PREDICTION OF SINGULAR VARS AND AN APPLICATION TO GENERALIZED DYNAMIC FACTOR MODELS

SIEGFRIED HÖRMANNa AND GILLES NISOLb

a Institute of Statistics, Graz University of Technology, Graz, Austria
b ECARES, Université libre de Bruxelles, Brussels, Belgium

Vector autoregressive processes (VARs) with innovations having a singular covariance matrix (in short singular VARs) appear naturally in the context of dynamic factor models. Estimating such a VAR is problematic, because the solution of the corresponding equation systems is numerically unstable. For example, if we overestimate the order of the VAR, then the singularity of the innovations renders the Yule-Walker equation system singular as well. We are going to show that this has a severe impact on accuracy of predictions. While the asymptotic rate of the mean square prediction error is not impacted by this problem, the finite sample behaviour is severely suffering. This effect will be reinforced, if the predictor variables are not coming from the stationary distribution of the process, but contain additional noise. Again, this happens to be the case in context of dynamic factor models. We will explain the reason for this phenomenon and show how to overcome the problem. Our numerical results underline that it is very important to adapt prediction algorithms accordingly.

Received 16 May 2020; Revised 30 September 2020; Accepted 27 October 2020

Keywords: Dynamic factor model; forecasting; regularization; vector autoregressive processes.

JEL. Primary 62M15; 62G10; secondary 60G15; 62G20.

1. INTRODUCTION

The principal motivation of our article initiates from an application and numerical implementation of a recent method by Forni et al. (2015) – here denoted by FHLZ – for estimation and forecasting of generalized dynamic factor models (GDFMs). We noticed that predictions can be improved when we slightly adapt the original forecasting algorithm. In essence, our observation was that a regularization step in the estimation part is very important. In this article, we are going to see that this owes to an ‘explosive’ combination of two features of FHLZ:

(a) The method relies crucially on predicting singular VAR systems, that is, vector autoregressive processes with a singular noise covariance.

(b) These singular VAR systems are loaded with noisy predictor variables. We cannot directly plug in lagged values of the process to forecast, since the lagged values are latent and need to be estimated. This perturbation – even if it is very small and asymptotically negligible – has a huge impact on the forecasts, especially in combination with (a).

While improving FHLZ was the initial motivation for our research, we believe that our theory has some more general interest in time series forecasting. Hence the main theme of this article is not restricted to predicting or modelling GDFMs. Rather the primary contribution of this article is an analysis of how noisy predictors impact the quality of VAR forecasts (see Section 2). Using these findings we will then – as an important application – improve the FHLZ algorithm.

* Correspondence to: Siegfried Hörmann, Institute of Statistics, Graz University of Technology, Kopernikusgasse 24/III, 8010 Graz, Austria. Email: shoermann@tugraz.at

© 2020 The Authors. Journal of Time Series Analysis published by John Wiley & Sons Ltd.
This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.
To fix ideas, let us consider a generic \( d \)-dimensional VAR(1) setting, where \( Y_{t+1} = \Theta Y_t + Z_{t+1} \). Here \( (Z_t) \) is some white noise sequence, potentially with singular covariance. Let \( \hat{\Theta} \) be some estimator (e.g. the OLS estimator) for the coefficient matrix \( \Theta \). Then we are interested in the prediction error which, on the one hand, is caused by estimation of \( \Theta \) and which, on the other hand, is due to an additional perturbation \( \epsilon \in \mathbb{R}^d \) of the predictor variable \( Y_t \):

\[
P(h) = \hat{\Theta}(Y_t + \epsilon) - \Theta Y_t.
\]

Notice that \( \epsilon \) is denoting some generic noise which might also depend on \( t \). Possible time dependence is not important for our theory, so it will typically be ignored. When \( \epsilon = 0 \) then the difference in (1.1) corresponds to the difference between the ‘usual’ prediction and the optimal (but infeasible) prediction \( \Theta Y_t \). In this article, we focus on one-step-ahead forecasting – this setting being easily generalized.

As we indicated above, noisy predictors are inherent in the GDFM framework (see Section 3.2) and even very small noise can have a dramatic impact on the quality of the forecasts. However, noisy predictors are also not uncommon in other context and to the best of our knowledge this problem has not been exploited in the literature. We give an illustrative example of data that we have been working with in the Appendix C.

The rest of the article is structured as follows: in Section 2 we outline the theory for singular VARs. In Section 3 some important background on GDFMs is provided. Moreover, we explain the FHLLZ prediction method. At the end of this section we point out how the theory from Section 2 is related to this method. Then, in Section 4, we present some numerical exercises which illustrate the dramatic improvement our approach can yield. Proofs, details of the implementation and several tables are provided in the Appendix.

## 2. SINGULAR VAR PROCESSES

We review some basics of VAR’s and explain how singularity of the noise impacts their estimation and prediction.

### 2.1. The Yule-Walker Equations

Let us consider some generic \( d \)-dimensional VAR(p) process of the form

\[
Y_{t+1} = (\Theta_1 Y_t + \cdots + \Theta_p Y_{t-p+1}) = Z_{t+1},
\]

where \( Z_t \sim \text{WN}(0, \Omega) \) (white noise). Denote by \( \gamma(h) \) the lag-\( h \) autocovariance matrix. We consider the case where the autoregression is singular, hence where \( |\Omega| = 0 \). Chen et al. (2011) have studied stationary solutions for this class of processes and have shown a number of interesting and subtle properties of the solutions of the related Yule-Walker equations. Let us briefly recall some basic facts related to the Yule-Walker equations which serve also to fix notation. The equations are obtained by multiplying (2.1) by \( Y'_{t+1-h} \), \( h = 0, \ldots, p \) and taking expectations. Defining

\[
\begin{align*}
Y_{t,p} &= (Y'_t, \ldots, Y'_{t-p+1})' \quad \text{and} \quad \Gamma_p = \text{Cov}(Y_{t,p}),
\end{align*}
\]

we can summarize them as

\[
(I_d - \Theta_p)\Gamma_{p+1} = (\Omega, O, \ldots, O),
\]

with \( I_d \) being the identity matrix in \( \mathbb{R}^d \) and \( \Theta_p = (\Theta_1, \ldots, \Theta_p) \). (Here and in the sequel boldface characters are used when we stack vectors or juxtapose matrices.) In particular, (2.2) yields the sub-equations \( \Theta_p = \gamma_p \Gamma_p^{-1} \) and \( \Omega = \gamma(0) - \gamma_p \Gamma_p^{-1} \gamma'_p \). Replacing the population covariances with the empirical ones, gives rise to the well-known Yule-Walker moment estimators for \( \Theta_p \) and \( \Omega \). In particular, when \( p = 1 \) the equations simplify to \( \Theta = \gamma(1)\gamma^{-1} \) and \( \Omega = \gamma(0) - \gamma(1)\gamma(0)^{-1} \gamma(1)' \).
Moreover, if Proposition 2.1. in the equation system may have a severe impact on the accuracy of the solution. VAR system, and the resulting Yule-Walker equations will be numerically unstable – small approximation errors respectively, of some square matrix framework consists of 1,000 Monte Carlo experiments for each combination of \((p,d)\). We proceed as follows:

1. Generate a \((d \times p d)\)-matrix \((\Theta_0), \ldots, \Theta_{\Theta_0}\) with i.i.d. standard normal entries. Here \(\Theta_{\Theta_0}\) are blocks of dimension \(d \times d\). Define \(\Theta_i = \frac{\Theta_{\Theta_0}}{||\Theta_{\Theta_0}||_F}\), where \(|| \cdot ||_F\) is the Frobenius norm. The parametrization ensures that the VAR has a stationary solution.

2. Generate a \((d \times (d - 1))\)-matrix \(R\) with i.i.d. standard normal entries. Set \(\Omega_0 = RR^t\) and then \(\Omega = \Omega_0/||\Omega_0||\).

3. Generate a \(d\)-vector \(r\) with i.i.d. standard normal entries and set \(\hat{R} = (R,r)\). Set \(\hat{\Omega}_0 = \hat{R}\hat{R}^t\) and then \(\hat{\Omega} = \hat{\Omega}_0/||\hat{\Omega}_0||\).

4. Compute \(\Gamma_p\) from the parameters \([\Theta_1, \ldots, \Theta_p, R]\) and \(\hat{\Gamma}_p\) from the parameters \([\Theta_1, \ldots, \Theta_p, \hat{R}]\).

5. Compute \(\rho := \log(C(\Gamma_p)/C(\hat{\Gamma}_p))\).

The values obtained for the \(\rho\)’s are given as histograms in Figure I. Depending on the choice of \(p\) and \(q\) the mode is somewhere between 3 and 5. For example, if \(p = 5\) and \(d = 3\) then the median of the ratios \(C(\Gamma_p)/C(\hat{\Gamma}_p)\) in the 1000 randomly chosen setups is beyond 200. (The average is close to 2000!)
perturbation of the noise covariance $\Omega$, rendering this matrix degenerate, leads to a huge increase of the condition number of $\Gamma_p$.

More theoretical results about singular VAR systems can be found in Anderson et al. (2012, 2016) and references therein.

