Ridge VARX

Model Estimation

F. Stroes

January 20, 2022

1 Introduction

In the following section the theory underlying the Ridge VARX package is outlined. Notation is similar to that in Lütkepohl (2013) which was referenced to obtain the original estimators for the VAR model. Given these estimators, contemporaneous exogenous variables and a ridge penalty on the estimation criterion can easily be added. I present the modified estimators in the next sections. It is important to note that in the book of Lütkepohl (2013), The columns of the regressor matrix contain observations of regressors that are in the rows. This is a transpose of the notation that is more commonly used.

1.1 Stationarity

The theory in the following sections only applies if the sample consists of weak sense stationary data. Although weak sense stationarity is not easily tested for itself, the presence of a unit root which is a common cause of non-stationarity can easily be tested for. If the data is non-stationary, differencing should be applied before fitting the model.

1.2 VARX introduction

In real world settings the explicit goal in time series analysis is often to forecast a single time series using explanatory variables. The need to forecast multiple time series at once can likely occur in the same exercise, for instance when one or more of the explanatory variables need to be forecasted themselves.

The VAR formulation of the problem produces forecasts of several time series simultaneously. This is achieved by putting all variables of interest in a vector. This vector is then forecasted, using past values of the vector. Hence, *'Vector Auto Regressive'*.

Formally we predict a $(k \times 1)$ column vector containing the values of k endogenous time series at time t.

$$y_t = Bz_t + u_t$$

Where B is a time-invarying matrix of coefficients and z_t is a vector containing the values of each regressor at time t.

The sequence $\{u_t\}$ is a sequence of white noise innovations. The normality of $\{u_t\}$ is implicitly assumed when one chooses to minimize the sum of squared residuals. And the assumption of the white noise property of $\{u_t\}$ depends on the assumption of correct model specification. The latter is more problematic in the case of penalized estimation. This is discussed in the section on p-values.

In a VAR model, only lagged values of $\{y_t\}$ are contained in $\{z_t\}$, The remaining X thus stands for *'exogenous'*. In this context, entries of z_t that are predicted in the vector of the VARX model are defined as endogenous. And variables that are not influenced by (historical) values of the vector are defined as exogenous.

The model can alternatively be written as a single equation describing the entire sample, by adding columns for each value of $t \in \{1, 2, ..., T\}$.

$$Y = BZ + U$$

The parameter matrixs B is unknonw and needs to be estimated, this is esily accomplished using the multivariate least squares (MLS) solution.

$$\hat{B} = YZ^T(ZZ^T)^{-1}$$

1.3 Ridge introduction

In short samples (if T is small relative to K, where K is the total number of parameters in B) over-fitting can be a problem. In these cases a penalized or *'regularized'* estimator will in general perform better out of sample, provided that the regularization parameter is properly tuned. Ridge can be shown to decrease the parameter estimation variance, at the cost of producing a small bias. I added a Ridge penalty to the estimator found in Lütkepohl (2013) to obtain:

$$\hat{B} = YZ^T(ZZ^T + \lambda I)^{-1}$$

Besides the potentially improved out of sample performance. the Least Squares criterion with a Ridge penalty is well posed (has an analytical solution) even in settings where there are more explanatory variables than observations in the sample.

1.4 Estimation

In their original form both the original MLS and Ridge estimator require a matrix inversion. Matrix inversions are generally avoided in numerical optimization, due to concerns regarding machine precision for large matrices. Luckily, the estimation problem can easily be reformulated in a way that does not require this troublesome matrix inversion, Doing so, I rewrote the problem as:

$$\underbrace{I_k \otimes (ZZ^T + \lambda I)}_{\text{A}} \underbrace{\text{vec}(B^T)}_{\text{x}} = \underbrace{\text{vec}(ZY^T)}_{\text{b}}$$

To have control over the numerical optimization the scipy.sparse.linalg.cg solver is used. This solver requires x and b to be vectors, in solving $Ax = b$.

For the positive definite and symmetric matrix A, the conjugate gradient method provides an efficient solution to the problem with near double precision accuracy for the parameter estimates.

