonal (see assumption R2 of Doz et al. (2012)). The reason for this diagonal assumption is discussed below. Nonetheless, the QML approach is consistent for large N and T along any path of N and T tending to infinity. The rate of consistency for estimating the factors is $min(\sqrt{T}, \frac{n}{\log(n)})$(see Doz et al. (2012)). PCs are also consistent however the QML approach can provide efficiency improvements over PCs in finite samples.

2.3 Current reasoning on the diagonal assumption

For small N DFMs, maximum likelihood estimation using a general cross sectional covariance matrix is not an issue (see e.g. Diebold et al. (2021)). Large N DFMs however are known to suffer from the curse of

dimensionality. There is a wealth of literature that mentions that the diagonal assumption is adopted due to the very high number of parameters which need to be estimated, for example Poncela et al. (2021). However, in depth technical reasons for this are not easy to find. Overall, the literature indicates that the curse of dimensionality in DFMs is because firstly the QML type approaches need to invert $N \times N$ covariance matrices in the Kalman filter, and secondly that estimating the covariance matrices adds to the number parameters making estimation unfeasible. Some examples where this is discussed in the literature are given below.

Barigozzi (2018) states that Σ is assumed to be diagonal since there are problems with the Kalman smoother when inverting the matrix if N is large. This mis-specification does not affect consistency however may lead to loss of efficiency in finite samples. Indeed it is well known that in general the Kalman filter is subject to the curse of dimensionality since it involves taking the inverse of $N \times N$ terms such as $\Lambda P \Lambda' + \Sigma$ where P $(K \times K)$ is the estimated covariance of the state (see e.g. Hamilton (1994) Eq. 13.2.16). Taking the inverse of this $N \times N$ matrix during factor estimation is problematic for large N. Bai and Li (2016) states that the number of variables is comparable or even greater than the sample size, this is because the covariances are included as parameters which need to be estimated. Bai and Li (2012) shows consistency of quasi-maximum likelihood estimators under different identifying restrictions / normalisations in the context of static factor models. They assume diagonal idiosyncratic error covariance as the analysis would be too complex otherwise. Bai and Li (2016) adopts a similar approach but allowing for dynamics in the factors. They state that the model is not identifiable for large N with unrestricted Σ because the number of parameters to be estimated is too large. Doz et al. (2012) also states that parsimony is achieved by restricting to diagonal Σ, but that once this restriction is relaxed there is no obvious way to model the measurement equation correlation because there is no natural order to the correlations.

None of these methods are feasible for large N and a general Σ matrix. Bayesian approaches also seem to rely on Kalman type approaches or band (i.e sparse) matrices (see e.g. Chan et al. (2019) chapter 18) when estimating factors for example in Gibbs sampling. Attempts to incorporate non-diagonal idiosyncratic covariance matrix are not common in the literature. These approaches seem to focus on penalised likelihood methods (see Doz and Fuleky (2019) and Barigozzi and Luciani (2022)).

3 New general and diagonal covariance approaches

3.1 Likelihood derivation for the general approach

The general covariance approach aims to relax the assumption of diagonal Σ. This is desirable intuitively as some of the correlation in x_t may be attributed to the dynamic factors and some may be attributed to the measurement equation error term, which is not dynamic. The non-diagonal Σ aims more towards full maximum likelihood since fewer assumptions are required compared to existing approaches such as the Doz et al. (2012) quasi-maximum likelihood (QML) approach, which is named "quasi" due to the diagonal assumption. Non-diagonal Σ should lead to more optimal estimates of the factors, having a higher log-likelihood and leading to better forecasts in finite samples. The model assumptions are the same as Section 2 except that the approach here relaxes the assumption that Σ is diagonal.

The log-likelihood for the full sample can be calculated as follows. The likelihood is based on the innovations in x_t (i.e. $x_t - E(x_t|x_{t-1}, x_{t-2}, ..., x_1)$) which is common in the literature, see e.g. Watson and Engle (1983) and Doz et al. (2012). Substituting Eq. (2) into Eq. (1) gives:

$$x_t = \Lambda(\Phi f_{t-1} + \varepsilon_t) + \eta_t = \Lambda \Phi f_{t-1} + \Lambda \varepsilon_t + \eta_t, \tag{5}$$

so can be written:

$$x_t = \Lambda\Phi f_{t-1} + w_t, \tag{6}$$

where $w_t = \Lambda\varepsilon_t + \eta_t \sim N(0, \Omega)$ is $N \times 1$ and Ω is $N \times N$. Given initial estimates of the parameters, the covariance matrices can be estimated as

$$\hat{\Sigma} = \frac{1}{T}\hat{\eta}'\hat{\eta} = \frac{1}{T}(X - \hat{F}\hat{\Lambda}')'(X - \hat{F}\hat{\Lambda}') \tag{7}$$

and

$$\hat{\Sigma_f} = \frac{1}{T}\hat{\varepsilon}'\hat{\varepsilon} = \frac{1}{T}(\hat{F} - \hat{F}^{-}\hat{\Phi})'(\hat{F} - \hat{F}^{-}\hat{\Phi}). \tag{8}$$

Given the parameter and factor estimates, $\hat{w}_t$ is estimated as $\hat{w}_t = \hat{\Lambda}\hat{\varepsilon}_t + \hat{\eta}_t$ where $\hat{\eta}_t = x_t - \hat{\Lambda}\hat{f}_t$ and $\hat{\varepsilon}_t = \hat{f}_t - \hat{\Phi}\hat{f}_{t-1}$. The log-likelihood can then be written:

$$L(\hat{F}, \hat{\Lambda}, \hat{\Phi}) = \frac{1}{2}(T \log|\hat{\Omega}^{-1}| - TN\log(2\pi) - \sum_{t=1}^{T} \hat{w}_t'\hat{\Omega}^{-1}\hat{w}_t), \tag{9}$$

where $\hat{\Omega}$ is the the estimated $N \times N$ covariance matrix of w_t, i.e. $\hat{\Omega} = \frac{1}{T}\hat{w}'\hat{w}$ where $\hat{w} = \hat{\varepsilon}\hat{\Lambda}' + \hat{\eta}$ is $T \times N$. $\hat{\Omega}$ is assumed to be positive definite, hence the present paper only the covers the case where $T > N$.

It is assumed that ε and η are independent, i.e. orthogonal to each other. They could in principle be allowed to be correlated but the benefit of doing so is unclear and existing literature invariably assumes they are uncorrelated.

