



Since the seminal paper of Sims (1980) vector autoregressive models have become a key instrument in macroeconomic research. This post presents the basic concept of VAR analysis and guides through the estimation procedure of a simple model. When I started my undergraduate program in economics I occasionally encountered the abbreviation *VAR* in some macro papers. I was fascinated by those waves in the boxes titled *impulse responses* and wondered how difficult it would be to do such research on my own. I was motivated, but my first attempts were admittedly embarrassing. It took me quite a long time to figure out which kind of data can be analysed, how to estimate a VAR model and how to obtain meaningful impulse responses. Today, I think that there is nothing fancy about VAR models at all once you keep in mind some points.

Univariate autoregression

VAR stands for *vector autoregression*. To understand what this means, let us first look at a simple univariate (i.e. only one dependent or endogenous variable) autoregressive (AR) model of the form $y_t = a_1 y_{t-1} + e_t$. In this model the current value of variable y depends on its own first lag, where a_1 denotes its parameter coefficient and the subscript refers to its lag. Since the model contains only one lagged value the model is called autoregressive of order one, short AR(1), but you can easily increase the order to p by adding more lags, which results in an AR(p). The error term e_t is assumed to be normally distributed with mean zero and variance σ^2 .

Stationarity

Before you estimate such a model you should always check if the time series you analyse are stationary, i.e. its means and variances are constant over time and do not show any trending behaviour. This is a very important issue and every good textbook on time series analysis treats it quite – maybe too – intensively. This has some reasons. A central point is that if you estimate models with non-stationary data, you will get improper test statistics that might lead you to choose the wrong model which is clearly undesirable.

There is a series of statistical tests like the Dickey-Fuller, KPSS or Phillips-Perron test to check whether a series is stationary. However, a common practise of practitioners is to plot a series and look whether it moves around a constant mean value, i.e. a horizontal line. If this is the case, it is likely to be stationary. However, both statistical and visual tests have their drawbacks and you should always be careful with those approaches. Additionally, you might want to check what the economic literature has to say about the stationarity of particular time series like ,e.g., GDP, interest rates or inflation. This approach is particularly useful if you want to determine whether a series is trend or difference stationary. (<https://de.mathworks.com/help/econ/trend-stationary-vs-difference-stationary.html>).

Autoregressive distributed lag models

Regressing a macroeconomic variable solely on its own lags like in an AR(p) model might be a quite restrictive approach. Usually, it is more appropriate to assume that there are further factors that drive a process. This idea is captured by models which contain lagged values of the dependent variable as well as contemporaneous and lagged values of other, i.e. exogenous, variables. Again, these exogenous variables should be stationary. For an endogenous variable y_t and an exogenous variable x_t such an *autoregressive distributed lag*, or ADL, model can be written as

$$y_t = a_1 y_{t-1} + b_0 x_t + b_1 x_{t-1} + e_t.$$

In this ADL(1,1) model a_1 and e_t are defined as above and b_0 and b_1 are the coefficients of the contemporaneous and lagged value of the exogenous variable, respectively.

The forecasting performance of such an ADL model is likely to be better than for a simple AR model. However, what if the exogenous variable depends on lagged values of the endogenous variable too? This would mean that x_t is endogenous too and there is further space to improve our forecasts.

Vector autoregressive models

At this point the VAR approach comes in. A simple VAR model can be written as

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix} \begin{pmatrix} y_{1t-1} \\ y_{2t-1} \end{pmatrix} + \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{pmatrix}$$

or, more compactly,

$$y_t = A_1 y_{t-1} + \epsilon_t,$$

where $y_t = \begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix}$, $A_1 = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$ and $\epsilon_t = \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{pmatrix}$.

Note: Yes, you should familiarise yourself with some (basic) matrix algebra (addition, subtraction, multiplication, transposing, transposition, inversion and the determinant), if you want to work with VARs.

Basically, such a model implies that *everything depends on everything*. But as can be seen from this formulation, each row can be written as a separate equation, so that $y_{1t} = a_{11}y_{1t-1} + a_{12}y_{2t-1} + \epsilon_{1t}$ and $y_{2t} = a_{21}y_{1t-1} + a_{22}y_{2t-1} + \epsilon_{2t}$. Hence, the VAR model can be rewritten as a series of individual ADL models as described above. In fact, it is possible to estimate VAR models by estimating each equation separately.