1.5 p-Values of the package

Note

The package is primarily intended for forecasting, not for hypothesis testing. The p-values of the package are therefore based on the normality of the parameter estimates and the assumption of correct dynamic model specification. The first of these assumptions, normality, is unrealistic in short samples and the latter (correct model specification) is violated when penalized estimation is used and a bias is introduced in the auto regressive parameters. The p-values are therefore best used only with long samples. If more accurate estimates of the p-values are needed, the parameter estimation variance can be used to make adjustments for incorrect dynamic model specification and low degrees of freedom independently of the package.

To get p-values for the estimated parameters, An estimate of the variance of $\hat{B}(\lambda)$ is needed. To find this estimator, I applied the well known method for finding the variance of the Ridge estimator, starting from the original Least Squares estimator to the estimator as it is defined in Lütkepohl (2013). Using this method, the estimator is found by taking the known MLS estimator for the model parameters:

$$\hat{B} = YZ^T(ZZ^T)^{-1}$$

which has variance equal to $(ZZ^T)^{-1} \otimes \Sigma_U$, given a sample of T longitudinal observations.

And noting that the Ridge estimator $\hat{B}(\lambda)$ is equal to $\hat{B}W_\lambda^T$ when

$$W_\lambda = (ZZ^T + \lambda I)^{-1} ZZ^T$$

And therefore the variance can easily be seen to be:

$$\text{var}(B(\lambda)) = W_\lambda (ZZ^T)^{-1} W_\lambda^T \otimes \Sigma_U$$

Finally Σ_U can be replaced by its consistent in sample estimator. In the MLS case Σ_U will be a diagonal matrix under the assumption of correct dynamic model specification. And an unbiased in sample estimator can be found to be:

$$\frac{T}{T - (1 + pk + m)} UU^T$$

The degrees of freedom in the LS fitted model are trivially seen to be equal to the amount of fitted parameters per equation $(1 + pk + m)$. In the ridge case however, the variability of the predictions decreases as the amount of regularization increases. An estimator for the degrees of freedom in the ridge case $\widehat{\text{df}}(\lambda)$ needs to be used.

If the number of parameters per equation is bigger than T the Woodburry identity provides us with an equivalent formulation of $(ZZ^T + \lambda I)^{-1}$ that requires a smaller matrix inversion and is therefore preferred.

1.6 Degrees of freedom Ridge

In for instance Dijkstra (2014) it is shown how to get the degrees of freedom of a model with Ridge regularization. The degrees of freedom are obtained by first finding the matrix $H(\lambda)$ for which: $ZB = M(\lambda)Y$ and then taking the trace of $H(\lambda)$.

For the ridge estimator of the VARX model I found:

$$H(\lambda) = Z^T (ZZ^T + \lambda I)^{-1} Z$$

With this matrix one can compute:

$$\widehat{\text{df}}(\lambda) = \text{tr}(H(\lambda))$$

Note that $\widehat{\text{df}}(\lambda) = \text{tr}(H(\lambda))$ which is not affected by the columns in Y. Therefore, same result is obtained regardless of the equation that we are considering. And $\widehat{\text{df}}(\lambda)$ can be used as a scalar value in:

$$\widehat{\Sigma}_U(\lambda) = \frac{T}{T - \widehat{\text{df}}(\lambda)} UU^T$$

1.7 Ridge parameter tuning

Finding a good value for the regularization parameterλ is not trivial. What makes a good value depends on the sample and can only be determined after setting a, to some extent, arbitrary standard for when our model is 'good'. One way of doing this is by looking for the model that minimizes Akaike's information criterion (AIC).

For the likelihood of the entire sample I assumed a multivariate normal distribution to be appropriate in accordance with the previous assumption of normality of the innovations.

AIC can then be minimized for the $h = 1$ period ahead forecast by minimizing:

$$\ln(|\tilde{\Sigma}_U(\lambda)|) + \frac{2(\widehat{\mathrm{df}}(\lambda))K}{T}$$

Where $\tilde{\Sigma}_U(\lambda)$ is the unadjusted ML estimator of the in sample error variance for any given value of λ.

References

Dijkstra, T. K. (2014). Ridge regression and its degrees of freedom. *Quality & Quantity*, *48*(6), 3185–3193.

Lütkepohl, H. (2013). *Introduction to multiple time series analysis*. Springer Science & Business Media.