### 2.2. Estimation and Prediction

In the sequel we fix $p$ and suppress it in the notation. For example, we write $\Gamma$ and $Y_t$ instead of $\Gamma_p$ and $Y_{t,p}$. We denote by $A \otimes B$ the Kronecker product between two matrices and let $\mathcal{N}_m(\mu, \Sigma)$ be an $m$-dimensional normal random vector with mean $\mu$ and covariance matrix $\Sigma$. We define $Y^*_t$ as the optimal but infeasible prediction, that is, $Y^*_t = (\Theta_1,\ldots,\Theta_p)Y_t$. We assume that $\Gamma$ and its estimator are full rank. However, we focus on a situation in which $C(\Gamma)$ is large.

The derivations of the previous section indicate that $|\Omega| = 0$ will have an impact on the estimation of the VAR parameters. The matrix $\Gamma$ and thus also $\hat{\Gamma}$ might be close to singular and the resulting estimator $(\hat{\Theta}_1,\ldots,\hat{\Theta}_p)$ is likely to be of poor quality. This does not necessarily imply, however, that the prediction is poor, too. Define

$$P_t(\varepsilon) := (\hat{\Theta}_1,\ldots,\hat{\Theta}_p)(Y_t + \varepsilon) - Y^*_t. \quad (2.5)$$

Then $P_t(0) = \hat{Y}_{t+1} - Y^*_t$ is the difference between our actual prediction and the optimal but infeasible prediction. The quantity $P_t(\varepsilon)$ takes into account that our predictor variable is contaminated with some noise $\varepsilon$ and we are interested to see how this noisy variable impacts the forecast.

As we will see in Proposition 2.2, $|\Omega| = 0$ is not relevant for the order of magnitude of $\|P_t(0)\|$. However, it can be very relevant for $\|P_t(\varepsilon)\|$, even if $\|\varepsilon\|$ is comparably small.

Let us first rewrite our VAR model in state-space form. To this end define

$$\Theta = \begin{pmatrix} \Theta_1 & \Theta_2 & \cdots & \Theta_{p-1} & \Theta_p \\ I & 0 & \cdots & O & O \\ O & I & \cdots & O & O \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ O & O & \cdots & I & O \end{pmatrix} \quad \text{and} \quad Z_t = \begin{pmatrix} Z_t \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$
and thus
\[ Y_{t+1}' = \Theta Y_t' + Z_{t+1}Y_t'. \]

Taking expectations on both sides of (2.6) we get
\[ \Delta = \Theta \Gamma, \quad \text{where} \quad \Delta = E(Y_{t+1}'Y_t'), \]
which is the state-space from of the Yule-Walker equations. On the other hand, summing both sides in (2.6) over \( t \in \{1, \ldots, T-1\} \) and scaling by \( T \) we obtain
\[ \hat{\Delta} = \Theta \hat{\Gamma} + R, \]
where \( \hat{\Delta} := \frac{1}{T} \sum_{t=1}^{T-1} Y_t'Y_t' \), \( \hat{\Gamma} = \frac{1}{T} \sum_{t=1}^{T-1} Y_t'Y_t' \) and \( R := \frac{1}{T} \sum_{t=1}^{T-1} Z_{t+1}Y_t' \). This gives rise to the least squares estimator of \( \Theta \), namely
\[ \hat{\Theta} := \hat{\Delta} \hat{\Gamma}^{-1}, \]
and accordingly \((\hat{\Theta}_1, \ldots, \hat{\Theta}_p) := \hat{\gamma} \hat{\Gamma}^{-1} \), where \( \hat{\gamma} := \Pi_d \hat{\Delta}, \) with \( \Pi_d = (I_d, 0, \ldots, 0) \in \mathbb{R}^{dp \times dp} \). Combining (2.8) and (2.9) leads to \( \hat{\Theta} - \Theta = R \hat{\Gamma}^{-1} \). We remark that the corresponding Yule-Walker estimator from the state-representation is very similar. Here the summation index in \( \hat{\Gamma} \) is running form 1 to \( T \).

It follows that
\[ P_t(0) = \hat{Y}_{t+1} - \Pi_d \Theta Y_t = \Pi_d (\hat{\Theta} - \Theta) Y_t = \Pi_d R \hat{\Gamma}^{-1} Y_t. \]

If our predictor \( Y_t \) is perturbed with noise \( \epsilon \) we have
\[ P_t(\epsilon) = \Pi_d \hat{\Theta} \hat{Y}_t - \Pi_d \Theta Y_t = P_t(0) + \Pi_d \hat{\Theta} \epsilon, \]
with \( \hat{Y}_t = Y_t + \epsilon \). The term \( H(\epsilon) := \Pi_d \hat{\Theta} \epsilon \) represents the additional prediction error due to a noisy predictor. We further split this term into the parts \( H_1(\epsilon) := \Pi_d \Theta \epsilon \) and \( H_2(\epsilon) := \Pi_d (\hat{\Theta} - \Theta) \epsilon \), so that
\[ P_t(\epsilon) = P_t(0) + H_1(\epsilon) + H_2(\epsilon). \]

In Proposition 2.2 we investigate the order of magnitude of the terms \( P_t(0) \) and \( H_i(\epsilon), i = 1, 2 \).

**Proposition 2.2.** Assume that \( \epsilon = (\epsilon_1', \ldots, \epsilon_p')' \) and \( \epsilon_j = (\epsilon_{j1}, \ldots, \epsilon_{jd})' \) with \( \epsilon_{jk} \) being i.i.d. with zero mean and variance \( \sigma^2 \). Moreover, we assume that \( \epsilon \) is independent of the innovations \( (Z_t) \) and hence of the VAR process. We assume that \( (Z_t) \) are i.i.d. and that \( |\Gamma| > 0 \). It holds that
\[ \sigma^2_0 := \frac{1}{T} \sum_{t=1}^{T-1} E[|P_t(0)|^2] = \frac{1}{T} \sum_{t=1}^{m} E[|N_{t,j}|^2]; \]
\[ \sigma^2_1(\epsilon) := E[|H_1(\epsilon)|^2] = \sigma^2 \sum_{j=1}^{p} \text{tr}(\Theta_{t} \Theta_{t}'); \]
\[ \sigma^2_2(\epsilon) := E[|H_2(\epsilon)|^2] = \frac{\sigma^2}{T} \sum_{t=1}^{m} E\left(\frac{|N_{t,j}|^2}{\lambda_{t}}\right). \]
Here the variables \((N_{1T}, \ldots, N_{mT})^T \xrightarrow{D} N_{md}(0, I_d \otimes \Omega)\) (convergence in distribution) and \(\hat{\lambda}_e\) are the eigenvalues in the spectral decomposition \(\hat{\Gamma} = \sum_{e=1}^{m} \hat{\lambda}_e \hat{e} \hat{e}'\).

We would like to stress that the error terms in this proposition are exact and so we can also use these results to assess the finite sample behaviour. We can interpret the result as follows:

1. The term \(\sigma_0^2(e)\) is the usual mean square prediction error when we have no noise contaminating the predictor variables. Due to the limiting distribution of the \(N_{eT}\) we get that \(\|N_{eT}\|^2\) are asymptotically i.i.d. and follow a generalized \(\chi^2\)-distribution \(Z^T \Omega Z\) where \(Z \sim N_{d}(0, I_d)\). If \(\omega_1 \geq \omega_2 \geq \cdots \geq \omega_k\) are the eigenvalues of \(\Omega\), then notice that \(Z^T \Omega Z \sim \sum_{i=1}^{d} \omega_i Z_i^2\). This means that small eigenvalues are not relevant for the distribution \(\|N_{eT}\|^2\).

The distribution essentially depends only on the large eigenvalues of \(\Omega\).

2. The term \(\sigma_1^2(e)\) is a bias term which comes from the noisy predictor variable. It is proportional to \(\sigma_e^2\) and it is not influenced by \(\Omega\).

3. The term \(\sigma_2^2(e)\) measures the error when predicting outside of the stationary distribution. It is proportional to \(\sigma_e^2/T\), a factor which obviously is small when the sample size is large, or when \(\sigma_e^2\) is small. However, it crucially depends on the condition number \(C(\hat{\Gamma})\) via the reciprocal eigenvalues of \(\hat{\Gamma}\). If \(|\Omega| = 0\) these eigenvalues can be very small as we have seen and may dramatically inflate \(\sigma_2^2(e)\) in finite sample. This will become impressively visible in our empirical part.

### 2.3. A Regularization Approach

In view of the previous results we suggest to use a regularized estimator \(\hat{\Theta}^{(K)}\), which is defined as \(\hat{\Theta}^{(K)} := \hat{\Theta} P_K\), where \(P_K := \sum_{l=1}^{K} \hat{e}_l \otimes \hat{e}_l\), with some appropriate choice of \(K \in \{1, \ldots, m\}\). Simply speaking, eigenvectors of \(\hat{\Gamma}\) belonging to small eigenvalues are removed from our estimator. If we denote \(P_K^* = I_m - P_K\), then we obtain by the results of the previous section that \(\hat{\Theta}^{(K)} - \hat{\Theta} P_K = \hat{\Gamma}^{(K)} - P_K\) and thus that \(\hat{\Theta}^{(K)} - \hat{\Theta} = \hat{\Gamma}^{(K)} P_K - \hat{\Theta} P_K^*\). The prediction error \(P_{1[K]}(e) := \Pi_d \left[ \hat{\Theta}^{(K)} Y_t - \hat{\Theta} Y_t \right]\) decomposes into

\[
P_{1[K]}(e) = \Pi_d \left[ \hat{\Theta}^{(K)} - \hat{\Theta} \right] Y_t + \hat{\Theta}^{(K)} e \\
= \Pi_d \left[ \hat{\Gamma}^{(K)} P_K Y_t \right] + \Pi_d \left[ \hat{\Theta} P_K e \right] + \Pi_d \left[ \hat{\Gamma}^{(K)} - P_K \right] Y_t \\
= : P_{1[K]}(0) + H_{1[K]}(e) + H_{2[K]}(e) + H_{3[K]}.
\]

The term \(H_{3[K]}\) is the additional bias that is introduced by the regularization. The meaning of the other error terms is the same as in the previous section. This bias term is solely related to the predictor without noise.