Under the assumption that ε and η are orthogonal to each other, Ω is estimated by

$$\hat{\Omega} = \frac{1}{T}\hat{w}'\hat{w} = \frac{1}{T}(\hat{\varepsilon}\hat{\Lambda}' + \hat{\eta})'(\hat{\varepsilon}\hat{\Lambda}' + \hat{\eta}) = \frac{1}{T}(\hat{\Lambda}\hat{\varepsilon}'\hat{\varepsilon}\hat{\Lambda}' + \hat{\eta}'\hat{\eta} + \hat{\Lambda}\hat{\varepsilon}'\hat{\eta} + \hat{\eta}'\hat{\varepsilon}\hat{\Lambda}') \tag{10}$$

$$= \hat{\Sigma} + \hat{\Lambda}\hat{\Sigma}_f\hat{\Lambda}' + \hat{\Lambda}\hat{\Sigma}_{\varepsilon\eta} + (\hat{\Lambda}\hat{\Sigma}_{\varepsilon\eta})' = \hat{\Sigma} + \hat{\Lambda}\hat{\Sigma}_f\hat{\Lambda}', \tag{11}$$

where $\hat{\Sigma}_{\varepsilon\eta} = \frac{1}{T}\hat{\varepsilon}'\hat{\eta}$ $(K \times N)$ is a matrix of the covariances between ε and η which is zero under the assumption that ε and η are orthogonal to each other. The last term of Eq. (9) can be simplified (see Hamilton (1994) Eqs. 11.1.32 & 11.1.33):

$$\sum_{t=1}^{T} \hat{w}_t'\hat{\Omega}^{-1}\hat{w}_t = trace(\sum_{t=1}^{T} \hat{w}_t'\hat{\Omega}^{-1}\hat{w}_t) \tag{12}$$

$$= trace(\sum_{t=1}^{T} \hat{\Omega}^{-1}\hat{w}_t\hat{w}_t') \tag{13}$$

$$= trace(\hat{\Omega}^{-1}(T\hat{\Omega})) \tag{14}$$

$$= trace(TI_N) = TN, \tag{15}$$

where I_N is the identity matrix. This simplification is well known in the context of vector autoregression, however it is rare for dynamic factor models. Eq. (14) is obtained since $\hat{\Omega} = \frac{1}{T}\hat{w}'\hat{w} = \hat{\Sigma} + \hat{\Lambda}\hat{\Sigma}_f\hat{\Lambda}'$ using the orthogonality assumption. The simplification is hence mathematically valid under the orthogonality assumption. Forming the likelihood under a certain null hypothesis in this way is standard.

Noting that $\log|\hat{\Omega}^{-1}| = -\log|\hat{\Omega}|$ the log-likelihood then simplifies to:

$$L(\hat{F}, \hat{\Lambda}, \hat{\Phi}) = \frac{1}{2}(-T\log|\hat{\Omega}| - TN\log(2\pi) - TN), \tag{16}$$

where $\hat{\Omega} = \hat{\Sigma} + \hat{\Lambda}\hat{\Sigma}_f\hat{\Lambda}'$. Note that the calculation of the log-likelihood in this way avoids the need to invert $\hat{\Omega}$.

Evaluation of Eq. (16) is not feasible when $N > T$. This is because the determinant of the $N \times N$ matrix $\hat{\Omega} = \frac{1}{T}\hat{w}'\hat{w}$ is zero since $\hat{\Omega}$ is of rank T which is less than N. The general covariance approach should be easily adapted to handle this by estimating $|\hat{\Omega}|$ as the product of the non-zero eigenvalues of $\hat{\Omega}$, as advocated by Srivastava and von Rosen (2002). The Moore-Penrose pseudo inverse can be used to show that the log-likelihood summation term simplification shown above still holds in this case with the difference that the summation term equals T^2 instead of TN. The $N > T$ case is not a focus of the present paper but would be an interesting topic for further research.

3.2 Likelihood derivation for the diagonal approach

The parameters can also be concentrated out using a diagonal cross sectional covariance matrix to form a restricted version of the general covariance approach above. This restriction is the only difference between the new general covariance approach and the new diagonal covariance approach. In the same way that the general covariance case assumes zero correlation between η and ε, so too does the diagonal equivalent.

Recall from section 3.1 above, we had (see Eq. (16)) for the full log-likelihood:

$$L(\hat{F}, \hat{\Lambda}, \hat{\Phi}) = \frac{1}{2}(-T\log|\hat{\Omega}| - TN\log(2\pi) - TN), \tag{17}$$

where Ω was estimated by

$$\hat{\Omega} = \hat{\Sigma} + \hat{\Lambda}\hat{\Sigma}_f\hat{\Lambda}'. \tag{18}$$

In the same way as the cross-equation error correlations were set to zero to implement the independent η and ε assumption in Eq. (11) for the general approach, the off diagonal elements of Ω can be set to zero to implement the diagonal assumption, i.e. Ω is estimated by

$$\hat{\Omega} = \hat{D} + \hat{\Lambda}\hat{\Sigma}_f\hat{\Lambda}', \tag{19}$$

where D is a diagonal matrix containing the diagonal elements of $\hat{\Sigma}$. The log-likelihood simplification (see Eqs. (12) - (15)) works out in a similar manner.

3.3 Concentrating out the parameters and solving for the factors

For both the new general covariance and diagonal covariance approaches, the coefficients are estimated by maximum likelihood. This enables concentrating out the parameters from the log-likelihood so that it only depends on the factors.

The formulae for the coefficients for the general covariance approach are:

$$\hat{\Lambda} = X'\hat{F}(\hat{\varepsilon}'\hat{\varepsilon} + \hat{F}'\hat{F})^{-1} \tag{20}$$

$$\hat{\Phi} = \hat{F}'\hat{F}^{-}(\hat{F}^{-\prime}\hat{F}^{-})^{-1} \tag{21}$$

So the formula for $\hat{\Phi}$ is the same as the OLS formula, but $\hat{\Lambda}$ includes an additional $\hat{\varepsilon}'\hat{\varepsilon}$ term which is an interaction term coming from the measurement and state equation being linked through the log-likelihood. The diagonal model also uses $\hat{\Phi} = \hat{F}'\hat{F}^{-}(\hat{F}^{-\prime}\hat{F}^{-})^{-1}$ but is more complicated for $\hat{\Lambda}$ which is solved for iteratively. See the Appendix for derivations and further details. Based on the estimated coefficients, the covariance matrices are also concentrated out using Eqs. (7) & (8)

Concentrating out is well known in optimisation problems in general and involves rewriting one or more of the parameters to be optimised as a function of a different variable. In the case here, the coefficient matrices are concentrated out as functions of the factors, by maximum likelihood as explained above. The covariance matrices are also concentrated out through the resulting residuals, so Ω itself is also concentrated out.

