Regularized Estimation in High-Dimensional Vector Autoregressive Models

Research Proposal

submitted to:
London School of Economics (Statistics Department)

submitted by:
Ratmir Miftachov

Contents

1 Methodology

Vector Autoregressions (VARs) are capable of capturing complex dynamics among the variables of a system, cross sectional as well as temporal. They have shown to be particularly useful for describing the behavior of macroeconomic or financial time series data. Although this proposal is concerned with the macroeconometric framework as an area of application, the method of interest is more universally applicable. A major objective of VAR models is to perform an accurate prediction. Thus, the most obvious research question immediately emerges as:

- Can we improve the prediction accuracy using a novel regularization method, compared to well-known benchmark models?

A major disadvantage of large VAR models is the Curse of Dimensionality. Often, the data provided to the researcher in the macroeconometric literature is measured on a relatively low frequency, resulting in a sparse number of observations. Additionally, the number of variables in a VAR Model can get large quite quickly. The first publication on VAR models, Sims (1980), already mentions this emerging issue. Consequently, the author suggests the need of "mean-square-error shrinking devices" to circumvent the problem of overfitting. However, if the researcher ignores this requirement, it results in severely distorted predictions. Thus, a need for shrinkage methods emerges in order to reduce the MSFE. These methods range from bayesian type of approaches, to factor model approaches as well as to machine learning techniques like regularization. Particularly, this research proposal seeks to contribute on the bridge between regularization techniques and the literature on VAR models.

We introduce the novel feature weighted elastic net (fenet) regularization, following the working paper of Tay, Aghaeepour, Hastie and Tibshirani (2020). This method can be seen as a generalization of the elastic net regularization, allowing the researcher to incorporate external information on the data in the minimization problem. The objective of including additional information is to improve prediction performance. This type of side-information, a priori known to the researcher, is also known as "co-data" in the literature. By being a hybrid model, which nests the ridge and the lasso penalization, the fenet includes the main advantages of both models. First, the l_1 norm enables the coordinate descent algorithm to conduct variable selection. Second, the l_2 norm allows to handle highly collinear variables better than the bare lasso model.

This proposal is concerned with the implementation of the feature weighted elastic net penalization into the VAR model. Hereby, the VAR system reduces to a seemingly unrelated regression (SUR) model, since each equation has the same set of regressors. Thus, we validate the tuning parameters and estimate the coefficients equation-wise. To the best of our knowledge, we are the first to extend the fenet penalization to the VAR model, formally, as well as empirically.

By the construction of the fenet penalization, it assigns an additional weighting to each variable in the regularization term of Lagrange function 2. This weighting function is dependent on a pre-specified auxiliary matrix Z as well as an additional hyperparamter θ^k, resulting in a score function. Prior to the estimation, the researcher separates each coefficient into a group by using the auxiliary matrix. This separation translates the subjective belief that variables in the same group have correlated signal on the dependent variable. Thus, if a certain group of variables is important, the resulting weighting factor will be smaller for the whole group. As a consequence, the algorithm assigns the respective coefficients a relatively large magnitude and is less likely to shrink them to zero. Contrary to the group lasso, the fenet penalization does not conduct variable selection on a group level, but rather on an individual level.

As mentioned in Tay et al. (2020), the implementation of a score for each variable in the Lagrange function, is a general idea, which requires further investigation. In addition, we believe that the structure of VAR models brings even more need for the idea of co-data than a basic linear regression model. Our hypothesis is that prediction accuracy can be improved by including external information of the data into the Lagrange function for each equation of the VAR model. Consequently, we acknowledge the flexibility of imposing additional structure on the penalization term and propose different grouping schemes for the fenet VAR model. The research question narrows down to:

- Which grouping schemes can we impose and do we find superiority among them, dependent on the structure of the data?

All three schemes are motivated by major publications in the bayesian VAR literature and the regularized VAR literature. An attractive characteristic of our model is that we do not need to specify the degree of shrinkage among the variables, while simultaneously being able to impose additional structure on the penalization. The degree of shrinkage on the imposed structure is estimated automatically by an extended pathwise Coordinate Descent algorithm. In principle, the researcher is free to group the set of coefficients separately in each SUR equation as she wishes. The assigned grouping structures are allowed to overlap within an equation, leading to even more flexibility.