**Proposition 2.3.** Assume the same setting and same notation as in Proposition 2.2. It holds that

\[
\sigma_{0[K]}^2 := \frac{1}{T} \sum_{t=1}^{T-1} E \|H_{1[K]}\|^2 = \frac{1}{T} \sum_{t=1}^{K} E \|N_{eT}\|^2; \\
\sigma_{1[K]}^2(e) := E \|H_{1[K]}(e)\|^2 = \sigma_e^2 E \text{tr} \left( \hat{\Theta} P_K \hat{\Theta}' \right); \\
\sigma_{2[K]}^2(e) := E \|H_{2[K]}(e)\|^2 = \frac{\sigma_e^2}{T} \sum_{t=1}^{K} E \left( \frac{\|N_{eT}\|^2}{\hat{\lambda}_e} \right); \\
\sigma_{3[K]}^2 := \frac{1}{T} \sum_{t=1}^{T-1} E \|H_{3[K]}\|^2 = E \text{tr} \left( \hat{\Theta} \left( \sum_{k=1}^{m} \hat{\lambda}_e \hat{e}_k \hat{e}_k' \right) \hat{\Theta}' \right).
\]
The variables $N_{k,t}$ have the same properties as stated in Proposition 2.2.

We see that the terms $\sigma^2_{0|K}$ and $\sigma^2_{1|K}(e)$ are of a similar order of magnitude as $\sigma^2_0$ and $\sigma^2_1(e)$ respectively, if $K$ is close to $m$. On the contrary, a large $K$ implies that small reciprocal eigenvalues are included in the expansion of $\sigma^2_{2|K}(e)$, to which this term is very sensitive. The terms $\sigma^2_{0|K}$, $\sigma^2_{1|K}(e)$ and $\sigma^2_{2|K}(e)$ are growing with $K$, while $\sigma^2_{3|K}$ decays. The tuning of $K$ will be a bias-variance trade-off problem. In this article, we consider two methods for choosing $K$:

The TRIM approach: Select $K$ which minimizes $\Sigma(K) : = \sigma^2_0 + \sigma^2_{1|K}(e) + \sigma^2_{2|K}(e) + \sigma^2_{3|K}$. In practice $\Sigma(K)$ is unknown and requires an empirical counterpart. So we compute estimators of the $\Sigma(K)$'s. Details of how we tackle this in our implementation are given in Appendix B.

The CV approach: Select $K$ by cross-validation. Again the implementation details are given in Appendix B.

The numerical performance of both approaches and a comparison with the usual OLS estimator are provided in Section 4. From a theoretical point of view, it would be interesting to give bounds on $\Sigma(m) - \Sigma(K)$. (Note that $\Sigma(m) = \sigma^2_0 + \sigma^2_1(e) + \sigma^2_2(e)$.) Since we are interested in the non-asymptotic case, this appears a very difficult challenge which is beyond the scope of this article.

We conclude this section by noting that other regularization methods could be used as well. For example, in our numerical exercises we have also worked with a ridge penalty but found our approach to be favourable. We have not explored other regularization methods, and leave this open for further research.

3. PREDICTION OF GDFMS

The literature on dynamic factor models has been expanding at a very rapid pace during the previous decades, mostly due to the success of such models to predict macro-economic variables within a large cross-sectional panel dataset and relatively small time range (cf. Stock and Watson (2011) for a review). Until today dynamic factor models and so-called GDFMs receive a lot of attention due to their wide range of applicability and their empirical success (see e.g. Forni et al. (2018)).

In a GDFM observation vectors $X_t$ (which usually are very high dimensional) can be decomposed into an unobserved common component $x_t$ – capturing most of the serial and cross-sectional dependence – and an unobserved idiosyncratic component $\xi_t$ – being weakly correlated in time and cross-section. We then have

$$X_t = x_t + \xi_t, \quad t \in \mathbb{Z}. \tag{3.1}$$

If a generalized dynamic $q$-factor model holds, then $x_t$ is driven by $q$ serially orthogonal white noise sequences. We refer to Barigozzi (2019) for a comprehensive exposition of dynamic factor models. We recall key features of the GDFM (3.1) and the way the FHLZ prediction algorithm works. We do not present here all technicalities, but refer to corresponding articles Forni et al. (2015, 2017). In Section 3.1 we also explain the FHLZ estimation framework, while in Section 3.2 we outline the prediction step. We show how the findings in Section 2 relate to forecasting GDFMs and how they can be used to improve the corresponding prediction method.

3.1. GDFM Setup and FHLZ Estimation Procedure

Consider infinite dimensional vectors $X^\omega_t = (X_{1t}, X_{2t}, \ldots)'$, $t \in \mathbb{Z}$, and let us assume that our observations $X_t = (X_{1t}, X_{2t}, \ldots, X_{Nt})'$ are the first $N$ components of it. For a lighter notation we suppress the cross-sectional dimension in our variables and parameters. For any $N$, our observations are assumed to form a second-order stationary process. Moreover, all processes appearing in the sequel are assumed to be zero mean. We let $\gamma_X(h) = EX^\omega_thX^\omega_t$ be the auto-covariance of $(X_t)$ at lag $h$ and let $\hat{\gamma}_X(h)$ be its sample version. We denote by $F_X(\theta) : = \sum_{h \geq 0} \hat{\gamma}_X(h)e^{-i\theta h}$ the corresponding spectral density matrix at frequency $\theta \in [-\pi, \pi]$. The series defining $F_X(\theta)$ is supposed to converge for almost all $\theta$. The eigenvalues of $F_X(\theta)$ are denoted $\lambda_X(j\theta)$, $j \geq 1$, and are indexed in decreasing order. The analogous notation is used for the common component $\chi$ and the idiosyncratic component $\xi$. With
the superscript $\infty$ we indicate corresponding quantities based on $X_t^{\infty}$. For example, we use $\nu_\xi^{\infty}(h)$ or $F_{\mathcal{X}}^{\infty}(\theta)$. The eigenvalues $\lambda^{\infty}_k(\theta|\theta)$ are interpreted as the limit of $\lambda_k(\theta|\theta)$ when $N \to \infty$.

In a GDFM it is assumed that the common components are given as

$$
\chi_t = b_{1i}(L)u_{1t} + \ldots + b_{qi}(L)u_{qt}, \quad (i \geq 1),
$$

(3.2)

where $(u_{ij} : 1 \leq j \leq q, t \in \mathbb{Z})$ is the $q$-dimensional time series of factors. This vector time series is assumed to be an orthonormal white noise independent of the idiosyncratic component $\xi_t$ defined in (3.1) at any lead and lag. The $b_{ij}(L) = \sum_{\ell \geq 0} b_{ij\ell} L^\ell$ denote the backshift-series of factor loadings. The idiosyncratic components are assumed to be in some sense weakly dependent in cross-section. Roughly speaking, it is assumed that the cross-sectional average of $\xi_t$ converges to zero. Hence, all relevant cross-sectional dependences are contained in the common components. Thus, every unit of idiosyncratic component is forecasted independently of the other unit via a univariate model, typically an autoregressive model. The final forecast is the sum of the common component forecast and the idiosyncratic forecast (Section 8 of Barigozzi (2019)). In the theory part of this article and in the simulation study we focus on the forecast of the common component.

A key result in Forni and Lippi (2001) is that under stationarity conditions and if $\lambda^{\infty}_k(q|\theta) = \infty$ for almost all $\theta \in [-\pi, \pi]$, while $\lambda^{\infty}_k(q + 1|\theta) \leq B < \infty$ for almost all $\theta \in [-\pi, \pi]$, then for any $N \geq 1$ the process $(X_t)$ can be represented as in (3.1) and (3.2) where the filter coefficients are square summable, that is $\sum_{\ell \geq 0} b_{ij\ell}^2 < \infty$ for $i \geq 1, \quad 1 \leq j \leq q$. The idiosyncratic component and the common component are also stationary and their spectral densities exist. All eigenvalues of $F_\mathcal{X}(\theta)$ remain bounded, while the first $q$ eigenvalues of $F_{\mathcal{X}}(\theta)$ diverge with $N \to \infty$. This decomposition is unique.

We now turn to the explanation of the FHLZ prediction method, a state-of-the-art prediction method for GDFMs. It requires that the filters in the GDFM are rational:

$$
b_{ij}(L) = \frac{c_{ij}(L)}{d_{ij}(L)}, \quad i \geq 1, \quad 1 \leq j \leq q.
$$

(3.3)

where $c_{ij}(L)$ and $d_{ij}(L)$ are lag polynomials of degree $\leq s_1$ and $\leq s_2$ respectively. For simplicity of the presentation, we assume that $s_1$ and $s_2$ are uniform bounds with respect to the index $i$. This latter assumption is milder than the classical assumption that $b_{ij}(L)$ is a finite lag polynomial. The FHLZ framework encompasses important models that the anterior literature could not consider. For instance, it contains the important special case in which every cross-sectional unit $X_{it}$, $t = 1, \ldots, T$ is an autoregressive process.

