Variable Targeting and Reduction in High-Dimensional Vector Autoregressions

Tucker McElroy (U.S. Census Bureau)

Frontiers in Forecasting
February 21-23, 2018

Disclaimer

This presentation is released to inform interested parties of research and to encourage discussion. The views expressed on statistical issues are those of the authors and not necessarily those of the U.S. Census Bureau.

Collaborator: Thomas Trimbur (Census)

Outline

- Motivation for VAR Variable Selection
- Methodology
- An Application: QWI

Lecture slides are interposed with R sessions.

[GO TO TEACH0]

Motivation: considerations

- **Framework**: many time series variables; which ones should I use to help forecast the important (core) variables?
- **Variable Selection**: for multivariate regression, or Vector Autoregression (VAR), zeroing parameters does not eliminate variables. How to select variables (i.e., reduce dimension)? Sample size T may be less than number of variables N.
- **Sparsity**: if only a few variables are useful for forecasting the core variables, many parameters should be zero. How to enforce parameter sparsity? (Too much gives underfit, too little gives overfit.)

Motivation: desiderata

1. **Dimension Reduction**: we want to throw away irrelevant variables, i.e., variables that don't help forecasting.
2. **Variable Targeting**: suppose there are core variables that we wish to target, for forecasting applications. These should impact our model fitting and variable selection criterion.
3. **Variable Preservation**: we want forecasts for the core variables, not some linear combination of core variables.

Motivation: example

Data Background: our core variables are GDP and UR. Quarterly Workforce Indicators (QWI) measure employment, hires, separations, job construction, job destruction, and earnings, across 19 different private industry sectors.

Task: forecast GDP and UR (annual rate), utilizing 114 auxiliary labor variables.

[GO TO TEACH1]

Methodology: proposal

- **How to focus on core?** Use a modified Yule-Walker (YW) estimator that only involves the forecast performance of core series (so forecast performance of auxiliaries is ignored).
- **Which auxiliaries?** Eliminate any auxiliary that does not *Granger-cause* the core variables, i.e., if it does not help forecasting, chuck it.
- **How to get sparsity?** Replace with zeros any estimated parameters with low t-statistics, such that the likelihood ratio test is not significant.

Methodology: alternatives

LASSO/SCAD: a useful way to get sparsity; could be used in conjunction with initial variable selection above. Involves nonlinear optimization.

Dynamic Factor Analysis or **Random Projections**: dimension reduction is achieved by transforming variables, which hinders interpretability. How to do variable targeting and variable preservation?

Methodology: VAR background

A VAR(p) for N-dimensional $\{y_t\}$ satisfies

$$\Phi(L)\, y_t = \epsilon_t, \qquad \epsilon_t \sim WN(0, \underline{\sigma})$$

with L the lag operator and $\Phi(z) = I_N - \sum_{j=1}^{p} \Phi^{(j)}\, z^j$, where I_N is an $N \times N$ identity matrix, and each $\Phi^{(j)}$ denotes a coefficient matrix with real-valued entries.

Methodology: VAR background

If $p = 1$ and $y_t = [x_t', z_t']'$ is partitioned into core and auxiliary variables, then

$$\begin{bmatrix} x_t \\ z_t \end{bmatrix} = \begin{bmatrix} \Phi_{xx} & \Phi_{xz} \\ \Phi_{zx} & \Phi_{zz} \end{bmatrix} \begin{bmatrix} x_{t-1} \\ z_{t-1} \end{bmatrix} + \epsilon_t.$$

If $\Phi_{xz} \equiv 0$ then $\{z_t\}$ *does not Granger-cause* $\{x_t\}$.

Methodology: VAR background

Let $\widehat{y}_{t+1}$ be the one-step ahead linear forecast based on data $\{y_t, y_{t-1}, \ldots\}$; then

$$\widehat{y}_{t+1} = \sum_{j=1}^{p} \Phi^{(j)} \, y_{t+1-j}.$$

The forecast error covariance matrix is

$$\mathrm{MSE}_{t+1|t} = \mathbb{E}[(\widehat{y}_{t+1} - y_{t+1})\,(\widehat{y}_{t+1} - y_{t+1})'].$$

The usual fitting criterion is to minimize $\det \mathrm{MSE}_{t+1|t}$.

Methodology: objective function

Proposition

Consider the VAR(p) model for $\{y_t\}$ consisting of core $\{x_t\}$ and auxilliary $\{z_t\}$ variables. Suppose that $\{z_t\}$ does not Granger-cause $\{x_t\}$, and suppose that we fit the model so as to minimize the determinant of the forecast error covariance matrix of the core variables. Then the parameter estimates are given by the solution to the YW equations arising from $\{x_t\}$ alone.

Methodology: variable selection

Using Proposition 1: sequentially add one auxiliary variable at a step to the core variables, fit this augmented model, and test (with Wald statistic) whether this has any improvement over prior step, i.e., test whether the new auxiliary Granger-causes the core or previous auxiliary variables.

Procedurally: first determine best candidate auxiliary variables, by fitting each alone together with the core variables, and obtaining p-values for Wald statistics of Granger causality. (So low p-values indicate strongest rejection of non-causality, i.e., these variables have the most predictive impact on the core.)

