1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
26
27
28
29
30
31
32
33
34
35
36
37
38
39
40
41
42
43
44
45
46
47
48
49
50
51
52
53
54
55
56
57
58
59
60

Response to Reviewers: VARX-L: Structured Regularization for Large Vector Autoregressions with Exogenous Variables

October 16, 2016

We thank the editors and reviewers for their constructive comments. We have carefully gone through these comments and addressed every concern in our newest revision.

Reviewer 1 (beginning with I have refereed...)

1. The paper mainly uses the lag structure to construct the penalty functions. While this is a natural idea, one could also consider grouping variables according to which sector they belong to. This is relevant in particular for the Stock and Watson type data set considered where the variables are grouped into various sectors. It is then natural to conjecture that lags of variables from the same sector are more important than lags of variables from other sectors. Could and/or should one combine grouping with respect to lag structure with grouping with respect to the type of variable?

This is an interesting observation that we have considered in our applications. It is certainly possible to group by sector/economy as in Stock/Watson. Moreover groupings can be applied to a wide variety of localized domain-specific problems. In the absence of any information about the application domain or the relationship between the series, we recommend grouping based upon the intrinsic VARX structure. We have added some clarifying remarks on page 9 of the manuscript:

In addition to the Own/Other and Lag structures, one could consider a variety of application-dependent partitions. This idea was briefly explored in Song and Bickel [2011] who posit potentially segmenting financial or economic series based upon industry or sector characteristics in addition to a lag based grouping. However, since in many applications we lack a priori information about the potential relationships of the included series, we prefer to structure our groupings around the intrinsic structure of the VARX.

3. The covariance matrix $\Sigma_u^{(l,j)}$ has a determinant of zero when $k > T$ making AIC and BIC infeasible. Perhaps it is already quite unstable even when k is large compared to T yet still $k \geq T$. This can also partially explain the rather bad performance of AIC and BIC in the simulations.

ScholarOne, 375 Greenbrier Drive, Charlottesville, VA, 22901
1

We agree that Σ_u is unstable under these circumstances. We attempt to address this concern by adding a small ridge penalty to improve the conditioning of Σ_u . In addition, across most of our simulations and applications, we operate under settings in which $k < T$. Our AIC and BIC algorithms also allow for the choice of $p = s = 0$, which simply returns the conditional mean. We have added a citation to our information criterion minimization routine as well as some further clarifications on page 16.

4. It is not correct to say that Scenarios in which the sample mean forecasts well relative to more sophisticated procedures imply weak temporal dependence. Think of the case $X_t = X$ for all $t \geq 1$ for some random variable X . In that case the sample average will be a perfect predictor of the future while the correlations between all pairs of X_t and X_s .

We agree; our aim was to raise the point that strong performance of the sample mean provides evidence that the models in question have low predictive power relative to a white noise process. As you stated, there could be a number of potential relationships in which the sample mean performs well. We have clarified this statement to read:

Strong relative performance of the sample mean indicates that the sophisticated models of interest have low predictive power relative to white noise; the cause of which could be a number of factors, including weak temporal dependence or dependence of a nonlinear nature.

5. I think it would be useful to also include a factor model in the forecast comparison as this is another type of very commonly used model in macroeconomic forecasting. I am not asking for anything fancy, just a rather plain version.

We've implemented a factor model, but found that it performed very poorly across all applications; generally on par with the sample mean. Its performance has been added to Tables 2,4,6, and 9. We do not include the factor model in our simulations as we do not expect its performance to be informative, as it is designed to excel in low rank as opposed to sparse regimes.

6. In connection to the above comment, the MC setting play heavily against AIC and BIC. BIC is designed for consistent model selection. This means that if BIC is restricted to choose from models of increasing dimension, it will always tend to choose a VAR(4) even if the first three lags are zero and only the fourth is non-zero. As many of the settings have gaps in the lag structure these play against BIC which chooses a VAR(4) which is very heavily parameterized when estimated by least squares resulting in imprecise forecasts. Of course, trying all subsets of the lags with BIC is not feasible when there are many lags so one may also argue that this is an inherent problem of BIC which it pays the price for.