Let $i = (i_1, \ldots, i_p)$, with $1 \leq i_1 < i_2 < \cdots < i_p \leq N$, and for some vector $x \in \mathbb{R}^N$ denote by $x_i = (x_{i_1}, \ldots, x_{i_p})$. It was then shown in Forni et al. (2015) that for almost all such parametrizations $c_{ij}$ and $d_{ij}$ of the GDFM, any subvector $\chi_t^{i}$ has the following VAR representation:

$$
\chi_t = A_{i1} \chi_{t-1} + \ldots + A_{ip} \chi_{t-p} + R_t u_t, \quad t = p_0 + 1, \ldots, T,
$$

(3.4)

where $A_{ij}, j = 1, \ldots, p_0$ are $(q + 1) \times (q + 1)$ real matrices, $R_t$ is an $(q + 1) \times q$ matrix and $p_0 = p_0(\theta) \leq p_{\text{max}} := s_1 q + s_2 q^2$. Set $A_i := (A_{i1}, A_{i2}, \ldots, A_{ip})$. For the sake of simplicity, we will assume an arbitrary but fixed $N$ of the form $N = (q + 1)n$.

If $j \to i_j$ denotes a permutation of $\{1, \ldots, N\}$, the above result implies that one can obtain $n$ VAR equations (3.4) by selecting the subsets of indexes $\{i_1, i_2, \ldots, i_q\}, \{i_{q+2}, i_{q+3}, \ldots, i_{2q+1}\}, \ldots, \{i_{(n-1)q+1}, \ldots, i_{nq+1}\}$. Reassembling these $n$ equations, a VAR representation of the whole vector $\chi_t$ can be obtained:

$$
\chi_t = A_1 \chi_{t-1} + \ldots + A_p \chi_{t-p} + R_t u_t, \quad t = p + 1, \ldots, T,
$$

where $p$ denotes the maximum among the true orders of the $n$ VAR equations. Some remarks are in order.
Remark 3.1. The orders $\rho_{i0}$ of these VAR model are unknown and need to be estimated.

Remark 3.2. Let $\Omega_i = \text{Var}(R_iu_i)$. Then apparently $\Omega_i$ is of rank $q$, which means that $(\chi_i^e)$ is a singular VAR process. That this feature is not specifically restricted to above model assumptions is stressed in Anderson et al. (2012): ‘In almost all empirical econometric modelling and forecasting exercises using GDMF’s, the latent variables are modelled as a singular AR model,…’.

We are now ready to outline the estimation part of the FHLZ algorithm. For details we refer to Forni et al. (2017).

**THE FHLZ ESTIMATION ALGORITHM:**

1. Estimate the spectral density matrices of the observations $(X_t)$, for example with the Bartlett lag-window estimator:

$$\hat{P}_X(\theta) := \sum_{-B \leq h \leq B} \left(1 - \frac{|h|}{B}\right) \hat{\gamma}_X(h)e^{-i\theta}.$$  \hfill (3.5)

2. Derive estimators $\hat{P}_X(\theta)$ and $\hat{P}_U(\theta)$. Assumptions on the model imply that a spectral expansion of $\hat{P}_X(\theta)$ along the first $q$ eigenvectors consistently estimates the population spectral density $F(\theta)$ and an expansion on the subsequent eigenvectors consistently estimates $F(\theta)$.

3. Estimate the auto-covariance matrices at lag $h$ of the common component through

$$\hat{\gamma}_X(h) = \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{P}_X(\theta) e^{ih} d\theta.$$  

In practice the integral is computed by averaging over a set of frequencies.

4. Select corresponding components of this matrix to obtain the autocovariance matrix $\hat{\gamma}_X(h)$ of $\chi_t$.

5. For the $n$ equations (3.4), select the respective orders $\hat{\rho}_{i0}$ and use the Yule-Walker equations to estimate the coefficients of these VARs. For each of the $n(q+1)$-dimensional VARs we thus obtain the estimators $\hat{A}_{ij}$, $j = 1, \ldots, \hat{\rho}_{i0}$. Let $\hat{\rho}$ be the maximum of the $n$ estimated orders $\hat{\rho}_{i0}$.

6. Combine the elements of the previous step to obtain estimators $\hat{A}_1, \ldots, \hat{A}_{\hat{\rho}}$. Define $\hat{E}_t := X_t - (\hat{A}_1X_{t-1} + \ldots + \hat{A}_{\hat{\rho}}X_{t-\hat{\rho}})$, $t = \hat{\rho} + 1, \ldots, T$.

7. The residuals $\hat{E}_t$ are estimators of $E_t = Ru_t + A(L)\xi_t$, where $A(L) = I - A_1L - \ldots - A_{\hat{\rho}}L^\hat{\rho}$. Now $R$ and $u_t$ are estimated through a regular PCA on the residual process $(\hat{E}_t)$.

Remark 3.3. Note that since we do not directly observe the common components, the estimator of the VAR parameters in 5 is not exactly obtained as in Section 2.2. We obtain the second-order moments via the spectral analysis and use those moments for solving the Yule-Walker equations.

### 3.2. PHLZ Predictions and How to Improve Them

With the steps 1–7 combined, we obtain estimators $\hat{A}$ and $\hat{R}$ and $\hat{\mu}_t$. The next step is the actual prediction step. As we have mentioned previously, we focus on the forecast of the common component. Let us describe this in detail and how it relates to our Section 2.

We have already pointed out in Remark 3.2 that the $(q+1)$-dimensional subvectors $\chi_t$ define singular VAR processes, which we write in a more compact form as $A_1(L)\chi_t = R_iu_i$, where $A_1(L) = I - A_1L - \ldots - A_{\hat{\rho}}L^\hat{\rho}$. Hence $\chi_t = (A_1(L))^{-1}R_iu_i$. Would we observe $\chi_{t-h}$, $h = 0, 1, 2, \ldots$, our forecast would be naturally defined as

$$\hat{\chi}_{t+h} = (I - \hat{A}_1(L))\hat{\chi}_{t+1} = \hat{A}_1\hat{\chi}_{t+1} + \ldots + \hat{\hat{A}}_{\hat{\rho}}\hat{\chi}_{t-\hat{\rho}+1}.$$  

Since we do not observe $\chi_t$ we use $\hat{\chi}_{t+h} = (\hat{A}_1(L))^{-1}\hat{R}_i\hat{\mu}_t$ and define instead

$$\hat{\chi}_{t+1} = (I - \hat{A}_1(L))\hat{\chi}_{t+1} = \hat{A}_1\hat{\chi}_{t+1} + \ldots + \hat{\hat{A}}_{\hat{\rho}}\hat{\chi}_{t-\hat{\rho}+1}.$$  \hfill (3.6)
Computing predictions of each subsystem in this way leads to the forecast \( \hat{Y}_{t+1} \) of the entire vector. In the FHLZ algorithm steps 1–7 are repeated \( N_{\text{rep}} \) times, typically 30 times, for different partitions of the cross-section in \((q + 1)\)-dimensional subvectors. The resulting predictions for each component are then averaged.

A look at (3.6) shows that we find ourselves in the context of predicting singular VARs with noisy predictors. A look at (3.6) shows that we find ourselves in the context of predicting singular VARs with noisy predictors.

The key point of our proposed modification is to alter the estimation step 5 (Yule-Walker estimation of the sub-VAR systems) and use either TRIM or CV. The other steps remain unchanged. The details for these implementation methods are given in Appendix B.

4. SIMULATIONS

We will run two sets of simulations. In the first one, we study the impact of spectral trimming on the prediction of VAR processes of different dimensions, order, persistence, sample size and rank. In the second one, we study the impact of the trimming on the prediction of a dynamic factor model and we compare it to the VAR processes of different dimensions, order, persistence, sample size and rank. In the second one, we study the

4.1. Forecasting VARs

We generate a VAR(p) process \((Y_t : 1 \leq t \leq T) (p = 1, 2, 3)\) of dimension \(d = 3, 4, 5\) with sample size \(T = 120, 240\), where the first eigenvalue of the companion matrix is \(\rho = 0.1, 0.4, 0.8\) and the innovation process has rank \(d - L\), \(0 \leq L \leq 2\). Our procedure is as follows: first, we draw \(pd^2\) coefficients, independently from a uniform distribution on \([-1, 1]\). Let \(\Theta_{0i}, i = 1, \ldots, p,\) be the resulting coefficients matrices of dimension \(d \times d\). If \(\lambda_{\max}\) is the largest eigenvalue of the corresponding companion matrix, we define the actual coefficient matrices by \(\Theta_i = (\rho / \lambda_{\max}) \Theta_{0i}\) and set \(\Theta_p = (\Theta_1, \ldots, \Theta_p)\). This parametrization ensures that the large eigenvalue of the companion matrix of \(\Theta_p\) is \(\rho\). Hence, with \(|\rho| < 1\) the process will be stable. The innovations process \((Z_t : 1 \leq t \leq T)\) is the product of a \(d \times (d - L)\) matrix whose coefficients are independently drawn from \(U[-1, 1]\) and a \((d - L) \times T\) matrix whose coefficients are independently drawn from a standard normal distribution. The matrix of coefficients \(\Theta_p\) and the innovations \((Z_t)\) permit to generate the VAR process \((Y_t)\). From this process we compute the LS estimator \(\hat{\Theta}_p = (\hat{\Theta}_1, \ldots, \hat{\Theta}_p)\). We then generate perturbed observations \(\tilde{Y}_t\) by adding noise \(\epsilon_t\) to \(Y_t\). The elements of \(\epsilon_t\) are drawn independently from a standard normal distribution with standard deviation being 0.01\(\hat{\epsilon}\), where \(\hat{\epsilon}\) is the mean standard deviation across the components of the time series \(Y_t\). Hence, the scale of the noise is 1% of the scale of the observations.

