Published in Proceedings of the Demand Forecasting Workshop of the 32nd International Conference on Machine Learning which should be cited to refer to this work

Learning coherent Granger-causality in panel vector autoregressive models

Magda Gregorová[†§] MAGDA.GREGOROVA@HESGE.CH
Alexandros Kalousis[†§] ALEXANDROS.KALOUSIS@HESGE.CH
Stéphane Marchand-Maillet[§] STEPHANE.MARCHAND-MAILLET@UNIGE.CH
[†]University of Applied Sciences and Arts of Western Switzerland, Geneva, Switzerland
[§]University of Geneva, Switzerland

Abstract

We consider the problem of forecasting multiple time series across multiple cross-sections based solely on the past observations of the series. We propose to use panel vector autoregressive model to capture the inter-dependencies on the past values of the multiple series. We restrict the panel vector autoregressive model to exclude the cross-sectional relationships and propose a method to learn models with sparse Granger-causality structures coherent across the panel sections. The method extends the concepts of group variable selection and support union recovery into the panel setting by extending the group lasso penalty (Yuan & Lin, 2006) into matrix output regression setting with 3d-tensor of model parameters.

1. Introduction

In many demand forecasting applications we wish to forecast multiple time series across several cross-sections. For example, for managing the supply of a large retail chain we wish to be able to forecast the demand for the individual retail products (bread, milk, umbrellas, etc.) across the supply units (supermarkets or some higher geographical units).

In this paper we assume that for the modelling and predicting the future demand the only data available to the forecaster is the past demand data (more specifically past sales data[1]). In result, the model needs to be build only over these without any exogenous variables (such as macroeconomic indicators, weather forecasts, etc.) entering.

[1]The discussion of the problems related to estimating demand based on the sales data is out of the scope of this paper.

Accepted contribution for the *ICML2015 Workshop on Demand Forecasting*.

To cater for the cross-sectional structure of the data, we will use the panel extension of the vector autoregressive model (PVAR) which is a well established model in the time series literature, e.g. (Lütkepohl, 2005), and which can capture the linear (inter)-dependencies on past observed values and therefore allows for extrapolating the demand based solely on the historical data (see section 1.1).

The VAR representation of the models can also provide a useful insight into the Granger-causal (G-causal) relationships amongst the multiple time series. In brief, time series X is said to G-cause time series Y if Y can be predicted better using the past values of X (see section 1.2). The graph of such relationships is naturally captured within the parameters matrix of a VAR model. When learning VARs, zero constraints can be imposed on the parameters matrix to restrict the model learning to G-causal graphs corresponding to the specific domain theory (e.g. macroeconomic theory). However, the problem of discovering the G-causal graphs in the absence of such domain driven assumptions is not addressed by the state-of-the-art methods in a principled manner and most often the learned VARs have fully connected G-graphs.

The graph of such relationships is naturally captured within the parameters matrix of a VAR model. When learning VARs, zero constraints can be imposed on the parameters matrix to restrict the model learning to G-causal graphs corresponding to the specific domain theory (e.g. macroeconomic theory). However, the problem of discovering the G-causal graphs in the absence of such domain driven assumptions is not addressed by the state-of-the-art methods in a principled manner and most often the learned VARs have fully connected G-graphs.

In this paper we develop methods for learning panel VARs with consistent sparse G-causal graphs across the panels. The sparsity motivation follows the Occam's razor principal for a preference of simple models: VARs typically suffer from the small-sample-high-dimensionality problem and sparse learning helps to tackle the overfitting issues; sparse G-causal graphs have clear interpretational advan-

tages. The panel consistency is motivated by the specific structure of the panel problem itself. Essentially, each of the cross-sections in the panel is an instantiation of a "similar" multivariate random process. Obviously, assuming the "same" generating process across the panel is far too naïve to be realistic (though still sometimes used). Therefore we allow the models to differ across the sections, while we assume that the underlying G-causal relationships are likely to be shared by (at least some of) the models. This assumptions also adds power to the model learning and G-causal graphs discovery following the multi-task learning paradigms (e.g. (Evgeniou et al., 2005)).