Mathematically,

$$L(\Omega(F, \Lambda, \Phi)) = L(\Omega(F, \Lambda(F), \Phi(F))) = l(\Omega(F)) = l(F), \tag{22}$$

where $l(F)$ is now the concentrated log-likelihood. The task at hand is to solve:

$$\hat{F}^{updated} = \arg\max_{F}(l(F)), \tag{23}$$

where $l(F)$ is as per Eq. (16) and $\hat{F}^{updated}$ is the new value of F given the initial (e.g. PCs) factors. The parameters can be recovered once the new factors have been estimated, using the maximum likelihood coefficients explained above.

An identifying normalisation is required in order to identify Eq. (3), since rotating the factors should have no impact on the resulting common component $F\Lambda'$. Mathematically, $F\Lambda' = FRR^{-1}\Lambda'$ for any invertible $K \times K$ matrix R. Identifying normalisations are common in factor estimation in general and there are many available (see e.g Bai and Li (2012) & Bai and Li (2016)). The new diagonal and general approaches here adopt the common normalisation $F'F/T = I_k$.

The optimisation algorithm used is Matlab's `fminunc` function. This is a gradient based method which uses finite difference derivates. `fminsearch`, which uses the derivative free simplex method, seems to give gives nearly identical results but `fminunc` is faster. The tolerance is set to 0.0001 in terms of the first-order optimality measure, which is the maximum absolute value in the gradient vector, i.e. the optimisation stops when the gradient is nearly zero.

4 Out of sample Monte Carlo simulation test set-up

4.1 Simulation model

The simulation model is based as much as is reasonably possible on Doz et al. (2012)[1]. The main difference is that Doz et al. (2012) has autocorrelation in the idiosyncratic term. No autocorrelation is assumed here because none of the new approaches outlined here account for this autocorrelation, and in reality if there were autocorrelation, more (lagged) factors would likely be added until there was minimal autocorrelation. The models would be mis-specified if there was significant autocorrelation in the simulation model. Hence the simulation model here has no autocorrelation of the idiosyncratic term.

[1] A similar set up is used in Stock and Watson (2002).

Factors f_t and data x_t are simulated according to the following sequence:

$$\Lambda_{ij} \sim i.i.d.N(0,1) \text{ for } i = 1...N, j = 1..K \tag{24}$$

$$\alpha_i = \frac{\beta_i}{1-\beta_i}\frac{1}{1-\rho^2}\sum_{j=1}^{k}\Lambda_{ij}^2 \text{ with } \beta_i \sim i.i.d.\ U(u, 1-u) \tag{25}$$

$$\mathcal{T}_{ij} = \tau^{|i-j|}(1-d^2)\sqrt{\alpha_i\alpha_j} \text{ for } i,j = 1...N \tag{26}$$

$$\Phi = diag(\rho) \tag{27}$$

$$f_t = \Phi f_{t-1} + \varepsilon_t \text{ where } e_t \ \sim i.i.d.\ N(0, I_k) \tag{28}$$

$$x_t = \Lambda f_t + \eta_t \text{ where } \ \eta_t \ \sim i.i.d.\ N(0, \mathcal{T}) \tag{29}$$

where $diag(\rho)$ is a diagonal matrix with all diagonal elements equal to ρ, $\mathcal{T}_{ij}$ is the i,j element of $\mathcal{T}$ and $U(u, 1-u)$ means uniformly distributed between u and $1-u$. Note that the simulated measurement and state equation coefficient matrices are Λ and Φ as defined by Eqs. (24) & (27) respectively. The simulation model is defined by the 4 parameters ρ, τ, u & k. The Toeplitz matrix $\tau^{|i-j|}$ defines the correlation matrix and the other two terms in the $\mathcal{T}_{ij}$ matrix are for scaling the covariance matrix according to signal to noise ratio parameters. For example, for $\tau = 0.5$ as used in the baseline test case below, the correlation $\tau^{|i-j|}$ of η_t is

$$\begin{pmatrix} 1 & \frac{1}{2} & \frac{1}{4} & \cdots & \frac{1}{2^{N-1}} \\ \frac{1}{2} & 1 & \frac{1}{2} & & \\ \frac{1}{4} & \frac{1}{2} & 1 & & \\ \vdots & & & \ddots & \\ \frac{1}{2^{N-1}} & & & & 1 \end{pmatrix}$$

and for $\tau = 0$ the correlation of η_t is the identity matrix. The u parameter controls the signal to noise ratio since β_i represents the ratio between the variance of the idiosyncratic component for individual i and the total variance in the data for individual i. For example for $u = 0.1$, as used in the test cases below, the β_i terms are uniformly distributed between 0.1 and 0.9 so the idiosyncratic component accounts for an average of 50% of the variance of the total data. The parameter ρ controls the autoregressive coefficients of the state equation and k is the number of dynamic factors.

The data is simulated 500 times (the same number as in Doz et al. (2012)) for each set of simulation parameters and the the correct number of factors is assumed to be known in advance.

Note that for a given τ as $N \to \infty$, more and more low correlations are added to the Toeplitz correlation matrix $\tau^{|i-j|}$ so the average off-diagonal correlation tends to zero. This is to ensure that the degree of cross sectional dependence (i.e. correlation) is weak. However, this means that the model in some sense tends towards a strict factor model error structure as $N \to \infty$. The direct approach, however should be robust when the degree of cross sectional dependence is strong.

4.2 Out of sample methodology

For each simulated data set the number of out of sample periods, T^{oos}, is chosen to be 30. In four of the results figures in Section 5.2, the in sample (i.e. calibration) period, T, is 100, and for one of the test cases, it is 200. For out of sample tests in general, sufficient out of sample data should be used in order to form meaningful conclusions. In machine learning, it seems 15-30% of the data is expected, and this is followed here. However the number of out of sample periods is fairly arbitrary here since as long as enough simulations are performed, very similar results should be produced if the number of out of sample periods is increased

because the parameters are all fixed. An expanding in sample window is chosen for advancing through the out of sample period. Note that forecasting is only ever performed one period ahead, which is in line with the structure of the models in this paper (i.e. the factors have autoregressive order 1).

PCs are used as initial factors for the diagonal and general covariance approaches. PCs requires that input data is standardised, in order not to bias the estimated factors due to differences in variance of the individual time series. The input data for the in sample PCs calculation for each out of sample point is standardised using the in sample mean and variance of the in sample period as required. Out of sample data is standardised according to leave-one-out standardisation, this is explained further along with reported metrics below.

4.3 Approaches tested

The following five approaches are compared:

- PCs plus VAR in the factors (labelled as "PCVAR" in the charts) - factors are estimated by PCs, then a VAR in the PC factors is formed using OLS.
- Two step approach (labelled as "TwoStep" in the charts) - the Doz et al. (2011) approach (see Section 2.2)
- QML (labelled as "QML" in the charts) - the Doz et al. (2012) approach (see Section 2.2)
- New diagonal correlation approach (labelled as "NewDiagonal" in the charts) - this is the new approach which assumes diagonal cross sectional covariance (see Section 3.2).
- New general correlation approach (labelled as "NewGeneral" in the charts) - this is the new approach which assumes general cross sectional covariance (see Section 3.1).