We agree that BIC will generally perform poorly in such a scenario; we consider our proposed methods as a viable alternative. We have added a simulation scenario (Scenario 5 on page 31 and in section 3) in which there is no sparsity in order to compare performance under a framework better suited for information criterion based methods.

8. *What do you mean by the sparsity being completely random in scenario 1 of the simulations? Is every coefficient zero with a certain probability?*

We generated the coefficient matrix in such a manner that 10 percent of coefficients are nonzero. These coefficients were selected at random. We have added this information to the paper.

9. *Is stationarity imposed/ensured in the simulations or could some of the models be explosive?*

We generated the coefficient matrices using random numbers corresponding to the specified sparsity patterns until a stationary matrix was recovered. We added a citation describing our process.

Review 2 (beginning “This paper addresses a very...”)

1. *Given that the homoscedasticity assumption the authors make about the white noise vector in equation (2.1) on page 5 is very strong, I would like to see robustness results in the simulations for settings where u_t is allowed to show some sort of heteroscedasticity: this is a more realistic assumption in many economic and financial applications.*

We have added an additional simulation scenario that generates a VAR with a non-diagonal covariance matrix (Simulation Scenario 6 on pages 31-32 and in 3). We also cite another of our working papers which conducts an extensive simulation study of penalized VAR methods across a variety of covariance structures.

2. *How do the methods perform in settings where there are no true sparsity structures? Are they still able to perform and compare well against the different alternatives?*

We added a simulation scenario (Scenario 5 on page 31 and in 3) in which there is no sparsity and find that the VARX-L procedures still substantially outperform all benchmark models.

1.. *Given the emphasis put on the feasibility of the methods in large (high) dimensions, it is a pity that the authors report results only for moderately sized applications. I suggest that the authors consider a large dimensional application with a number of series in the order of hundreds. In my opinion the stronger forecasting power of the methods will be even more evident in such dimensions.*

We have added an additional application section in which we forecast Koop’s *medium- large* ($k = 40$) set of indicators using the *large* set as exogenous predictors $m = 128$. In the absence of parameter restrictions, this results in 26,880 variable least squares model.

Table 1 displays the strong performance of our methods in this application.

Table 1: One-step ahead MSFE and four-step ahead MSFE of $k = 40$ macroeconomic indicators (relative to sample mean) with $m = 128$ exogenous predictors $p = 4, s = 4$.

Model/VARX-L Penalty Structure	One-step ahead Out of Sample Relative MSFE	Four-step ahead Out of Sample Relative MSFE
Basic	0.7810	0.9824
Lag Group	0.8550	0.9756
Own/Other Group	0.7587	0.9680
Sparse Lag Group	0.7899	0.9575
Sparse Own/Other Group	0.7917	0.9952
Endogenous-First	0.8278	0.9657
VARX with lags selected by AIC	1.9789	2.5105
VARX with lags selected by BIC	0.9999	0.9999
BGR's Bayesian VAR	0.9062	0.9646
Factor Model	0.9758	1.0166
Sample Mean	1.0000	1.0000
Random Walk	1.6763	1.6763

2. Tables 2 to 4 are interesting but not able to fully describe the differences among the approaches: Are the differences found significant in statistical terms? I would like to see results of model confidence sets (MCS) to be able to discriminate among the models that perform the best and the ones that are clearly outperformed.

We implemented the model confidence sets procedure for our stationary macroeconomic applications and summarized the results in Table 6. In the nonstationary application, Assumption 2 of Hansen et al. [2011] (stationarity of the pairwise loss differentials) is not satisfied, so we do not report the MCS for that application. Following the default setting in the R package, we use a significance level of 0.15. This is an unusual significance level, but neither the R package nor Hansen [1998] offer any guidance toward its selection.

We report the model sets of equal predictive ability in Table 2.