We separate the sample into training set and test set. The training set contains the first [0.87] observations. For \(t \in T_{\text{test}} = \{[0.87] + 1, [0.87] + 2, \ldots, T\}\) and for iteration \(i\), we compute the following quantities:

1. \(\delta_{t1} = ||\hat{\Theta}_p Y_t - \Theta_p Y_t||^2\),
2. \(\delta_{t2} = ||\hat{\Theta}_p \tilde{Y}_t - \Theta_p Y_t||^2\),
3. \(\delta_{t3} = ||\hat{\Theta}_p P_K \tilde{Y}_t - \Theta_p Y_t||^2\), \(K \) selected by TRIM,
4. \(\delta_{t4} = ||\hat{\Theta}_p P_K \tilde{Y}_t - \Theta_p Y_t||^2\), \(K \) selected by CV.

The norm considered is the Euclidian norm of \(R^d\). Hence \(\delta_{t1}\) represents the classical value for a prediction error (compared to the optimal unfeasible prediction \(\Theta_p Y_t\)) of a VAR. It corresponds to \(||P_0(0)||^2\). The second value \(\delta_{t2}\) represents the case where we have noisy predictors, but where we ignore this in our forecasting method.
corresponds to \( \|P_t(e)\|^2 \). The terms \( \delta_{t3} \) and \( \delta_{t4} \) also correspond to the prediction errors with noisy predictors, but when we use our proposed regularization methods, namely \( \|P_t(e)\|^2 \), with \( K \) chosen by TRIM or CV.

Based on 1000 replications, we compute the ratios:

\[
RE(k) := \frac{\sum_{i=1}^{1000} \sum_{t \in \text{test}} \delta_{itk}}{\sum_{i=1}^{1000} \sum_{t \in \text{test}} \delta_{it1}}, \quad k = 2, 3, 4.
\]

For simplicity we write \( RE = RE(2) \), \( RE_{\text{TRIM}} = RE(3) \) and \( RE_{\text{CV}} = RE(4) \). Hence, \( RE \) represents the inflation of the prediction error due to the perturbation \( e \), while \( RE_{\text{TRIM}} \) and \( RE_{\text{CV}} \) additionally take into account the proposed spectral trimming.

The values of \( RE(k) \) for \( \rho = 0.4 \) and \( T = 240 \) are displayed in Table I. Other constellations of \( \rho \) and \( T \) are given in Tables D1–D5 in Appendix D. We notice that in average regularized predictions are more accurate than the standard method for all cases and that we get tremendously more accurate results in some situations.

When \( L = 0 \) (full rank noise) the effect of the noisy predictor variables is very small (\( RE \approx 1 \)). Still, the forecasts based on the regularized estimators are significantly better, especially with larger values of \( p \).

When \( L \neq 0 \) (singular VAR) a regularization is absolutely necessary. The prediction error can get completely out of control. The bigger \( L \), the more relevant is the regularization. A lower rank of the noise covariance implies a larger condition number of \( \Gamma_p \) which causes often extreme instability of the predictions. For the same reason (see Proposition 2.1), the naive prediction becomes worse the smaller \( \rho \) and the larger \( p \). While the OLS estimator is the best linear unbiased estimator, it is far not optimal for predictions.

We also observe that CV is often leading to considerably better results than TRIM. An advantage in favour of TRIM is the low computational costs compared to CV. Moreover, the method TRIM will be useful when we work with GDFMs. Here, CV cannot be directly used as we explain in Appendix B and requires TRIM as a preliminary step.

### 4.2. Comparison of FHLZ, TRIM and CV in the Context of GDFMs

We now use spectral trimming in the prediction algorithm of a dynamic factor model. Again we call the approaches TRIM or CV as compared to FHLZ. We simulate from the data generating process (DGP) described by (3.1) and (3.3). More precisely, we choose \( c_{ij}(L) = a_{ij} \) and \( d_{ij}(L) = 1 + a_{ij}L \) where \( a_{ij} \sim U[-0.8, 0.8], i = 1, \ldots, N, \) and \( j = 1, \ldots, q \). This DGP is the same as Forni et al. (2017) chose for their simulation.

| \( d \) | \( p = 1 \) | \( p = 2 \) | \( p = 3 \) |
|---|---|---|---|
| \( L = 0 \) | \( RE = 1 \) | \( RE = 1 \) | \( RE = 1 \) |
| \( RE_{\text{TRIM}} \) | 0.98 | 0.78 | 0.28 |
| \( RE_{\text{CV}} \) | 0.91 | 0.62 | 0.39 |

| \( L = 1 \) | \( RE = 1 \) | \( RE = 1 \) | \( RE = 1 \) |
|---|---|---|---|
| \( RE_{\text{TRIM}} \) | 0.96 | 0.78 | 0.28 |
| \( RE_{\text{CV}} \) | 0.9 | 0.6 | 0.3 |

| \( L = 2 \) | \( RE = 1 \) | \( RE = 1 \) | \( RE = 1 \) |
|---|---|---|---|
| \( RE_{\text{TRIM}} \) | 0.94 | 0.76 | 0.26 |
| \( RE_{\text{CV}} \) | 0.85 | 0.6 | 0.3 |

| \( L = 3 \) | \( RE = 1 \) | \( RE = 1 \) | \( RE = 1 \) |
|---|---|---|---|
| \( RE_{\text{TRIM}} \) | 0.92 | 0.75 | 0.25 |
| \( RE_{\text{CV}} \) | 0.85 | 0.6 | 0.3 |

| \( L = 4 \) | \( RE = 1 \) | \( RE = 1 \) | \( RE = 1 \) |
|---|---|---|---|
| \( RE_{\text{TRIM}} \) | 0.9 | 0.7 | 0.2 |
| \( RE_{\text{CV}} \) | 0.85 | 0.6 | 0.3 |

| \( L = 5 \) | \( RE = 1 \) | \( RE = 1 \) | \( RE = 1 \) |
|---|---|---|---|
| \( RE_{\text{TRIM}} \) | 0.88 | 0.65 | 0.2 |
| \( RE_{\text{CV}} \) | 0.85 | 0.6 | 0.3 |

### Table I. Relative prediction errors for \( \rho = 0.4, T = 240 \). 1000 replications
TABLE II. Median of $Q_{TRIM}$ (and standard deviation of it) across repetitions

| $N/T$ | 60   | 120  | 240  | 480  |
|-------|------|------|------|------|
| $q=1$ |      |      |      |      |
| 30    | 0.06 (0.13) | 0.26 (0.28) | 0.26 (0.35) | 0.23 (0.34) |
| 60    | 0.07 (0.11) | 0.26 (0.19) | 0.24 (0.31) | 0.21 (0.24) |
| 120   | 0.09 (0.07) | 0.27 (0.17) | 0.23 (0.23) | 0.21 (0.28) |
| 240   | 0.09 (0.06) | 0.28 (0.15) | 0.24 (0.22) | 0.28 (0.27) |
|        |      |      |      |      |
| $q=2$ |      |      |      |      |
| 30    | 0.55 (0.24) | 0.48 (0.28) | 0.46 (0.3)  | 0.44 (0.31) |
| 60    | 0.5 (0.2)  | 0.43 (0.21) | 0.48 (0.23) | 0.49 (0.22) |
| 120   | 0.48 (0.14) | 0.46 (0.17) | 0.61 (0.17) | 0.63 (0.18) |
| 240   | 0.51 (0.14) | 0.59 (0.16) | 0.74 (0.12) | 0.75 (0.12) |

We take $u_t \sim \mathcal{N}(0, 1), j = 1, \ldots, q, t = 1, \ldots, T$. For the idiosyncratic components, we take $\xi_t \sim \mathcal{N}(0, 1)$. Every experiment $(T, N, q)$ consists of 500 repetitions of the same procedure with different drawings of the random values (including the parameters). Based on the DGP described above, we make one-step ahead predictions for the common components $\hat{\chi}^{(j)}_T$ with $j \in \{ \text{FHLZ, TRIM, CV} \}$. These predictions are solely based on the information provided by $X_t$, $i = 1, \ldots, N$, and $t = 1, \ldots, T-1$. We then compare these procedures through the relative mean square error between the forecast and the theoretically optimal forecast $\chi^{(j)}_T$:

$$Q^{(j)} := ||\hat{\chi}^{(j)}_T - \chi^{(opt)}_T||^2$$

with $j \in \{ \text{FHLZ, TRIM, CV} \}$ at every repetition.