In summary, there are five approaches tested in this paper, three existing diagonal approaches, one new diagonal approach (which has the same assumptions as the existing diagonal approaches but a different estimation approach) and one new general covariance approach (which has a different estimation approach to the existing diagonal approaches and relaxes the diagonal assumption). Note that for the two step and QML approaches, the code was downloaded from the authors' website[2].

4.4 Reported metrics

The most relevant metric for the diagonal and general approaches is the log-likelihood, however out of sample likelihoods are not common in the literature. Instead the approaches are compared by the prediction errors. The metrics employed are not the objective function of any of the approaches tested, so are unbiased in the sense that they relate to no model in particular. They are simply a diagnostic tool which have obvious practical relevance.

In sample root mean squared prediction error (RMSPE): The in sample root mean square prediction error (RMSPE) of the predictions is formed by using the measurement and state dynamics equations. It is the equivalent of the standard root mean square error (RMSE) but in the context of the dynamic factor model structure. This is calculated as follows. First, given the model-estimated factors $\hat{f}_t$, the measurement equation and state equation parameters are obtained using the full data sample. Second, for $t = 1 \dots T$, $\hat{f}_{t-1}$

[2]See https://dataverse.harvard.edu/dataset.xhtml?persistentId=doi:10.7910/DVN/ZKNTUA

is used in the state dynamics equation (Eq. (2)) to get the model estimate of f_t given information available at $t-1$, denoted $\hat{f}_{t|t-1}$. Third, the estimated $\hat{f}_{t|t-1}$ (obtained via the second step) for $t = 1 \dots T$ are used to calculate an estimate of the data x_t according to the measurement equation (Eq. (1)). The resulting estimate of x_t uses only information up to time $t-1$ (aside from the parameter estimates which use the full data from the relevant in sample period). This estimate is denoted $\hat{x}_{t|t-1}$. Lastly, the final RMSPE is formed as the average over individuals i of the square root of the time average of the squared deviations of $\hat{x}_{t|t-1}$ and x_{t+1}. Mathematically, for a certain model:

$$RMSPE = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\frac{1}{T}\sum_{t=1}^{T}(X_{ti} - \hat{X}_{ti})^2}, \tag{30}$$

for $i = 1, \dots, N$ and $t = 1, \dots, T$ where:

$$\hat{X}_{ti} = (\hat{F}^{-}\hat{\Phi}\hat{\Lambda}')_{ti}, \tag{31}$$

which incorporates both the second and third step explained above in the calculation of $\hat{X}_{ti}$. In the same way as in earlier sections of this paper, $\hat{X}$ is the stacked version of of $\hat{x}_{t|t-1}$. What is reported in the results is the in sample RMSPE corresponding to the in sample period of the first out of sample point.

Out of sample root mean squared prediction error (RMSPE): This is the out of sample equivalent of the in sample RMSPE explained above. The input data for the in sample PCs calculation for each out of sample point is standardised using the in sample mean and variance of the in sample period as usual. Each out of sample point is standardised according to the mean and variance of the in sample period corresponding to the out of sample point. If the in sample period is 100, so the 101th data point is "standardised" using the mean and variance of the period $t = 1, \dots, 100$. The 102nd data point is standardised using the mean and variance of the period $t = 1, \dots, 101$ etc.. Each out of sample point within the out of sample period is standardised according to a slightly different mean and variance. This leave-one-out standardisation has the benefit of using only in sample data to calculate the mean and variance which is used for standardisation so is more aligned to what would be the case in reality if models were re-estimated on a periodic basis during the out of sample period. Also, the input data for calculation of PCs is always correctly standardised, so there is no bias due to differences in variance of the individual time series. The resulting standardised data for the out of sample period is denoted X^{oos} and is $T^{oos} \times K$.

The forecast for a certain out of sample point is formed by using the relevant in sample coefficient matrices multiplied by the estimated factors for just the last time point of the in sample period. Mathematically, for out of sample point t:

$$\hat{x}^{oos}_{t|t-1} = \hat{\Lambda}\hat{\Phi}\hat{f}_{t-1}, \tag{32}$$

where $\hat{f}_{t-1}$ is the estimated value of the factors at the end of the in sample period corresponding to out of sample point t, and $\hat{\Lambda}$ and $\hat{\Phi}$ are the coefficients calculated using the in sample period corresponding to out of sample point t. A separate calibration is used for each out of sample point within the out of sample period. The final RMSPE is formed as the average over individuals i of the square root of the time average of the squared deviations of $\hat{x}^{oos}_{t|t-1}$ and X^{oos}_{ti}. Mathematically, for a certain model:

$$RMSPE^{oos} = \frac{1}{N}\sum_{i=1}^{N}\sqrt{\frac{1}{T}\sum_{t=1}^{T^{oos}}(X^{oos}_{ti} - \hat{X}^{oos}_{ti})^2}, \tag{33}$$

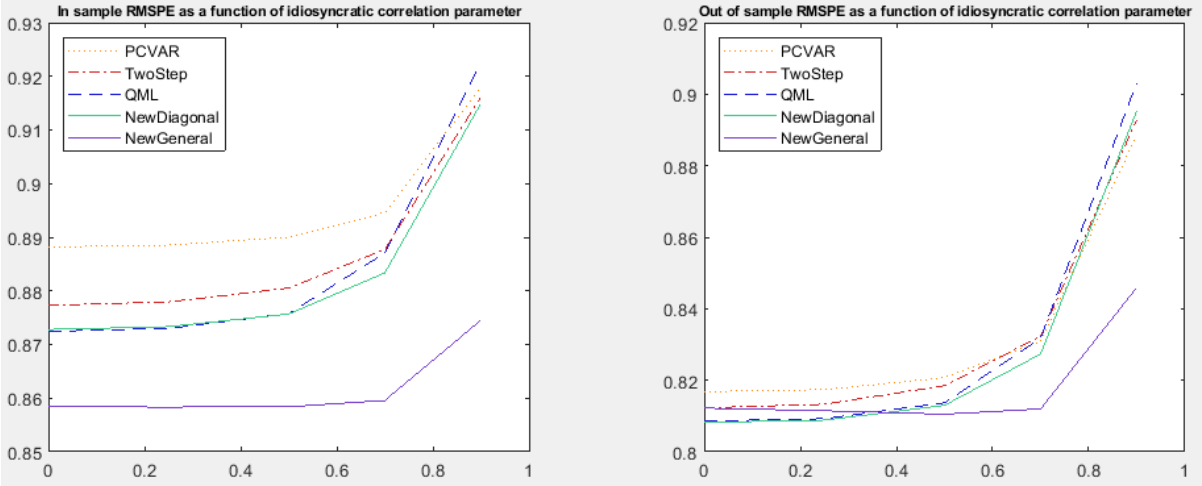
where $\hat{X}^{oos}$ is the stacked version of $\hat{x}^{oos}_{t|t-1}$.