Table 2: Model Sets M of Equal Predictive Ability ($\alpha = 0.15$). Within each set of models, we cannot reject the null hypothesis of equal predictive ability, though they achieve superior forecasting performance relative to all excluded models.

Application	One-Step M	Four-Step M
Stock and Watson [2005] US Macroeconomic Data ($k = 20, m = 20$)	Own/Other VARX-L Sparse Own/Other VARX-L	Own/Other VARX-L Sparse Own/Other VARX-L
Stock and Watson [2005] US Macroeconomic Data ($k = 40, m = 128$)	Own/Other VARX-L	Sparse Lag VARX-L Own/Other VARX-L Endogenous-First VARX-L BGR's Bayesian VAR
Canadian Macroeconomic Data ($k = 4, m = 20$)	Sparse Own/Other VARX-L Sparse Lag VARX-L	Sparse Own/Other VARX-L Basic VARX-L Sparse Lag VARX-L Own/Other VARX-L

2. Another interesting question that is not addressed in the Tables is about the number and type of variables that are found to be active and important for forecasting. A more detailed discussion is needed.

We agree that this is an important facet of the structured regularization framework. We have added in Tables 3,5,7, and 10 the average sparsity ratio (number of coefficients set to zero /number of potential coefficients) which gives an idea as to the relative complexity of the models returned by the VARX-L procedures.

3. Is there a way to directly compare the results (accuracy) provided in Table 4 with those reported in Table 2? Given that the application is exactly the same, it would be interesting to see which approach performs the best (non-stationarity vs. stationarity).

This is an interesting observation. In our particular application, the two sets of results are not directly comparable due to the nonstationary data encompassing a longer period as the stationarity transformation suggested by Stock and Watson [2005] is not appropriate for one of the included series, which forces us to truncate the stationary application in Quarter 3 of 2007.

One advantage of not applying stationarity transformations is that it allows us to utilize more of our data. The data used in section 3.3 extends to Quarter 4 of 2008, but one series, non-borrowed depository institutional reserves (FMRNBA), becomes negative in early 2008 due in part to changes in both monetary policy and Federal Reserve accounting [Ip, 2008]. The stationarity transformation guidelines provided by Stock and Watson [2005] for this series propose taking the first difference of logs, which is obviously not appropriate for negative values.

If the data was truly equivalent across both applications, we could compare their performance by transforming the data back to its original scale at every point in time during the forecast evaluation stage and reporting the loss. However, it is unclear what type of loss function is suitable to jointly measure performance on the untransformed scale, and a subjective choice is probably necessary for each new application, in general.

Added Simulation Scenarios

Scenario 5: No within-lag Sparsity

Next, we consider a scenario in which the entire 5×40 coefficient matrix is dense. Although the coefficient matrix is not sparse, as in the other scenarios, the information criterion based methods can impose sparsity by truncating the maximum lag orders $\hat{p}$ and $\hat{s}$. The relative magnitudes of the coefficient matrix are depicted in Figure 1. This scenario serves to give an advantage to the procedures that do not impose any within-lag sparsity: the information criterion based methods as well as BGR’s Bayesian VAR. The results from this scenario are depicted in Table 3.

Table 3: Out of sample MSFE of one-step ahead forecasts after 100 simulations: Scenario 5. Standard errors are shown in parentheses.

Model/VARX-L Penalty Structure	MSFE	MSFE Relative to Sample Mean
Basic	0.0780 (0.0010)	0.6142
Lag Group	0.0731 (0.0008)	0.5724
Own/Other Group	0.0736 (0.0009)	0.5761
Sparse Lag Group	0.0747 (0.0009)	0.5848
Sparse Own/Other Group	0.0747 (0.0009)	0.5849
Endogenous-First	0.0736 (0.0009)	0.5766
VARX with lags selected by AIC	0.1207 (0.0016)	0.9444
VARX with lags selected by BIC	0.1157 (0.0022)	0.9053
BGR’s Bayesian VAR	0.1148 (0.0018)	0.8982
Sample Mean	0.1278 (0.00296)	1.0000
Random Walk	0.2020 (0.00369)	1.5806

Table 4: Out of sample MSFE of one-step ahead forecasts after 100 simulations: Scenario 6. Standard errors are shown in parentheses.

