VAR modeling, regularization and the SparseTSCGM package

Stijn de Vos

03 - 11 - 2017

We start with a small review of vector-autoregressive models. Afterwards, regularization is discussed and the SparseTSCGM package is demonstrated.

Introduction

Suppose we have four items called 'Positive Affect' (x_1), 'Anhedonia' (x_2), 'Irritability' (x_3) and 'Weightloss' (x_4). Measuring these items multiple times gives a time-series dataset where each observation consists of four elements. Written mathematically, the observation at time t consists of four numbers $(x_{1t}, x_{2t}, x_{3t}, x_{4t})$. As a shorthand notation, people often write $X_t = (x_{1t}, x_{2t}, x_{3t}, x_{4t})$. X_t is called a (four-dimensional) *vector*.

This is why it's called **vector**-autoregressive (VAR) modeling; we regress a vector X_t on past 'versions' of itself, instead of a single variable. The simplest VAR model only regresses X_t on its previous measurement X_{t-1}. This VAR model is said to have a **lag** of 1. In formula form, this model looks as follows:

$$
\begin{aligned}
x_{1t} &= a_{11}x_{1(t-1)} + a_{12}x_{2(t-1)} + a_{13}x_{3(t-1)} + a_{14}x_{4(t-1)} + \epsilon_{1t} \\
x_{2t} &= a_{21}x_{1(t-1)} + a_{22}x_{2(t-1)} + a_{23}x_{3(t-1)} + a_{24}x_{4(t-1)} + \epsilon_{2t} \\
x_{3t} &= a_{31}x_{1(t-1)} + a_{32}x_{2(t-1)} + a_{33}x_{3(t-1)} + a_{34}x_{4(t-1)} + \epsilon_{3t} \\
x_{4t} &= a_{41}x_{1(t-1)} + a_{42}x_{2(t-1)} + a_{43}x_{3(t-1)} + a_{44}x_{4(t-1)} + \epsilon_{4t}
\end{aligned}
$$

The numbers a_{ij} are the regression coefficients of the VAR model. The terms ϵ_{it} represent the (contemporaneous) noise at time t. VAR models usually assume that ϵ_{it} is normally distributed with mean 0 and some standard deviation: $\epsilon_{it} \sim \mathcal{N}(0, \sigma_i)$.

This all looks very nasty and complicated. Luckily, using vector notation, we can abbreviate the model description quite a bit. Using the vector notation we can write

$$
\begin{bmatrix} x_{1t} \\ x_{2t} \\ x_{3t} \\ x_{4t} \end{bmatrix} = \begin{bmatrix} a_{11} & a_{12} & a_{13} & a_{14} \\ a_{21} & a_{22} & a_{23} & a_{24} \\ a_{31} & a_{32} & a_{33} & a_{34} \\ a_{41} & a_{42} & a_{43} & a_{44} \end{bmatrix} \begin{bmatrix} x_{1(t-1)} \\ x_{2(t-1)} \\ x_{3(t-1)} \\ x_{4(t-1)} \end{bmatrix} + \begin{bmatrix} \epsilon_{1t} \\ \epsilon_{2t} \\ \epsilon_{3t} \\ \epsilon_{4t} \end{bmatrix}
$$

Or, even more convenient:

$$
X_t = AX_{t-1} + \epsilon_t
$$

You will have to take my word for it that this represents the same system of equations as written earlier. If you don't, or feel particularly bored, you can try looking up the Wikipedia page on matrix operations and confirm it for yourself!

The vector ϵ_t is comprised of four random variables, each with a normal distribution. Another way of saying this is that epsilon is distributed according to a *multivariate* normal distribution with mean vector $(0, 0, 0, 0)$ and covariance matrix

$$
\Sigma = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} & \sigma_{14} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} & \sigma_{24} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} & \sigma_{34} \\ \sigma_{41} & \sigma_{42} & \sigma_{43} & \sigma_{44} \end{bmatrix}
$$

Note that when we include more lags in the model, we have multiple regression coefficient matrices. For the sake of easy exposition, we stick to the lag-1 model for now. The most important important sets of parameters of this model are

- A, the matrix of regression coefficients, representing the associations between specific items over time;
- Σ, the covariance matrix representing contemporaneous noise at each measurement.

When we fit a VAR model, we are trying to find values for the elements of these matrices that fit your dataset the best.

Regularization

Conclusion