5 Out of sample results and discussion

5.1 Overview of results charts

The out of sample test results are presented below. The figures plot the in sample and out of sample RMSPE metrics (which are explained above) on the y axis as a function of the idiosyncratic correlation parameter (τ in the simulation model) on the x axis. This is obviously the most interesting parameter to look at for the general covariance approach. The other set of simulation parameters are set as $\rho = 0.9$ and $u = 0.1$, these are the same as some of the test cases in Doz et al. (2012). The values of τ (the x axis in the charts) are $0, 0.25, 0.5, 0.7$ and 0.9.[3] For figures 1-3, the in sample T is 100 and N is 10 which is line with Doz et al. (2012). The number of simulations per simulation run is 500 which is the same as in Doz et al. (2012). The results are produced for $K = 1, 2$ and 3, where K is the number of factors both in the simulation and used by the models (see figures 1-3). Also included are the cases for $K = 3$ but with $N = 20$ and in sample $T = 200$ respectively, so N and T are each doubled in turn (see figures 4 and 5).

Figure 1: Results for K = 1



[3] 0.9 is perhaps of less practical relevance as if the cross sectional correlation seems to be that high, it is likely a practitioner would add an extra factor to try to absorb some of the correlation.

Figure 2: Results for K = 2

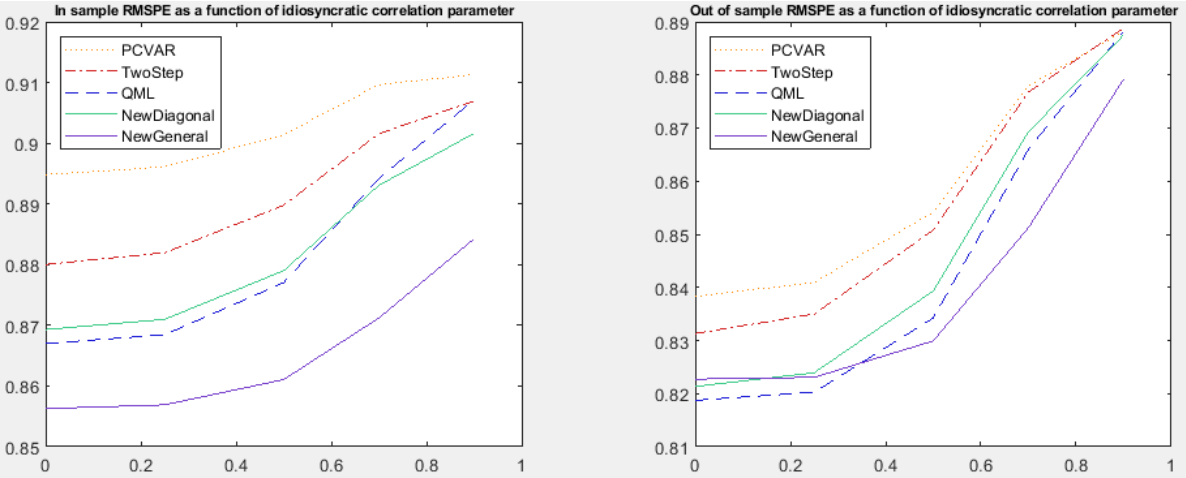


Figure 3: Results for K =3

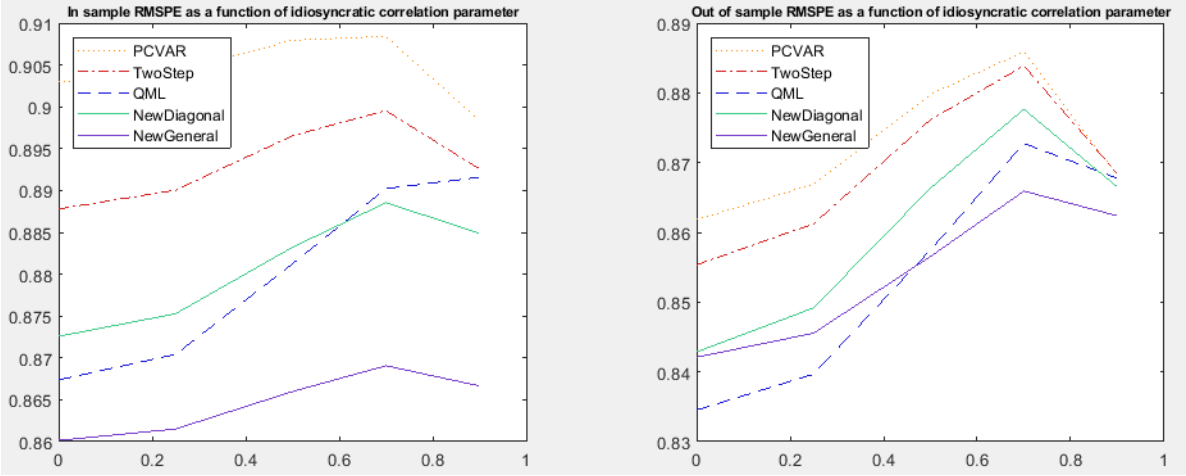


Figure 4: Results for $\mathbf{K = 3, N = 20, T = 100}$

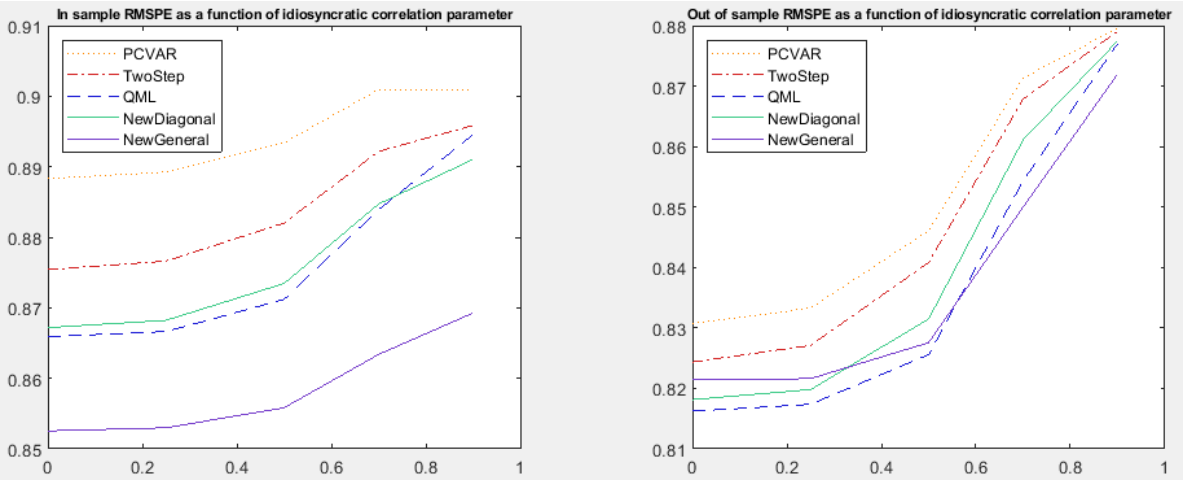
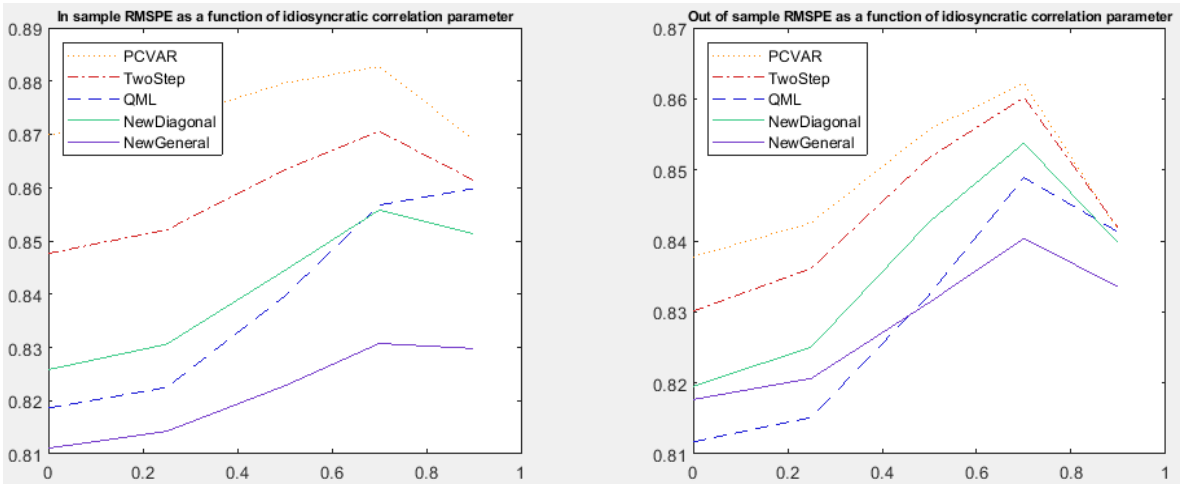


Figure 5: Results for $\mathbf{K = 3, N = 10, T = 200}$



5.2 Discussion of results

From a theoretical point of view, note that the focus here should be on comparing the new diagonal approach against the new general approach, since these are like for like, the only difference being that the latter allows for general cross sectional covariance. The existing three diagonal approaches (PCVAR, TwoStep and QML) are only included for comparing these new approaches to existing approaches.

The in sample RMSPEs for the new diagonal and general approaches are around the same compared to out of sample RMSPEs. The shape of these graphs are similar as well. These features indicate well a well specified model although one must bear in mind that RMSPE is not the objective function of any of the approaches, it is simply a diagnostic tool which has practical relevance. The combination low in sample RMSPEs and lack of overfitting suggest that the new general approach should be very useful for inference.

For the *in sample* RMSPE, QML is usually better (i.e. lower) than TwoStep which in turn is better than PCVAR. The new diagonal approach is around the same as QML and the general approach does best.