Looking a bit closer at the single equations you will notice, that there appear no contemporaneous values on the right-hand side (*rhs*) of the VAR model. However, information about contemporaneous relations can be found in the so-called *variance-covariance matrix* Σ . It contains the variances of the

endogenous variable on its diagonal elements and covariances of the errors on the off-diagonal elements. Latter contain information about contemporaneous effects. Like the error variance σ^2 in a single equation model it is essential for the calculation of test statistics and confidence intervals.

The covariance matrix (usually) is *symmetric*, i.e. the elements to the upper right of the diagonal (the 'upper triangular') mirror the elements to the lower left of the diagonal (the 'lower triangular'). This reflects the idea that the relations between the endogenous variables only reflect correlations and do not allow to make statements about causality, since the effects are the same in each direction. This is the reason why this model is said to be not uniquely *identified*.

Contemporaneous causality or, more precisely, the structural relationships between the variables is analysed in the context of so-called *structural* VAR (SVAR) models which impose special restrictions on the covariance matrix – and depending on the model on other matrices as well – so that the system is identified, i.e. there is only one unique solution for the model and it is clear, how the causalities work. The drawback of this approach is that it depends on the more or less subjective assumptions made by the researcher. For many researchers this is too much subjective information, even if sound economic theory is used to justify those assumptions.

In this article I consider a VAR(2) process of the form

$$\begin{pmatrix} y_{1t} \\ y_{2t} \end{pmatrix} = \begin{bmatrix} -0.3 & -0.4 \\ 0.6 & 0.5 \end{bmatrix} \begin{pmatrix} y_{1t-1} \\ y_{2t-1} \end{pmatrix} + \begin{bmatrix} -0.1 & 0.1 \\ -0.2 & 0.05 \end{bmatrix} \begin{pmatrix} y_{1t-2} \\ y_{2t-2} \end{pmatrix} + \begin{pmatrix} \epsilon_{1t} \\ \epsilon_{2t} \end{pmatrix}$$

and I generate the data in R with:

```
1 set.seed(123) # Reset random number generator for reasons of reproducibility
2 # Generate sample
3 t <- 200 # Number of time series observations
4 k <- 2 # Number of endogenous variables
5 p <- 2 # Number of lags
6 A.1 <- matrix(c(-.3,.6,-.4,.5),k) # Coefficient matrix of lag 1
7 A.2 <- matrix(c(-.1,-.2,.1,.05),k) # Coefficient matrix of lag 2
8 A <- cbind(A.1,A.2) # Companion form of the coefficient matrices
9
10 series <- matrix(0,k,t+2*p) # Raw series with zeros
11 for (i in (p+1):(t+2*p)){ # Generate series with e ~ N(0,0.5)
12   series[,i] <- A.1%*series[,i-1] + A.2%*series[,i-2] + rnorm(k,0,.5)
13 }
14
15 series <- ts(t(series[,-(1:p)])) # Convert to time series format
16 names <- c("V1","V2") # Rename variables
17 plot.ts(series) # Plot the series
```

Estimation

The estimation of the parameters and the covariance matrix of a simple VAR model is straightforward. For $Y = (y_1, \dots, y_T)$ and $Z = (z_1, \dots, z_T)$ with z as a vector of lagged values of y and possible deterministic terms the least squares estimator of the parameters is $\hat{A} = YZ(Z'Z)^{-1}$. The covariance matrix is then obtained from $\frac{1}{T-Q}(Y - \hat{A}Z)(Y - \hat{A}Z)'$, where Q is the number of estimated parameters. For basic applications these formulas are usually already programmed in standard statistics packages.

In order to estimate the VAR model I use the `vars` package by Pfaff (2008). The relevant function is `VAR` and its use is straightforward. You just have to load the package and specify the data (`y`), order (`p`) and the `type` of the model. The option `type` determines whether to include an intercept term, a trend or both in the model. Since the artificial data I generated does not contain intercepts, I choose to neglect it in the estimation by setting the `type` option to `"none"`.

```
1 | library(vars) # Load package
2 | var.1 <- VAR(series,2,type="none") # Estimate the model
```

Model comparison

A central issue in VAR analysis is to find the number of lags which yields the best results. Model comparison is usually based on information criteria like the AIC, BIC or HQ. Usually, the AIC is preferred over other criteria, due to its favourable small sample forecasting features. The BIC and HQ, however, work well in large samples and have the advantage of being a consistent estimator of the *true* order, i.e. they prefer the true order of the VAR model – in contrast to the order which yields the best forecasts – as the sample size grows.