Various degrees and forms of the panel similarity have been considered in the PVAR literature ((Canova & Ciccarelli, 2013) has a recent review). However, as to our knowledge we are the first to propose the joint G-causal learning in PVARs.

1.1. Panel vector autoregressive model

In this paper, we focus on the vector autoregressive models (VARs) and their panel extension. While there are certainly many other valid time series modelling approaches (e.g. state space models, stochastic dynamic equations, recurring neural networks, etc.), each having its advantages and disadvantages in particular settings, the aim of this paper is not a survey or a comparative analysis of all of these neither a search for the "ultimate" time series modelling approach. Instead, we show here how VARs, which are amongst the most common tools for modelling and forecasting sets of multiple time series, can be adapted to address the specific problem of forecasting cross-sectional multivariate time series, and we develop methods to address some of the weaknesses of VARs arising in such settings.

We first state the general form of a panel VAR model: for a set of K time series each observed across Z cross-sections at T synchronous equidistant time points we write the panel VAR for all $t \in \mathbb{N}_T,\ k \in \mathbb{N}_K$ and $z \in \mathbb{N}_Z$ as

$$y_{t,k,z} = \alpha_{k,z}(t) + \sum_{lij}^{pKZ} \left(w_{l,i,j}^{k,z}\, y_{t-l,i,j}\right) + \epsilon_{t,k,z}, \qquad (1)$$

where $\alpha_{k,z}(t)$ comprises all the deterministic components (constants, polynomial trends in time, seasonal dummies, etc.), p is the number of lags, and $\epsilon_{t,k,z}$ is a white noise process such that for each t, the $K \times Z$ matrix $\boldsymbol{\mathcal{E}}_t$ has a matrix-variate normal distribution with $E(\mathcal{E}_t) = \mathbf{0}$, $E(\mathcal{E}_t'\,\mathcal{E}_s) = 0$ (independence in time), $E(\mathcal{E}_t'\,\mathcal{E}_t) = \Sigma_K$, and $E(\mathcal{E}_t\,\mathcal{E}_t') = \Sigma_Z$.

Note that the deterministic part $\alpha_{k,z}(t)$ of the model is section/series specific and therefore allows for variations between the models (e.g. variation in time series mean levels across the sections). However, in the following we will focus on the stochastic part of the model and therefore will work with the detrended form of the model where the $\alpha_{k,z}(t) = 0$ and therefore can be dropped.

The PVAR in its most general formulation (1) is typically highly over-parametrised: the total number of parameters for each time series in each cross section (for each k and z) is KZp which is usually much higher than the number of observations T. To limit the over-parametrisation, we will restrict the model to cross-sectional independence by setting $w_{l,i,j}^{k,z} = 0$ when $j \neq z$ (when regressing on series from other cross-section). Such an assumption is rather strong but may often be realistic in real-life setting when we do not expect the cross-sections to interact[2]. In result, we reduce the panel VAR into a set of Z standard VAR models with Kp parameters per series.

$$y_{t,k,z} = \sum_{li}^{pK} \left(w_{l,i,z}^{k,z}\, y_{t-l,i,z}\right) + \epsilon_{t,k,z}, \qquad (2)$$

Though the number of estimated parameters is now much lower, it is still typically much larger than the number of observations ($Kp \gg T$) and therefore some form of regularization is needed. The regularization we explore in this paper constraints the model learning towards sparse G-causal graphs coherent across the panel sections.

1.2. Granger causality

In (Granger, 1969) the definition of causality is based on predictability of the series. A series X is said to Granger-cause another series Y if Y can be better predicted (in terms of having lower variance of the predictive error) using the past of Y than without it. This notion of causality can be extended to a set of series so that a set of series $\{Y_1, \ldots, Y_l\}$ is said to Granger-cause series Y_k if Y_k can be better predicted using the past values of the set.