For the *out of sample* RMSPE, QML is usually better than TwoStep which in turn is better than PCVAR. The new diagonal approach is usually between QML and TwoStep but closer to QML. The new general approach is usually slightly worse than the new diagonal approach for very low values of the idiosyncratic correlation parameter (τ in the simulation model), but does much better for high values of τ, most notably for $K = 1$ (see Figure 1). This makes sense because in general, approaches which relax a certain restriction should perform better when the restriction is invalid. Similarly, the new general approach is worse than QML for low values of τ, most notably for $K = 3$ (see Figure 3) but better than QML for high values of τ, most notably for $K = 1$. The new diagonal and the new general approaches perform less well overall compared to QML as the number of factors increases. This may plausibly be due to the approaches being less able to distinguish between the additional factor and cross sectional correlation, since each factor accounts for less of the total common component as more factors are added.

The $N = 20$ case (Figure 4) looks very similar to the $N = 10$ case (see Figure 3) but the new general approach does slightly worse relative to the other approaches for medium to high values of τ compared to when $N = 10$. This is likely because as N increases, the average of the off diagonal correlations decreases because of the Toepliz structure of the simulation model (see the comments at the end of Section 4.1). This means there is less cross sectional correlation for the general approach to utilise. The $T = 200$ case (see Figure 5) shows that the general model does slightly better relative to the other approaches compared to the $K = 3$ case, which has $T = 100$. This makes sense because more general approaches usually benefit from more historic data for estimation.

6 Conclusion

The main contribution of this paper to the literature is the introduction of a novel maximum likelihood approach which relaxes the assumption of diagonal cross sectional covariance matrix. The parameters are concentrated out so are jointly estimated along with the factors. Importantly, the cross sectional covariance matrix is also concentrated out so does not need to be explicitly estimated. The approach uses a neat simplification of the log-likelihood which makes estimation for large dimensional data feasible. Implementation of the general covariance approach is by numerical optimisation of the concentrated log-likelihood with respect to the factors. Also introduced is an equivalent model which assumes diagonal covariance, mainly for comparative reasons.

Out of sample tests using Monte Carlo simulations show the new general covariance approach performs well, with smaller prediction errors overall compared to a range of existing diagonal approaches. Understandably, the general approach does particularly well for high cross sectional covariance. This is most apparent for low numbers of factors. The general correlation approach benefits from a longer of history of data. This makes sense because more general approaches usually require more historic data for estimation.

This paper opens up the literature to new ways of estimating dynamic factor models and further improvements in inference and forecasting for big data. The approach introduced here is in its infancy so there are likely many improvements which could be made. There are also many possible extensions, for example allowing for more lags in the factor dynamics or new approaches to estimate the number of factors taking into account the cross sectional correlation. Testing on larger real world data sets, e.g. macroeconomic or financial data would be interesting and is anticipated to be covered in a subsequent paper.