2 Grouping Schemes

One possibility for the researcher is to orientate herself on the bayesian literature, and utilize the ideas of certain hyperparameter specifications for prior distributions. For example, Banbura et al. (2010) specify the hyperparameter in their prior such that the ad-hoc belief of more relevant recent lags than distant lags is included. The idea is initially proposed by the Minnesota Prior, in which a hyperparameter controls the degree of shrinkage for more distant lags. Song and Bickel (2011) implemented this belief in the regularization framework using a lasso type VAR model. However, there is a major distinction between the approach the previous authors take and ours. The researcher using fenet VAR is not able to arbitrarily set the degree of shrinkage imposed on the coefficients. In a large part of the bayesian literature, as well as in Song and Bickel (2011), the model requires both, to arbitrarily decide which coefficients to shrink differently (e.g. the more distant the lag, the higher the shrinkage) and to specify the shrinkage relation among each other (the relation in shrinkage differs dependent on the size of the hyperparameter). Often, either a harmonic or geometric decay is assumed for higher lag orders. On the contrary, the only arbitrary part in fenet VAR is to assign the coefficients in groups, which we assume to have common explanatory power for the dependent variable. This belief is quantified in an auxiliary matrix. It is not possible to assign an arbitrary amount of shrinkage on the prespecified groups of coefficients in this matrix. However, this can be seen as an attractive property of the fenet VAR. The relative amount of shrinkage for each group is determined by the algorithm, not the researcher. Thus, if our belief about the importance of a specific group of coefficients is true, we will observe a higher resulting score in the output. More precisely, the fenet VAR is able to avoid the potential mistake of imposing wrong shrinkage relations among the groups.

One objective of my Master Thesis was to find suitable grouping schemes for the fenet VAR. On this occasion, I have proposed the following three schemes. The first scheme, *Scheme 1*, is the consequence of the previously motivated grouping structure. Thus, we group all variables of the same lag order together with the belief that the resulting score will be higher for more recent lags than for more distant lags.

The second scheme, *Scheme 2*, includes a similar belief to *Scheme 1*. However, the researcher is only interested in assigning a different weight between all variables of the first lag order and all variables of the remaining lag orders. Thus, we separate the first lag of the whole set of variables into the same group. The remaining coefficients represent the second group.

Motivated by another belief of Banbura et al. (2010), namely assigning the diagonal elements less shrinkage as the off-diagonal elements in the VAR, we introduce *Scheme 3*. The belief is that own lags explain more variation than the remaining variables in each equation of the system. For example, a VAR system with five lags and three independent variables, should give less shrinkage on all five lags of the first variable in the first equation. The remaining ten coefficients should obtain relatively more shrinkage since they are assumed to contain less explanatory power for the first dependent variable. However, we modify this approach and separate the remaining ten coefficients one more time, such that each individual variable has an own group, including all five lags. In other words, we assume that all lags of the same variable have a common underlying feature. Thus, we specify as many separate groups as the number of variables in each equation of our system.

3 Integration

Within my Master Thesis, I conducted prediction exercises with *Scheme 1-3* using a data set of U.S. macroeconomic variables in a high-dimensional variable setting. The empirical results have shown that the prediction performance of each scheme is dependent on variable persistency, ratio of observations to number of variables as well as signal to noise ratio. However, due to the tight time constraint of the Master Thesis, I did not conduct a more rigorous investigation of the proposed grouping schemes. Subsequently, I am keen to continue my work from a methodological point of view in a simulation setup. Particularly, I intend to compare the proposed grouping schemes in terms of prediction accuracy, for a variation of the above mentioned data properties. If a specific scheme dominates in terms of prediction accuracy, we are interested in finding an explanation for it. In addition, we want to know the subsequent implications and limitations for empirical applications. The emerging research questions are:

- Do we detect an improvement in prediction accuracy by moving into larger model sizes?

- Does prediction accuracy increase in case of additional imposed structure on the penalization term (comparison to the nested elastic net)?