We estimate the spectral density of the different components using a lag window of length $[T^{1/2}]$. We limit the computation of the inverse of $\hat{A}(L)$ to 20 lags. The number of factors $q$ is estimated by the criterion $IC_2$ of Hallin and Liska (2007) with their penalization function $p_1$, with the constant $c$ being selected from the grid $\{0, 0.01, 0.02, \ldots, 2\}$, the different sample sizes being constant at $T$ and the different cross-section dimensions $N_j$ being $\{[27N/36], [28N/36], \ldots, N\}$. Finally, we do $N_{rep} = 30$ permutations.

The comparison between FHLZ and TRIM is shown in Table II. A comparison of TRIM and CV is shown in Table D6 of Appendix D. We see a huge improvement when using our proposed procedures. Depending on the choice of $q, T, N$ the median prediction error over the 500 simulations can be as low as 6% of the one produced by FHLZ. One can also observe that CV gives a further improvement compared to TRIM. To disentangle the effect of the trimming and order selection we have run all methods with the BIC criterion as it is proposed in FHLZ. Using a modified version (see Section B.4) further improves the accuracy of all methods. The relative accuracy between them remains mostly unchanged.

5. CONCLUSION

We have investigated the problem of forecasting singular VARs. A particular and important application is the forecasting of GDFMs via the FHLZ prediction algorithm. The latter is founded on the fact that $(q+1)$-dimensional subvectors of a $q$-factor GDFM have a singular VAR structure.
We have seen that predictions for singular VARs have to be computed with great care, since estimators can be rather unstable. The problem is even more severe if our predictor variables are perturbed and sampled outside the stationary distribution of the underlying process. This particular situation occurs in the context of GDFMs but is also common in other settings.

In our article we explain the mathematical reason for the above phenomenon and demonstrate the specific effect in terms of several simulations experiments. We conjecture that this issue has not been investigated in the literature since it is asymptotically negligible. However, it is extremely relevant in finite sample. The instability which is leading to poor predictions, can be easily cured by some regularization. We have explored spectral trimming, but other regularization methods might also be considered. In numerical exercises we have shown that throughout all considered settings the method we propose gives statistically significant, sometimes even dramatic improvements.

ACKNOWLEDGEMENTS

We are grateful to Marc Hallin and Manfred Deistler for valuable discussions during the preparation of this manuscript. This research was supported by the Fonds de la Recherche Scientifique – FNRS under Grant MCF/FC 24535233.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are simulated data. The code is made available by the authors upon request.

REFERENCES

Allen G, Sioutas C, Koutrakis P, Reiss R, Lurmann F, Roberts P. 2012. Evaluation of the TEOM method for measurement of ambient particulate mass in urban areas. *Journal of the Air & Waste Management Association* **47**: 682–689.

Anderson D, Deistler M, Chen W, Filler A. 2012. Autoregressive models of singular spectral matrices. *Automatica* **48**: 2843–2849.

Anderson D, Deistler M, Felsenstein E, Koelbl L. 2016. The structure of multivariate AR and ARMA systems: regular and singular systems; the single and the mixed frequency case. *Journal of Econometrics* **192**: 366–373.

Barigozzi M. 2019. Dynamic factor models. Lecture Notes.

Chen W, Anderson BDO, Deistler M, Filler A. 2011. Solutions of Yule-Walker equations for singular AR processes. *Journal of Time Series Analysis* **32**: 531–538.

Forni M, Giovannelli A, Lippi M, Soccorsi S. 2018. Dynamic factor model with infinite-dimensional factor space: forecasting. *Journal of Applied Econometrics* **33**: 625–642.

Forni M, Hallin M, Lippi M, Zaffaroni P. 2015. Dynamic factor models with infinite-dimensional factor space: one-sided representations. *Journal of Econometrics* **185**: 359–371.

Forni M, Hallin M, Lippi M, Zaffaroni P. 2017. Dynamic factor models with infinite-dimensional factor space: asymptotic analysis. *Journal of Econometrics* **199**: 74–92.

Forni M, Lippi M. 2001. The generalized dynamic factor model: representation theory. *Econometric Theory* **17**: 1113–1141.

Hallin M, Liska R. 2007. Determining the number of factors in the general dynamic factor model. *Journal of the American Statistical Association* **102**: 603–617.

Stock J, Watson M. 2011. Dynamic factor models. *Oxford Handbook of Economic Forecasting* **1**: 35–59.

APPENDIX A. PROOFS

*Proof of Proposition 2.1.* Define \( \Psi_{p+1} = (I_d, -\Theta_p) \), \( p \geq 1 \). From (2.1) we deduce that

\[
\Psi_p Y_{t-p} = Z_t + \Theta_p Y_{t-p}
\]
and thus that
\[
\text{Var} (\Psi_p Y_t) = \Psi_p \Gamma_p \Psi_p' = \Theta_p \gamma(0) \Theta_p'.
\]

Moreover, recall that for any vector \( w \in \mathbb{R}^{dp} \) we have that
\[
\lambda_{\min}(\Gamma_p) \leq \frac{w' \Gamma_p w}{w'w} \leq \lambda_{\max}(\Gamma_p).
\]

Consequently, for any \( v \in \mathbb{R}^d \) it holds that
\[
\lambda_{\min}(\Gamma_p) \leq \frac{v' \Psi_p \Gamma_p \Psi_p' v}{v' \Psi_p \Psi_p' v} = \frac{v' \Omega v + v' \Theta_p \gamma(0) \Theta_p' v}{v' \Psi_p \Psi_p' v} \leq \lambda_{\max}(\Gamma_p).
\]

Denote by \( g_1 \) and \( g_d \) the eigenvectors of \( \Omega \) belonging to its largest and to its smallest eigenvalue respectively. Then \( g_d' \Omega g_d = 0 \) and we obtain the following bounds:
\[
\lambda_{\min}(\Gamma_p) \leq \frac{g_d' \Theta_p \gamma(0) \Theta_p' g_d}{g_d' \Psi_p \Psi_p' g_d} \leq \|\Theta_p\|^2 \|\gamma(0)\|.
\]

(A.1)

For the second inequality we use that
\[
\Psi_p \Psi_p' = \text{Id} + \sum_{i=1}^{p-1} \Theta_i \Theta_i'
\]

and hence that
\[
g_d' \Psi_p \Psi_p' g_d \geq 1.
\]

Next we have
\[
\lambda_{\max}(\Gamma_p) \geq \frac{\lambda_{\max}(\Omega) + g_1' \Theta_p \gamma(0) \Theta_p' g_1}{g_1' \Psi_p \Psi_p' g_1} \geq \frac{\lambda_{\max}(\Omega)}{1 + \sum_{i=1}^{p-1} \|\Theta_i\|^2}.
\]

This yields (2.3).

For (2.4), we remark that if \( \alpha < 1 \), then it can be readily shown that
\[
\|\gamma(0)\| \leq \|\Omega\| \frac{1}{1 - \alpha}.
\]

\[\square\]

Proof of Proposition 2.2. We have
\[
P_s(0) = \frac{1}{T} \sum_{t=1}^{T-1} Z_{t+1} Y_t' \Gamma^{-1} Y_s = \frac{1}{\sqrt{T}} \sum_{t=1}^{T-1} \left( \frac{1}{\sqrt{T}} \sum_{t=1}^{T-1} Z_{t+1} Y_t' \hat{\epsilon}_\ell \right) \frac{Y_s' \hat{\epsilon}_\ell}{\sqrt{\lambda_\ell}};
\]

\[
H_1(\epsilon) = \sum_{\ell=1}^{p} \Theta_{\ell} \epsilon_{\ell};
\]

\[
H_2(\epsilon) = \frac{1}{T} \sum_{t=1}^{T-1} Z_{t+1} Y_t' \Gamma^{-1} \epsilon = \frac{1}{\sqrt{T}} \sum_{t=1}^{T-1} \left( \frac{1}{\sqrt{T}} \sum_{t=1}^{T-1} Z_{t+1} Y_t' \hat{\epsilon}_\ell \right) e' \hat{\epsilon}_\ell \sqrt{\lambda_\ell}.
\]
Set \( N_{\ell,T} := \frac{1}{\sqrt{T}} \sum_{t=1}^{T-1} Z_{t+1} Y_{t+1}^\ell \). By a routine martingale CLT argument we can show that for any vectors \( v_1, \ldots, v_m \in \mathbb{R}^m \) it holds that
\[
\left( \frac{1}{\sqrt{T}} \sum_{t=1}^{T-1} Z_{t+1}^\ell Y_{t+1}^\ell v_1, \ldots, \frac{1}{\sqrt{T}} \sum_{t=1}^{T-1} Z_{t+1}^\ell Y_{t+1}^\ell v_m \right) \xrightarrow{D} \mathcal{N}_{md} \left( 0, L \otimes \Omega \right),
\]
where \( L = \left( \langle v_i^T \Gamma v_j \rangle : 1 \leq i, j \leq n \right) \). The consistency of \( \hat{\Gamma} \) implies consistency of \( \hat{\lambda}_\ell \) and \( \hat{e}_\ell \), and hence with a Slutsky type argument it can be readily shown that
\[
\left( N_{1,T}, \ldots, N_{m,T} \right) \xrightarrow{D} \mathcal{N}_{dm} \left( 0, I_m \otimes \Omega \right).
\]

Now we have
\[
\frac{1}{T} \sum_{s=1}^{T-1} \| P_s(0) \|^2 = \sum_{\ell=1}^{m} \sum_{j=1}^{m} N_{\ell,T}^\prime N_{\ell,T} \frac{1}{\sqrt{\lambda_\ell}} \frac{Y_{s}^\ell \hat{e}_\ell}{\sqrt{\lambda_\ell}}.
\]
Noting that \( \frac{1}{T} \sum_{s=1}^{T-1} Y_{s}^\ell \hat{e}_\ell = \frac{\rho_\ell \hat{e}_\ell}{\sqrt{\lambda_\ell} \sqrt{\lambda_j}} \) the result for \( \sigma_0^2 \) is proven. The term \( \sigma_1^2(\ell) \) is immediate. Finally, using orthogonality of the \( \hat{e}_\ell \) it is easy to show that
\[
E \left[ \| H_2(\ell) \|^2 \right] = \frac{\sigma_2^2}{T} \sum_{\ell=1}^{m} \| N_{\ell,T} \|^2 \hat{\lambda}_\ell.
\]
Taking expectations finishes the proof.