The Granger-causal relationships can be described by a directed graph $G = \{\boldsymbol{\mathcal{V}}, \boldsymbol{\mathcal{E}}\}$ in which the set of vertices represents the time series in the system, and a directed edge $e_{l,k}$ from v_l to v_k means that time series l Granger-causes time series k.

In VARs, the Granger-causal relationships are captured within the parameters w of model (1). When $w_{l,i,j}^{k,z} \neq 0$ for any l we say that series i from panel j G-causes series k from panel z. Note that for the restricted model form (2) by putting $w_{l,i,j}^{k,z} = 0$ when $j \neq z$ we effectively exclude all G-causal relationships across the sections. In this way, we learn Z disconnected G-causal graphs, one for each section.

[2]Modelling of full unrestricted PVAR as well as of deterministic and stochastic trends are topics for future research

2. Learning coherent Granger-causal graphs

For better clarity we rewrite model (2) in the standard matrix form of a multi-output regression. For each cross-section $z \in \mathbb{N}_Z$ we write the VAR as

$$\mathbf{Y}_z = \mathbf{X}_z \mathbf{W}_z + \mathbf{E}_z \tag{3}$$

where $\mathbf{Y}_z$ is the $T \times K$ output matrix, $\mathbf{X}_z$ is the $T \times Kp$ input matrix so that each row t of the matrix is a Kp long vector with p lagged values of the K time series in the same cross section as inputs $\mathbf{x}_{t,\cdot} = (y_{t-1,1}, y_{t-2,1}, \ldots, y_{t-p,1}, y_{t-1,2}, \ldots, y_{t-p,K})'$[3] (dropping the fixed z indexes for ease of reading). $\mathbf{W}_z$ is the corresponding $Kp \times K$ matrix where each column is a model for a single time series. The $T \times K$ error matrix $\mathbf{E}_z$ is a random noise matrix with independent rows $\mathbf{e}_{t,\cdot} \sim N(\mathbf{0}, \mathbf{\Sigma})$.

This structure of the VAR models is particularly advantageous for the G-causal discovery since the G-causal restrictions ($w^{k,z}_{l,i,z} \neq 0$ for any l) translate directly into block-sparsity in the $\mathbf{W}_z$. A fictitious example of the $\mathbf{W}_z$ matrices is depicted in figure 1. The top part of the picture illustrates how the $\mathbf{W}$ can be organised into a 3d-tensor. The shaded squares are the non-zero blocks corresponding to the G-causal relationships between the series in the panel. The figure also illustrates how these non-zero blocks "drill" through the 3d-tensor so that the individual cross sections are coherent in terms of their block-sparsity. Also, note that the block diagonal elements of the $\mathbf{W}_z$ matrix capture the dependency of each series on its own history which falls out of the usual G-causal definition: as usual in multivariate time series analysis, series is always expected to depend on its own past and therefore the block-diagonal elements are non-zero.

From the above outlined link between the G-causal graphs and the VAR parameters matrix it should be clear that the problem of discovering sparse G-causal graphs can be seen as learning with group variable selection where the groups are formed by all lags of a single input time series. Similarly, discovering G-causal graphs coherent across the panel cross-sections corresponds to learning section models with coherent block-sparsity patterns in their parameter matrices $\mathbf{W}_z$.

We propose to learn the parameters of the restricted PVAR in eq. (2) by minimising a regularised loss problem

$$\underset{W}{\operatorname{argmin}}\, L(\mathbf{W}^{3d}) + \lambda R(\mathbf{W}^{3d}), \tag{4}$$

where $\mathbf{W}^{3d}$ is the 3d-tensor of the PVAR parameters (as in the top of figure 1), $L(\mathbf{W}^{3d})$ is the squared error loss for

[3] By convention, all vectors in this paper are column vectors.