The `VAR` function of the `vars` package already allows to calculate standard information criteria to find the best model. In my example I use the AIC:

```
1 | var.aic <- VAR(series,type="none",lag.max=5,ic="AIC")
```

Note that instead of specifying the order `p`, we now set the maximum lag length of the model and the information criterion used to decide which order should be used. The function then estimates all five models, compares them according to their AIC values and automatically selects the most favourable. Looking at `summary(var.aic)` we see that the AIC suggests to use an order of 2 which is the true order.

Looking at the results a bit more closely we can compare the parameter estimate of the model with the known true values:

```
1 | A # True values
2 | round(rbind(coef(var.aic)[[1]][,1],coef(var.aic)[[2]][,1]),2) # Rounded estimat
```

All the estimates have the right sign and are relatively close to their true values. I leave it to you to look at the standard deviations of `summary(var.aic)` to check whether the true values fall into the confidence bands of the estimates.

Impulse responses

Once we have decided for a final VAR model its estimated parameter values have to be interpreted. Since in such a model all variables depend on each other, individual parameter values only provide limited information. In order to get a better intuition of the model's dynamic behaviour, *impulse responses* are used. They give the reaction of a response variable to a one-time shock in an impulse variable. The trajectory of the response variable can be plotted which results in those wavy curves seen in many macro papers.

In R I use the `irf` function to obtain an impulse response function of Series 2 after a shock to Series 1. After specifying the model and the variables for which I want the impulse response I set the time horizon `n.ahead` to 20. The plot gives the response of series 2 for the periods 0 to 20 to a shock in series 1

in period 0. The function also automatically calculates so-called bootstrap confidence bands. (Bootstrapping is a common procedure in impulse response analysis. But you ought keep in mind that it has its drawbacks when you work with structural VAR models.)

```
1 | ir.1 <- irf(var.1,impulse="Series.1",response="Series.2",n.ahead = 20,ortho = F
2 | plot(ir.1)
```

The `ortho` option is important, because it says something about the contemporaneous relationships between the variables. In our example we already know that such relationships do not exist, because the true variance-covariance matrix – or simply covariance matrix – is diagonal with zeros in the off-diagonal elements. However, since the limited time series data with 200 observations restricts the precision of the parameter estimates, the covariance matrix has positive values in its off-diagonal elements which implies non-zero contemporaneous effects of a shock. To rule this out I set `ortho=FALSE`. The result of this is that the impulse response starts at zero in period 0. You could also try out the alternative and set `ortho=TRUE` which results in a plot that start below zero. I do not want to go into more detail here, but suffice it so say that the issue of so-called orthogonal errors is one of the central problems in VAR analysis and you should definitely read more about it, if you intend to set up your own VAR models.

Sometimes, it is interesting to see what the long-run effects of a shock are. To get an idea about that you can also calculate and plot the *cumulative* impulse response function to get an idea of the overall long-run effect of the shock:

```
1 | ir.2 <- irf(var.1,impulse="Series.1",response="Series.2",n.ahead = 20,ortho = F
2 | cumulative = TRUE)
3 | plot(ir.2)
```

We see that although the reaction of series 2 to a shock in series 1 is negative during some periods, the overall effect significantly positive.

References

Luetkepohl, H. (2007). *New Introduction to Multiple Time Series Analysis*. Berlin: Springer.

Bernhard Pfaff (2008). VAR, SVAR and SVEC Models: Implementation Within R Package *vars*. *Journal of Statistical Software* 27(4).

Sims, C. (1980). Macroeconomics and Reality. *Econometrica*, 48(1), 1-48.

ADVERTISEMENT

Advertisements




HORTONWORKS™

Hortonworks was
Ranked as a Leader in
the Forrester Wave™:
Big Data Warehouse

[LEARN MORE](#)

[Report this ad](#)



HORTONWORKS™

Hortonworks was
Ranked as a Leader in
the Forrester Wave™:
Big Data Warehouse

[LEARN MORE](#)

[Report this ad](#)