**Proof of Proposition 2.3.** The result for \( \sigma_{0|K}^2 \) and \( \sigma_{2|K}^2(\ell) \) is obtained exactly in the same way as in Proposition 2.2. To obtain \( \sigma_{1|K}^2(\ell) \) we basically only need to apply the definition. Finally, we have
\[
\frac{1}{T} \sum_{s=1}^{T-1} E \left[ \| H_{3|K} \|^2 \right] = E \left( \Pi_\ell \Theta (I_m - P_K) \hat{\Gamma} (I_m - P_K) \Theta' \Pi_\ell' \right)
\]
\[
= E \left( \Pi_\ell \Theta \left( \sum_{\ell=K+1}^{m} \hat{\lambda}_\ell \hat{e}_\ell e'_\ell \right) \Theta' \Pi_\ell' \right)
\]
\[
= E \left( \Theta_p \left( \sum_{\ell=K+1}^{m} \hat{\lambda}_\ell \hat{e}_\ell e'_\ell \right) \Theta_p' \right).
\]

**APPENDIX B. DETAILS ON THE NUMERICAL IMPLEMENTATION**

**B.1. Choosing \( K \) with TRIM**

We first assume that \( \sigma_\ell^2 \) is known or can be estimated (see e.g. Section B3). Our approach consists of the following steps:
1. We compute the LS estimator $\hat{\Theta}_p$. For the GDFM, we use the FHLZ estimator.

2. Since $E\|N_{\xi,T}\|^2 = \text{tr}(\Omega)$, we approximate $E\|N_{\xi,T}\|^2$ by $\text{tr}(\hat{\Omega})$ and $E\left(\frac{\|N_{\xi,T}\|^2}{N}\right)$ by $\frac{\text{tr}(\hat{\Omega})}{N}$. Here $\hat{\Omega}$ is the empirical covariance matrix of the residuals $Y_t - \hat{\Theta}_p Y_{t-1}$, which can generally be computed through

$$\hat{\Omega} = \hat{\gamma}(0) - \hat{\Theta}_p \hat{\Theta}_p' .$$

For FHLZ, we have for every sub-VAR system

$$\hat{\Omega}_i = \hat{\gamma}_i(0) - \hat{A}_i \hat{A}_i' .$$

3. We approximate $E[\text{tr}(\Theta (\sum_{t=K+1}^{m} \hat{A}_t \hat{e}_t \hat{e}_t') \Theta')]$ by $\text{tr}(\hat{\Theta}_p (\sum_{t=K+1}^{m} \hat{A}_t \hat{e}_t \hat{e}_t') (\hat{\Theta}_p)' )$ and $E \text{tr}(\Theta P \Theta')$ by $\text{tr}(\hat{\Theta}_p P_k (\hat{\Theta}_p)' )$.

4. The steps above, lead to estimators $\hat{\sigma}_i^2$, $0 \leq i \leq 3$. The sum of these estimators leads to $\hat{\Sigma}(K)$. Then, we select the trimming value $K = K(p)$ as

$$K(p) = \arg \min_{1 \leq K \leq dp} \hat{\Sigma}(K).$$

**B.2. Choosing $K$ with CV**

We suggest the following $C$-fold cross validation algorithm:

1. Divide the training samples into $C$ subsamples of equal size.
2. For $c = 1, \ldots, C$, estimate $\hat{\Theta}^{(K)}_p$, $K = 1, \ldots, m$ using the whole training sample except the $c$-th subsample.

   Compute the prediction error $e_{ckp}$ obtained by predicting the values of the $c$-th subsample using $\hat{\Theta}^{(K)}_p$.

3. Then we use $\hat{K} = \arg \min_{1 \leq K \leq m} \sum_{c=1}^{C} e_{ckp}^2$.

   If one ignores the order of the VAR, one can compute $\hat{\Theta}^{(K)}_p$ for different $p$ and deduce the $K$ and $p$ that minimize $\sum_{c=1}^{C} e_{ckp}^2$.

   Using the CV algorithm in context of a GDFM we need reference values of the common component to compute prediction errors. For the $j$-th iteration of the CV algorithm, we use as reference the average of the $N_{rep}$ TRIM estimators of $\hat{X}_t$ and the $j - 1$ first estimators obtained via CV. The reference is thus changed at every iteration.

**B.3. Estimation of the Noise Variance in a GDFM**

Using the TRIM algorithm requires an estimator of the noise variance $\sigma^2_{\epsilon}$. Our approach to this problem is to obtain a preliminary estimator $\hat{\sigma}^2_{\epsilon}$ using the FHLZ algorithm, which is then sequentially updated. More precisely, we proceed as follows: compute $\hat{X}^{(n)}_t = (\hat{A}^{(L)}_t)^{-1} \hat{R}^{(L)}_t \hat{u}_t$ for $t \leq T$ and for the $n$ selected sub-processes as it is described in Section 3.1. Repeat this step $N_{rep}$ times for different partitions of the cross-section. Thus for each $i$ and $t$ we have $N_{rep}$ different approximations for $X_t$ from which we calculate empirical variances. Those empirical variances in turn are averaged over time and cross-section and yield the preliminary estimator $\hat{\sigma}^2_{\epsilon}$. Now the actual procedure starts. With the preliminary estimator $\hat{\sigma}^2_{\epsilon}$, we are now able to apply the TRIM algorithm in step 5 of the FHLZ estimation algorithm. We do again $N_{rep}$ iterations. After the $k$th iteration we have $k$ different approximations for $X_t$ which we use to calculate an updated estimator for $\sigma^2_{\epsilon}$.
B.4. A Modified BIC Criterion

FHLZ have suggested to use the BIC criterion to choose the order $\hat{p}_0$. We propose to adapt this criterion to the singular VAR case. For a regular VAR of dimension $q + 1$, the criterion takes the form

$$BIC(p) = \sum_{j=1}^{q+1} \log(\lambda_j^i) + \log(T)(q + 1)^2p/T,$$

where $\lambda_j^i$ is the $j$th eigenvalue of the covariance matrix of the residuals of (3.4). Imposing that the rank of this covariance matrix is $q$ suggests to expand the sum above only from 1 to $q$. The difference between both can lead to very different results, due to the degeneracy of the logarithm in zero.

APPENDIX C. AN EXAMPLE FROM POLLUTION LEVEL MONITORING

Allen et al. (2012) investigate the difference of measurements of particulate matter PM10 or PM2.5 (the numbers 10 and 2.5 refer to the maximal diameter of the particles in micrometer) made with the Patashnick Tapered Element Oscillating MicroBalance (TEOM) method and time-integrated samplers (gravimetric methods which are officially approved by the U.S. Environmental Protection Agency). The TEOM method is attractive because it is comparably cheap and provides direct and thus near-continuous measurements of PM. However, during the TEOM measuring process there is a certain loss of semi-volatile particulate mass, and hence the measurements may differ from the more expensive gravimetric methods. The differences depend on aerosol composition of ambient air, which in turn are influenced by the temperature, relative humidity, and so on. This means that the actually recorded data by TEOM, let us say $\hat{Y}_t$, need to be preprocessed before they can be officially used (e.g. to record exceedances of limit values). Let us denote the processed and corrected data by $Y_t$. Quite naturally, when fitting a statistical model one will work with the processed data $Y_t$. If the processing is not instantaneous, however, timely predictions can only be computed with unprocessed data $\hat{Y}_t$. This means that the stationary model related to the process $(Y_t)$ is loaded with the predictor variables $\hat{Y}_t$. This leads to the question of how big the impact of the processing is, that is, how much $\hat{\Theta}Y_t$ deviates from $\hat{\Theta}Y_t$. In terms of (1.1) we have $\epsilon = Y_t - \hat{Y}_t$. As we show in this article, the impact can be surprisingly large even for very small $\epsilon$.