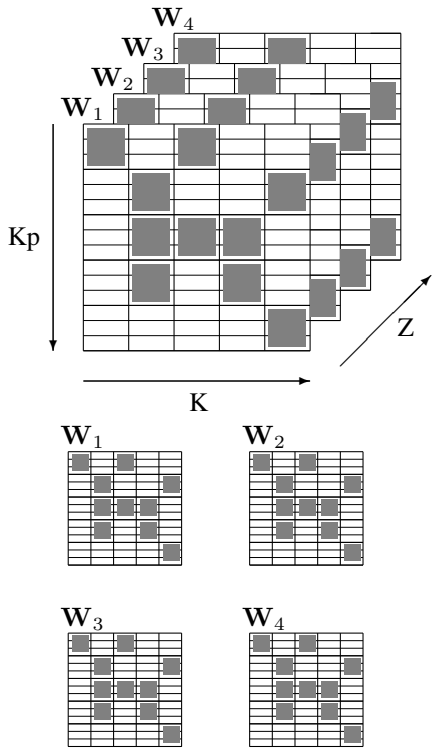

Figure 1. Schematic structure of the parameters of a panel autoregressive model with cross-sectional independece ($w^{k,z}_{l,i,j} = 0$ when $j \neq z$) and sparse G-causal graphs coherent across the panel sections.

the Z sectional models

$$L(\mathbf{W}^{3d}) := \sum_{z}^{Z} ||\mathbf{Y}_z - \mathbf{X}_z \mathbf{W}_z||_F^2, \tag{5}$$

$\lambda \geq 0$ is the regularization parameter and

$$R(\mathbf{W}^{3d}) := \sum_{k}^{K} \sum_{b \neq k}^{K} ||\tilde{\mathbf{W}}_{b,k}||_F + \sum_{k}^{K} ||\tilde{\mathbf{W}}_{k,k}||_F^2 \tag{6}$$

Here, the $\tilde{\mathbf{W}}_{b,k}$ is the $p \times Z$ matrix constructed by concatenating the vectors $\tilde{\mathbf{w}}_{b,k,z} = (w^{k,z}_{1,b,z}, w^{k,z}_{2,b,z}, \ldots, w^{k,z}_{p,b,z})'$ across the cross-sections z (one "drill" through the parameters tensor in figure 1), and $||.||_F$ is the matrix Frobenious norm.

The first term in (6) is a special case of the ℓ_1/ℓ_2 block-norm of (Yuan & Lin, 2006) known as *group lasso* adapted to the 3d-tensor models with p-large groups of variables. As such, it has also similar sparsity effects: it encourages common group-sparsity across the cross-section models. The ℓ_1/ℓ_2 block-norm is only applied to the non-diagonal blocks of each of the $\mathbf{W}_z$ in eq. 6, and standard ℓ_2

norm is used for the block-diagonals instead. This shrinks the block-diagonal parameters without encouraging sparsity which corresponds to the usual time series modelling assumption that each series depends on its own history.

We wish to emphasize here that by using the regularization term in eq. 6 we learn Z different models parametrised by the model matrices $\{\mathbf{W}_z : z \in \mathbb{N}_Z\}$, one for each cross-section. The models are, however, similar in the sense of having the same underlying G-causal structure reflected in the block sparsity of the parameters matrices.

Problem (4) with loss (5) regularized by (6) is a convex (though non-differentiable) problem that can be solved by standard optimisation approaches (e.g. proximal gradient methods).

2.1. Learning cross-sectional clusters

One of the major limitations of the model outlined in section 2 is the assumption of full coherency of the G-causal graphs across the sections. Indeed, a more realistic assumption would be that some cross-sections are more similar than others. Learning clusters of section models is thus a natural and certainly useful extension which could provide useful insight into the specific market characteristics of the supply chain units.

To allow for leaning the cross-sectional model clusters we propose to re-parametrise model (2) and its matrix multi-output regression form (3) so that each of the block sub-vectors of $\tilde{\mathbf{w}}_{b,k,z} = \gamma_{b,k,z}\tilde{\mathbf{v}}_{b,k,z}$, where $\tilde{\mathbf{v}}_{b,k,z}$ is a p-long vector (the same size as $\tilde{\mathbf{w}}_{b,k,z}$) and $\gamma_{b,k,z}$ is a scalar. Here, we will use the γ's to control the block-sparsity of the models. Essentially, $\gamma_{b,k,z} = 0$ implies $\tilde{\mathbf{w}}_{b,k,z} = \mathbf{0}$.