7 Appendix - Concentrating out the coefficients using maximum likelihood

7.1 Cross sectional coefficient, Λ

The coefficients are estimated by the first order conditions of the log-likelihood. Since we are looking at a maximum, we can ignore the log and only need to find the $N \times K$ matrix Λ which sets $\frac{\partial|\Omega|}{\partial\Lambda}$ to zero. A useful result here is the Jacobi formula that the derivative of the determinant is the adjoint. Mathematically, $\frac{\partial|M(q)|}{\partial q} = trace(adj(M(q))\frac{\partial M(q)}{\partial q})$ where M is a square matrix and q is a scalar (see e.g. Magnus and Neudecker (1999)). In the case where M is symmetric, the adjoint is the same as the cofactor matrix, since the adjoint is the transpose of the cofactor matrix. Note that in this Appendix, the hat accents denoting estimates are dropped for notational convenience.

For the general covariance case, we have, by the chain rule, where $i = 1, \dots N$ and $j = 1, \dots, K$:

$$\frac{\partial|\Omega|}{\partial\Lambda_{ij}} = trace(\frac{\partial|\Omega|}{\partial\Omega}\frac{\partial\Omega}{\partial\Lambda_{ij}}) = trace(C\frac{\partial\Omega}{\partial\Lambda_{ij}}) \tag{34}$$

$$= \frac{1}{T}trace(C\frac{\partial}{\partial\Lambda_{ij}}(\eta'\eta + g'g)), \tag{35}$$

where C is the cofactor matrix of Ω. Looking first at the $g'g$ part, where $g = \varepsilon\Lambda'$:

$$\frac{\partial g'g}{\partial\Lambda_{ij}} = \frac{\partial}{\partial\Lambda_{ij}}(\Lambda\varepsilon'\varepsilon\Lambda') = J_{ij}\varepsilon'\varepsilon\Lambda' + \Lambda\varepsilon'\varepsilon{J_{ij}}', \tag{36}$$

where J_{ij} is a $N \times K$ matrix a matrix of zeros except with a one in the i, j position. For $\eta'\eta$:

$$\frac{\partial}{\partial\Lambda_{ij}}(\eta'\eta) = \frac{\partial}{\partial\Lambda_{ij}}((X - F\Lambda')'(X - F\Lambda')) \tag{37}$$

$$= \frac{\partial}{\partial\Lambda_{ij}}(X'X + \Lambda F'F\Lambda' - \Lambda F'X - X'F\Lambda') \tag{38}$$

$$= J_{ij}F'F\Lambda' + \Lambda F'F{J_{ij}}' - J_{ij}F'X - X'F{J_{ij}}'. \tag{39}$$

Putting this together, and noting that some of the terms are the transpose of each other so are equivalent once the trace is taken:

$$\frac{\partial|\Omega|}{\partial\Lambda_{ij}} = \frac{2}{T}trace(C(\Lambda\varepsilon'\varepsilon{J_{ij}}' + \Lambda F'F{J_{ij}}' - X'F{J_{ij}}')). \tag{40}$$

It can easily be seen that arranging these elements into an $N \times K$ matrix gives:

$$\frac{\partial|\Omega|}{\partial\Lambda} = \frac{2}{T}C(\Lambda\varepsilon'\varepsilon + \Lambda F'F - X'F). \tag{41}$$

Solving for Λ, we obtain:

$$\Lambda = X'F(\varepsilon'\varepsilon + F'F)^{-1}. \tag{42}$$

In the diagonal case, $\Omega = \frac{1}{T}(diag(\eta'\eta) + g'g) = \frac{1}{T}(I_N \circ (\eta'\eta) + g'g)$, where the $diag()$ operator sets the off-diagonal elements to zero, and $\circ$ is the Hadamard or element-wise product and I_N is the $N \times N$ identity

matrix. Similarly:

$$\frac{\partial|\Omega|}{\partial\Lambda_{ij}} = \frac{2}{T}trace(C(\Lambda\varepsilon'\varepsilon {J_{ij}}' + I_N \circ (\Lambda F' F {J_{ij}}' - X' F {J_{ij}}'))). \tag{43}$$

Unfortunately it seems this does not have a straightforward analytical solution like for the general covariance case due to the element-wise product. However instead each $\frac{\partial|\Omega|}{\partial\Lambda_{ij}}$ is solved for separately holding all of the other Λ_{ij} fixed. This is iterated until reasonable convergence.

In order to solve for each $\frac{\partial|\Omega|}{\partial\Lambda_{ij}}$ in the diagonal case, the trace terms are written in matrix index notation:

$$\frac{\partial|\Omega|}{\partial\Lambda_{ij}} = \frac{2}{T}trace(C(\Lambda\varepsilon'\varepsilon {J_{ij}}' + I_N \circ (\Lambda F' F {J_{ij}}' - X' F {J_{ij}}'))) \tag{44}$$

$$= \frac{2}{T}\sum_{n=1}^{N}(C(\Lambda\varepsilon'\varepsilon {J_{ij}}' + I_N \circ (\Lambda F' F {J_{ij}}' - X' F {J_{ij}}')))_{nn} \tag{45}$$

$$= \frac{2}{T}\sum_{n=1}^{N}(C(J_{ij}\varepsilon'\varepsilon\Lambda' + I_N \circ (J_{ij} F' F\Lambda' - X' F {J_{ij}}')))_{nn} \tag{46}$$

$$= \frac{2}{T}\sum_{n=1}^{N}\{-(C \circ X' F {J_{ij}}')_{nn} + \sum_{m=1}^{K}(C J_{ij}\varepsilon'\varepsilon)_{nm}\Lambda_{nm} + \sum_{m=1}^{K} C_{nn}(J_{ij} F' F)_{nm}\Lambda_{nm}\}. \tag{47}$$

This is linear in Λ_{ij} and the only parts of $\frac{\partial|\Omega|}{\partial\Lambda_{ij}}$ which depend on Λ_{ij} are contained in the double summation terms, where $n = i$ and $m = j$, i.e. $\frac{2}{T}(C(J_{ij}\varepsilon'\varepsilon))_{ij}\Lambda_{ij}$ and $\frac{2}{T}C_{ii}(J_{ij}F'F)_{ij}\Lambda_{ij}$. Hence $\frac{\partial|\Omega|}{\partial\Lambda_{ij}}$ can be written as $\frac{\partial|\Omega|}{\partial\Lambda_{ij}} = A + B\Lambda_{ij}$ where:

$$B = \frac{2}{T}(C(J_{ij}\varepsilon'\varepsilon))_{ij} + \frac{2}{T}C_{ii}(J_{ij}F'F)_{ij}. \tag{48}$$

A could be calculated as the sums of the remaining terms but instead, for speedier calculation, A is calculated here as $A = \frac{\partial|\Omega|}{\partial\Lambda_{ij}} - B\Lambda_{ij}$. When A is calculated in this way it does not depend on Λ_{ij} because by construction $B\Lambda_{ij}$ contains all the parts of $\frac{\partial|\Omega|}{\partial\Lambda_{ij}}$ which depend on Λ_{ij}. Finally, the solved-for value of Λ_{ij} can then be calculated as $\Lambda_{ij} = -\frac{A}{B}$.