APPENDIX D. TABLES

| TABLE D1. $\rho = 0.1$, $T = 120$ |
|----------------------------------|
| $d$  | $p = 1$ | $p = 2$ | $p = 3$ |
|------|---------|---------|---------|
| $L = 2$  | RE | 344.63 | 1.3 | 1.05 | $>10^3$ | $>10^4$ | 490.86 | $>10^3$ | $>10^3$ | $>10^3$ |
|       | RE | 0.45  | 0.59  | 0.68  | 0.36  | 0.51  | 0.6   | 0.36  | 0.48  | 0.57  |
|       | RE | 0.35  | 0.29  | 0.25  | 0.17  | 0.14  | 0.11  | 0.14  | 0.08  | 0.06  |
|       | RE | 1    | 1.03  | 1    | 1.01  | 63.91 | 14.49 | 3.9   | $>10^3$ | $>10^3$ | $>10^3$ |
|       | RE | 0.72  | 0.79  | 0.82  | 0.66  | 0.74  | 0.78  | 0.65  | 0.72  | 0.76  |
|       | RE | 0.36  | 0.31  | 0.26  | 0.19  | 0.14  | 0.11  | 0.19  | 0.08  | 0.07  |
| $L = 1$  | RE | 1    | 1    | 1    | 1.05  | 1.03  | 1.04  | 2.06  | 1.54  | 1.11  |
|       | RE | 0.94  | 0.93  | 0.94  | 0.91  | 0.9  | 0.91  | 0.89  | 0.9  | 0.89  |
|       | RE | 0.4  | 0.33  | 0.27  | 0.19  | 0.15  | 0.11  | 0.11  | 0.08  | 0.06  |
| TABLE D2. $\rho = 0.4$, $T = 120$ |
|-----------------|--------|--------|--------|--------|
| $d$             | $p = 1$ | $p = 2$ | $p = 3$ |
| $L = 2$         | RE     | $234.88$ | 1.01 | 1 | $>10^3$ | $10^5$ | $>10^3$ | 35.53 |
|                 | RE$_{12}$ | 0.74 | 0.89 | 0.94 | 0.52 | 0.67 | 0.75 | 0.44 | 0.58 | 0.67 |
|                 | RE$_{CV}$ | 0.57 | 0.63 | 0.66 | 0.32 | 0.39 | 0.41 | 0.27 | 0.33 | 0.33 |
| $L = 1$         | RE     | 1 | 1 | 1 | 1.08 | 1.02 | 1.01 | 4.07 | 1.51 | 1.38 |
|                 | RE$_{12}$ | 0.96 | 0.98 | 0.99 | 0.8 | 0.86 | 0.87 | 0.74 | 0.8 | 0.83 |
|                 | RE$_{CV}$ | 0.78 | 0.72 | 0.73 | 0.5 | 0.5 | 0.47 | 0.39 | 0.39 | 0.36 |
| $L = 0$         | RE     | 1 | 1 | 1 | 1 | 1 | 1 | 1.01 | 1.01 | 1.01 |
|                 | RE$_{12}$ | 0.99 | 0.99 | 1 | 0.95 | 0.95 | 0.96 | 0.93 | 0.93 | 0.92 |
|                 | RE$_{CV}$ | 0.82 | 0.79 | 0.79 | 0.58 | 0.55 | 0.52 | 0.46 | 0.42 | 0.38 |

| TABLE D3. $\rho = 0.8$, $T = 120$ |
|-----------------|--------|--------|--------|--------|
| $d$             | $p = 1$ | $p = 2$ | $p = 3$ |
| $L = 2$         | RE     | 1.06 | 1.01 | 1 | 116.8 | 1.06 | 1.01 | $>10^3$ | 1.5 | 1.07 |
|                 | RE$_{12}$ | 0.96 | 0.99 | 0.99 | 0.77 | 0.87 | 0.9 | 0.65 | 0.75 | 0.8 |
|                 | RE$_{CV}$ | 0.88 | 0.93 | 0.95 | 0.66 | 0.74 | 0.75 | 0.5 | 0.58 | 0.61 |
| $L = 1$         | RE     | 1.02 | 1.01 | 1 | 1.01 | 1 | 1.01 | 1.04 | 1.01 | 1.01 |
|                 | RE$_{12}$ | 1 | 1 | 1 | 0.94 | 0.96 | 0.97 | 0.86 | 0.9 | 0.91 |
|                 | RE$_{CV}$ | 0.95 | 0.96 | 0.97 | 0.81 | 0.83 | 0.83 | 0.7 | 0.7 | 0.68 |
| $L = 0$         | RE     | 1.01 | 1.01 | 1 | 1 | 1.01 | 1 | 1 | 1 | 1 |
|                 | RE$_{12}$ | 1.01 | 1.01 | 1 | 0.99 | 0.99 | 0.99 | 0.97 | 0.96 | 0.96 |
|                 | RE$_{CV}$ | 0.99 | 0.98 | 0.98 | 0.89 | 0.88 | 0.86 | 0.8 | 0.77 | 0.72 |

| TABLE D4. $\rho = 0.1$, $T = 240$ |
|-----------------|--------|--------|--------|--------|
| $d$             | $p = 1$ | $p = 2$ | $p = 3$ |
| $L = 2$         | RE     | 549.93 | 1.28 | 1.03 | $>10^3$ | $10^4$ | 780.98 | $>10^3$ | $10^3$ | $10^5$ |
|                 | RE$_{12}$ | 0.44 | 0.59 | 0.68 | 0.35 | 0.51 | 0.62 | 0.34 | 0.5 | 0.59 |
|                 | RE$_{CV}$ | 0.33 | 0.34 | 0.31 | 0.22 | 0.19 | 0.17 | 0.2 | 0.12 | 0.1 |
| $L = 1$         | RE     | 1.03 | 1.01 | 1.01 | 138.68 | 10.93 | 3.42 | $>10^3$ | $10^3$ | $>10^3$ |
|                 | RE$_{12}$ | 0.73 | 0.78 | 0.83 | 0.66 | 0.74 | 0.78 | 0.66 | 0.74 | 0.77 |
|                 | RE$_{CV}$ | 0.44 | 0.37 | 0.33 | 0.24 | 0.19 | 0.17 | 0.16 | 0.12 | 0.1 |
| $L = 0$         | RE     | 1 | 1 | 1 | 1.05 | 1.02 | 1.02 | 1.24 | 1.05 | 1.5 |
|                 | RE$_{12}$ | 0.94 | 0.95 | 0.95 | 0.91 | 0.91 | 0.91 | 0.92 | 0.91 | 0.91 |
|                 | RE$_{CV}$ | 0.48 | 0.42 | 0.35 | 0.24 | 0.2 | 0.17 | 0.16 | 0.13 | 0.1 |

| TABLE D5. $\rho = 0.8$, $T = 240$ |
|-----------------|--------|--------|--------|--------|
| $d$             | $p = 1$ | $p = 2$ | $p = 3$ |
| $L = 2$         | RE     | 4.03 | 1.01 | 1.01 | 11.14 | 1.03 | 1.01 | $>10^4$ | 1.81 | 1.08 |
|                 | RE$_{12}$ | 0.92 | 0.96 | 0.98 | 0.73 | 0.8 | 0.84 | 0.58 | 0.65 | 0.68 |
|                 | RE$_{CV}$ | 1.03 | 1.01 | 1.01 | 1.01 | 1 | 1 | 1.03 | 1.02 | 1.01 |
| $L = 1$         | RE     | 1.02 | 1.01 | 1.01 | 0.96 | 0.96 | 0.97 | 0.89 | 0.91 | 0.92 |
|                 | RE$_{12}$ | 0.99 | 0.99 | 0.99 | 0.88 | 0.89 | 0.87 | 0.76 | 0.78 | 0.77 |
|                 | RE$_{CV}$ | 1.02 | 1.01 | 1.01 | 1 | 1 | 1 | 1 | 1 | 1 |
| $L = 0$         | RE     | 1 | 1 | 1 | 0.93 | 0.93 | 0.92 | 0.97 | 0.97 | 0.97 |
|                 | RE$_{12}$ | 1 | 1 | 1 | 0.93 | 0.93 | 0.92 | 0.97 | 0.97 | 0.97 |
|                 | RE$_{CV}$ | 1 | 1 | 1 | 0.93 | 0.93 | 0.92 | 0.97 | 0.97 | 0.97 |
TABLE D6. Median of $\frac{Q_{q0}^{CV}}{Q_{q0}^{TRIM}}$ (and standard deviation of it) across repetitions

| N/T  | 60   | 120  | 240  | 480  |
|------|------|------|------|------|
| 30   | 0.95 (0.23) | 0.91 (0.25) | 0.89 (0.24) | 0.87 (0.32) |
| 60   | 0.95 (0.16) | 0.89 (0.18) | 0.85 (0.19) | 0.86 (0.21) |
| 120  | 0.93 (0.13) | 0.89 (0.15) | 0.83 (0.16) | 0.81 (0.17) |
| 240  | 0.93 (0.13) | 0.87 (0.14) | 0.81 (0.15) | 0.77 (0.16) |

$q = 2$

| N/T  | 60   | 120  | 240  | 480  |
|------|------|------|------|------|
| 30   | 1.04 (0.19) | 0.99 (0.19) | 0.93 (0.2) | 0.92 (0.21) |
| 60   | 1.06 (0.15) | 1 (0.14)  | 0.92 (0.16) | 0.85 (0.18) |
| 120  | 1.06 (0.12) | 0.98 (0.13) | 0.88 (0.14) | 0.81 (0.15) |
| 240  | 1.07 (0.1) | 0.97 (0.1)  | 0.85 (0.13) | 0.77 (0.13) |