In the model in section 2 $\gamma_{b,k,z} = \gamma_{b,k}$, $\forall z \in \mathbb{N}_Z$ and $\forall b, k \in \mathbb{N}_K$. In the cluster version we wish to allow for $\gamma_{b,k,z} = \gamma_{b,k,c}$, $\forall z \in \mathbb{S}_c$ and $\forall b, k \in \mathbb{N}_K$, where $\mathbb{S}_c$ is the set of cross-section indices belonging to the same cluster so that $\cup_c \mathbb{S}_c = \mathbb{N}_Z$.

We will matricize the 3d-tenosor $\mathbf{\Gamma}^{3d}$ along the z dimension to construct a $K^2 \times Z$ matrix $\tilde{\mathbf{\Gamma}}$. The full consistency model in section 2 has $\tilde{\boldsymbol{\gamma}}_{i,z} = \tilde{\boldsymbol{\gamma}}_i$, $\forall z \in \mathbb{N}_Z$, $\forall i \in \mathbb{N}_{K^2}$ while, in the cluster version, we have $\tilde{\boldsymbol{\gamma}}_{i,z} = \tilde{\boldsymbol{\gamma}}_{i,c}$, $\forall z \in \mathbb{S}_c$, $\forall i \in \mathbb{N}_{K^2}$.

For leaning the panel VARs with sparse G-causality graphs coherent across the cross-section clusters we propose to assume that the model block-sparsity patterns within $\tilde{\mathbf{\Gamma}}$ lie in a low dimensional subspace so that the cross-sectional models can be seen as linear combinations of cluster prototypes with specific sparse G-causal structures.

We will reformulate the learning problem (4) as a minimisation with respect to the newly defined $\tilde{\mathbf{V}}$ and $\tilde{\mathbf{\Gamma}}$ of the regularised functional

$$\begin{aligned} &\textstyle\sum_z^Z ||\mathbf{Y}_z - \mathbf{X}_z(\tilde{\mathbf{\Gamma}}_\mathbf{z} \circ \tilde{\mathbf{V}}_z)||_F^2 + \\ &+ \textstyle\sum_z^Z \left(\lambda_1 ||\tilde{\mathbf{V}}_z||_F^2 + \lambda_2 ||\tilde{\boldsymbol{\gamma}}_\mathbf{z}||_1\right) \\ &\text{s.t. } rank(\tilde{\mathbf{\Gamma}}) \leq r, \end{aligned} \tag{7}$$

where $\tilde{\boldsymbol{\gamma}}_z$ is the z column of matrix $\tilde{\mathbf{\Gamma}}$, $\tilde{\mathbf{\Gamma}}_\mathbf{z}$ is the $\tilde{\boldsymbol{\gamma}}_\mathbf{z}$ vector reshaped and replicated to a $Kp \times K$ matrix to correspond in shape to $\tilde{\mathbf{V}}_z$ matrix, and $\circ$ indicates the element-wise product.

References

Canova, Fabio and Ciccarelli, Matteo. Panel vector autoregressive models: a survey. Working paper series 1507, European Central Bank, January 2013.

Evgeniou, Theodoros, Micchellli, Charels A., and Pontil, Massimiliano. Learning multiple tasks with kernel methods. *Journal of Machine Learning Research*, 6:615–637, 2005.

Granger, CWJ. Investigating causal relations by econometric models and cross-spectral methods. *Econometrica: Journal of the Econometric Society*, 37(3):424–438, 1969.

Lütkepohl, Helmut. *New introduction to multiple time series analysis*. Springer-Verlag Berlin Heidelberg, 2005.

Yuan, Ming and Lin, Yi. Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 68(1):49–67, 2006.