7.2 State coefficient, Φ

Similarly, with respect to element i, j of Φ, where $i = 1, \dots K$ and $j = 1, \dots, K$, noting that η does not depend on Φ:

$$\frac{\partial|\Omega|}{\partial\Phi_{ij}} = \frac{1}{T}\frac{\partial}{\partial\Phi_{ij}}|\eta'\eta + g'g| = \frac{1}{T}\frac{\partial}{\partial\Phi_{ij}}|\Lambda(F - F^{-}\Phi)'(F - F^{-}\Phi)\Lambda'| \tag{49}$$

$$= \frac{1}{T}\frac{\partial}{\partial\Phi_{ij}}|\Lambda(F'F + \Phi F^{-\prime}F^{-}\Phi' - \Phi F^{-\prime}F - F'F^{-}\Phi')\Lambda'| \tag{50}$$

$$= \frac{1}{T}trace(C\Lambda(J_{ij}F^{-\prime}F^{-}\Phi' + \Phi F^{-\prime}F^{-}{J_{ij}}' - J_{ij}F^{-\prime}F - F'F^{-}{J_{ij}}')\Lambda') \tag{51}$$

$$= \frac{2}{T}trace(C\Lambda(\Phi F^{-\prime}F^{-}{J_{ij}}' - F'F^{-}J_{ij}')\Lambda') \tag{52}$$

$$= \frac{2}{T}trace(\Lambda' C\Lambda\Phi F^{-\prime}F^{-}{J_{ij}}' - \Lambda' C\Lambda F'F^{-}J_{ij}'), \tag{53}$$

where J_{ij} is a 3×3 matrix of zeros except a 1 in the i, j position. As before, arranging the i, j elements in to a $K \times K$ matrix gives:

$$\frac{\partial |\Omega|}{\partial \Phi} = \Lambda' C \Lambda \Phi F^{-\prime} F^{-} - \Lambda' C \Lambda F' F^{-}. \tag{54}$$

Finally, solving this for Φ:

$$\Phi = F' F^{-} (F^{-\prime} F^{-})^{-1}, \tag{55}$$

which is the OLS estimator. Note that because the derivative of the $\eta'\eta$ term with respect to Φ is zero, the same formula for Φ applies for both the general and diagonal approaches.

8 Acknowledgements

I am grateful for comments by Zacharias Psaradakis, Ron Smith, Walter Beckert and Stephen Wright which have substantially improved this paper.

References

Jushan Bai and Kunpeng Li. Statistical analysis of factor models of high dimension. *The Annals of Statistics*, 40(1):436 – 465, 2012. doi: 10.1214/11-AOS966. URL https://doi.org/10.1214/11-AOS966.

Jushan Bai and Kunpeng Li. Maximum likelihood estimation and inference for approximate factor models of high dimension. *The Review of Economics and Statistics*, 98(2):298–309, 2016. ISSN 00346535, 15309142. URL http://www.jstor.org/stable/43830349.

Barigozzi and Luciani. Quasi maximum likelihood estimation and inference of large approximate dynamic factor models via the em algorithm. *arXiv:1910.03821*, 2022.

Matteo Barigozzi. Dynamic factor models. *Lecture notes, available at http://www.barigozzi.eu/*, 2018.

Joshua Chan, Gary Koop, Dale J. Poirier, and Justin L. Tobias. *Bayesian Econometric Methods*. Econometric Exercises. Cambridge University Press, 2 edition, 2019. doi: 10.1017/9781108525947.

Francis X. Diebold, Maximilian Göbel, Philippe Goulet Coulombe, Glenn D. Rudebusch, and Boyuan Zhang. Optimal combination of arctic sea ice extent measures: A dynamic factor modeling approach. *International Journal of Forecasting*, 37(4):1509–1519, 2021. ISSN 0169-2070. doi: https://doi.org/10.1016/j.ijforecast.2020.10.006. URL https://www.sciencedirect.com/science/article/pii/S0169207020301606.

Catherine Doz and Peter Fuleky. Dynamic factor models. Working Papers 2019-4, July 2019. URL https://ideas.repec.org/p/hae/wpaper/2019-4.html.

Catherine Doz, Domenico Giannone, and Lucrezia Reichlin. A two-step estimator for large approximate dynamic factor models based on kalman filtering. *Journal of Econometrics*, 164(1):188–205, 2011. URL https://EconPapers.repec.org/RePEc:eee:econom:v:164:y:2011:i:1:p:188-205.

Catherine Doz, Domenico Giannone, and Lucrezia Reichlin. A quasi—maximum likelihood approach for large, approximate dynamic factor models. *The Review of Economics and Statistics*, 94(4):1014–1024, 2012. ISSN 00346535, 15309142. URL http://www.jstor.org/stable/23355337.

James D. Hamilton. Time series analysis : James D. Hamilton, 1994, (Princeton University Press, Princeton, NJ), 799 pp., US $55.00, ISBN 0 - 691 - 04289 - 6. 11(3) : 494 - -495, *September* 1994. *URL* https://ideas.repec.org/a/eee/intfor/v11y1995i3p494-495.html.

Jan R. Magnus and Heinz Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics*. John Wiley, second edition, 1999. ISBN 0471986321 9780471986324 047198633X 9780471986331.

Pilar Poncela, Esther Ruiz, and Karen Miranda. Factor extraction using kalman filter and smoothing: This is not just another survey. *International Journal of Forecasting*, 37, 03 2021. doi: 10.1016/j.ijforecast.2021.01.027.

M. Srivastava and Dietrich von Rosen. Regression models with unknown singular covariance matrix. *Linear Algebra and its Applications*, 354:255–273, 10 2002. doi: 10.1016/S0024-3795(02)00342-7.

James H. Stock and Mark W. Watson. Forecasting using principal components from a large number of predictors. *Journal of the American Statistical Association*, 97(460):1167–1179, 2002. ISSN 01621459. URL http://www.jstor.org/stable/3085839.

Mark W. Watson and Robert F. Engle. Alternative algorithms for the estimation of dynamic factor, mimic and varying coefficient regression models. *Journal of Econometrics*, 23 (3):385–400, 1983. ISSN 0304-4076. doi: https://doi.org/10.1016/0304-4076(83)90066-0. URL https://www.sciencedirect.com/science/article/pii/0304407683900660.