Methodology: example

Variant: we may get better results, by doing variable selection for *each* core variable, and take union of all resulting auxiliaries at the end.

[GO TO TEACH2]

Methodology: sparsity

Fitting a Constrained VAR: now we want to replace small parameter estimates in each $\Phi^{(j)}$ matrix with a zero and refit. If we know *where* we want to place zeros, this amounts to a constrained YW estimator. (You can do OLS, or LASSO/SCAD as well...)

Who Gets a Zero?: we can compute t-statistics for YW parameter estimates quite easily. We propose to sort these by absolute t-statistic, starting with lowest (i.e., having the least evidence to reject a zero value); sequentially replace these entries with zeroes using constrained YW, at each step testing whether likelihood significantly differs from that of the unconstrained model.

Methodology: sparse details

Zero constraints: for r constaints, formulate via a $N^2p \times r$ dimensional selection matrix J, such that

$$\text{vec}[\Phi^{(1)} \dots \Phi^{(p)}] = J\,\psi \tag{1}$$

for an r-vector ψ. If $\Gamma(h) = \text{Cov}[y_{t+h}, y_t]$ is the autocovariance function, let $\underline{R}_{p+1}$ be block $(p+1)N \times (p+1)N$ dimensional with block entry jkth block entry $\Gamma(k-j)$ for $1 \leq j, k \leq p+1$. Partition this matrix as

$$\underline{R}_{p+1} = \begin{bmatrix} \Gamma(0) & \underline{R}_{\cdot} \\ \underline{R}'_{\cdot} & \underline{R}_{p} \end{bmatrix}.$$

Methodology: sparse details

The constrained YW is found by iteratively solving

$$\psi = \left(J' \left[\underline{R}_p \otimes \underline{\sigma}^{-1}\right] J\right)^{-1} J' \left[\underline{R}'_{\cdot} \otimes \underline{\sigma}^{-1}\right] \text{vec}(I_N)$$
$$\underline{\sigma} = \Gamma(0) - [\Phi^{(1)} \dots \Phi^{(p)}] \underline{R}_p [\Phi^{(1)} \dots \Phi^{(p)}]'.$$

In the second step, we use (1) to get the constrained VAR coefficients from ψ. The innovation covariance $\underline{\sigma}$ is also $\text{MSE}_{t+1|t}$, so (once converged) we can take its determinant to get the value of the objective function on the constrained model.

[GO TO TEACH3]

Methodology: model checking

Residuals: we can check the core residuals for non-normality (Shapiro-Wilks) or serial correlation (Portmanteau).

Displays: can examine sparse estimated VAR coefficients, and uncertainties. (Asymptotic variance for $\text{vec}[\Phi^{(1)} \dots \Phi^{(p)}]$ is $T^{-1}\underline{R}_p^{-1} \otimes \underline{\sigma}$. LASSO handles post-model selection uncertainty.)

[GO TO TEACH4]

Methodology: S-VAR

Structural VAR: another way of writing the process, where the innovations have covariance matrix I_N, but now there is a contemporaneous effect:

$$A^{(0)}\, y_t = \sum_{j=1}^{p} A^{(j)}\, y_{t-j} + \eta_t,$$

with $A^{(j)} = A^{(0)}\, \Phi^{(j)}$ for $1 \leq j \leq p$, $A^{(0)} = U^{-1}$, where U is upper triangular such that $U\, U' = \underline{\sigma}$, and $U\, \eta_t = \epsilon_t$.

Drawback: although it can help understand the impact of a shock, now the variables are contemporaneously related through $A^{(0)}$.

Methodology: impulse response

Impulse Response: write $A(z) = A^{(0)} - \sum_{j=1}^{p} A^{(j)} z^j$, and set $\Psi(z) = A(z)^{-1}$, yielding the VMA(∞) representation

$$y_t = \Psi(L)\, \eta_t.$$

Here $\Psi(z) = \sum_{\ell=0}^{\infty} \Psi^{(\ell)} z^\ell$ as a matrix power series. For any $1 \leq j, k \leq N$, we can plot $\Psi_{jk}^{(\ell)}$ as a function of $\ell \geq 0$, called the *impulse response plot*.

[GO TO TEACH5]

Methodology: evaluating performance

Forecast Performance: we choose a sample period, so that we can examine both in-sample and out-of-sample forecast performance (for each core variable). We can compare to a benchmark model, given by the best univariate AR model (order selected by AIC) for each core series.

Diebold-Marriano: to test whether competing forecasts outperform, we can cumulate out-of-sample squared forecast errors for both models, take the difference, and normalize by a HAC estimator (this is called the Diebold-Marriano test).

[GO TO TEACH6]

Conclusion

What did I learn?: a bit of VAR modeling, machine learning, and R! The *varhi* suite allows variable selection, fitting sparse VAR, and evaluating forecast performance. Everything is analytical (except some iterations in sparse VAR fitting) or recursive, making the computations fast – avoiding nasty likelihood surfaces that can occur with LASSO.

Contact: tucker.s.mcelroy@census.gov