- How should the researcher specify the auxiliary matrix Z and how does this choice affect the prediction outcome?

- Does a certain grouping scheme dominate? If yes, what is the reason for it? What does it imply for empirical applications?

Furthermore, several minor research questions emerge as a byproduct:

- What happens if we variate the specification of the weighting function $w_{ij}(\theta^k)$ which is used in the penalization term of equation 2? As already stated in the paper of Tay et al. (2020), there is no theoretical justification for using the present one. The current choice of the weighting function nests the elastic net penalization. However, are there other for the VAR model relevant weighting functions? What happens, if we distant the fenet penalization from the elastic net?

Based on my thesis, I have already first experience in the imposed research questions. Particularly, this working knowledge enables me to avoid pitfalls coming along the way of the research. For example, it shows that using a multi-parameter Cross-Validation procedure for hyperparameter validation is not a clever choice for the fenet VAR, since the computational costs are too expensive. A more feasible approach to validate the hyperparameter is to use information criteria like the BIC or the AIC. Given my background in econometrics, I developed a good intuition on the empirical application of my research: the practical implementation of potential discoveries would rather be a convenient task than a burden. In addition, my extensive experience in programming, especially in R, helps me to avoid hard times in the implementation of the simulation part of my work.

4 Details: Feature weighted elastic net VAR(p)

Let K be the number of variables and p the number of lags in a VAR(p) system. The kth equation can be expressed as

$$y_{kt} = v_k + \sum_{i=1}^{p}\sum_{j=1}^{K} \beta_{kj}^{i} y_{j,t-i} + u_{kt}. \tag{1}$$

Based on equation 1, we formulate the fenet VAR Lagrange function as

$$L(\theta^k, v_k, \beta_k)_{\lambda,\alpha} = \frac{1}{2}\sum_{t=1}^{T} u_{kt}^2 + \lambda \left[\sum_{i=1}^{P}\sum_{j=1}^{K} w_{ij}(\theta^k) \left[\alpha|\beta_{kj}^i| + \frac{1}{2}(1-\alpha)(\beta_{kj}^i)^2 \right] \right], \tag{2}$$

where $w_{ij}(\theta^k) = f(z_{ij}^{'}\theta^k) = \frac{\sum_{i=1}^{p}\sum_{j=1}^{K} \exp(z_{ij}^{'}\theta^k)}{pK\exp(z_{ij}^{'}\theta^k)}$ is the weight given to the i'th lag of variable j in equation k and $\theta^k = \left[\theta_1^k, \ldots, \theta_G^k\right]^{'}$ with G being the number of different sources of variable information for equation k. The auxiliary matrix $Z^{'} \in \mathbb{R}^{G \times pK}$ is expressed as $Z^{'} = \left[z_{11}, \ldots, z_{1K}, \ldots, z_{pK}\right]$. Consider the following minimalistic example using *Scheme 1* with $K = 3$ and $G = p = 2$. Then the auxiliary matrix Z as well as the score $z_{ij}^{'}\theta^k$ (e.g. lag $i = 1$ and variable $j = 1$) for equation k are

$$Z = \begin{bmatrix} z_{11}^{'} \\ z_{12}^{'} \\ z_{13}^{'} \\ z_{21}^{'} \\ z_{22}^{'} \\ z_{23}^{'} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \\ 0 & 1 \end{bmatrix} \qquad \text{and} \qquad z_{11}^{'}\theta^k = \begin{bmatrix} 1 & 0 \end{bmatrix} \begin{bmatrix} \theta_1^k \\ \theta_2^k \end{bmatrix} = \theta_1^k.$$

5 References

[1] Banbura, M., Giannone, D., Reichlin, L. (2010). Large Bayesian vector auto regressions. *Journal of Applied Econometrics*, 25(1), 71-92.

[2] Song, S., Bickel, P. J. (2011). Large vector auto regressions. *ArXiv:1106.3915v1.*

[3] Sims, C. (1980). Macroeconomics and Reality. *Econometrica*, 48, 148.

[4] Tay, J. K., Aghaeepour, N., Hastie, T., Tibshirani, R. (2020). Feature-weighted elastic net: Using features of features for better prediction. *Working Paper.*