Model/VARX-L Penalty Structure	MSFE	MSFE Relative to Sample Mean
Basic	0.7982 (0.0113)	0.1034
Lag Group	0.9089 (0.0133)	0.1177
Own/Other Group	0.8926 (0.0134)	0.1156
Sparse Lag Group	0.8833 (0.0122)	0.1144
Sparse Own/Other Group	0.8746 (0.0127)	0.1133
Endogenous-First	0.9283 (0.0135)	0.1202
VARX with lags selected by AIC	1.4237 (0.0232)	0.1844
VARX with lags selected by BIC	1.0895 (0.0185)	0.1411
BGR's Bayesian VAR	2.8876 (0.0782)	0.4022
Sample Mean	7.1789 (0.3739)	1.0000
Random Walk	4.2543 (0.1465)	0.5511

We observe that despite a lack of sparsity in the data generating process, all VARX-L methods substantially outperform the information criterion benchmarks as well as BGR's Bayesian VAR. The Lag Group VARX-L achieves the best performance, which is to be expected. Since the Lag Group VARX-L has fewer groups than the Own/Other structures and does not impose within-group sparsity, it has more of a tendency to employ ridge-like penalization as opposed to setting an entire group to zero. Within the class of VARX-L models, the Basic VARX-L performs the worst. This suggests that structured groupings are more robust in applications where the true model is not sparse.

Scenario 6: Non-diagonal Covariance

The previous five simulation scenarios impose a diagonal structure on Σ_u , the contemporaneous covariance matrix. Such a scenario may rarely occur in practice. In order to examine the robustness of the VARX-L procedures in the presence of a non scaled-identity covariance matrix, this scenario pairs the sparsity pattern from Scenario 1 with the covariance structure depicted in Figure 2.

Note that unlike the least squares VARX with lags selected by AIC or BIC, as well as BGR's Bayesian VAR, the VARX-L procedures do not explicitly incorporate Σ_u . Hence, one should expect the benchmark procedures to be at a slight advantage in this scenario.

We observe that in this scenario that despite the more complex error structure, the Basic VARX-L, which conforms to the true sparsity pattern achieves the best performance, following by the Sparse Own/Other Group VARX-L. As in every other scenario, all VARX-L models substantially outperform the benchmarks. BGR's Bayesian VAR performs substantially worse than both information criterion based methods, even though it explicitly incorporates the covariance of $\mathbf{u}_t$; its poor performance is likely the result of incorporating an unreliable covariance estimate. For an expanded discussion and simulation study involving non-diagonal covariance structures as well as a procedure to incorporate covariance in fitting VARX-L models, consult Nicholson et al. [2016].

Figure 2: Covariance Matrix: Simulation Scenario 6

References

Per Christian Hansen. *Rank-deficient and discrete ill-posed problems: numerical aspects of linear inversion*, volume 4. Siam, 1998.

Peter R Hansen, Asger Lunde, and James M Nason. The model confidence set. *Econometrica*, 79(2):453–497, 2011.

Greg Ip. Non-borrowed reserves: False alarm. *The Wall Street Journal*, February 2008. URL <http://blogs.wsj.com/economics/2008/02/08/non-borrowed-reserves-false-alarm/>.

William Nicholson, David Matteson, and Jacob Bien. Bigvar: Tools for modeling sparse high-dimensional multivariate time series. 2016. URL <http://www.wbnicholson.com/BigVAR.pdf/>.

Song Song and Peter Bickel. Large vector auto regressions. *arXiv preprint arXiv:1106.3915*, 2011.

James H Stock and Mark W Watson. An empirical comparison of methods for forecasting using many predictors. *Manuscript, Princeton University*, 2005.