Transformers for Time Series Forecasting

Marco Zanotti

University Milano-Bicocca

Contents

1. VARs

VAR Model

Vector autoregressive (**VAR**) models are popular choice for studying the joint dynamics of multiple time series.

Consider x_t as a p-dimensional vector of time-series at time t, a VAR model is just a multivariate normal regression of x_t on its own-lags

$$x_t = c + \sum_{i=1}^{k} \Pi_i x_{t-i} + \epsilon_t \qquad (1)$$

where Π_i are $p \times p$ coefficient matrices determining the dynamics of the system, c is a deterministic vector of p components and $\epsilon_t \sim N_p(0, \Sigma)$.

Example of VAR(1)

A first order VAR model ($p = 1$) in two variables would be given by

$$\begin{pmatrix} x_{1,t} \\ x_{2,t} \end{pmatrix} = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} + \begin{pmatrix} \pi_{11} & \pi_{12} \\ \pi_{21} & \pi_{22} \end{pmatrix} \begin{pmatrix} x_{1,t-1} \\ x_{2,t-1} \end{pmatrix} + \begin{pmatrix} \epsilon_{1,t} \\ \epsilon_{2,t} \end{pmatrix}$$

If, for example $\pi_{12} \neq 0$, this means that the history of x_2 helps explaining x_1.

Pros & Cons

- Main **advantage**: simplicity. VAR models require no special structure since the outcome variables are regressed on their own lagged variables.
- Main **disadvantage**: large number of regression coefficients, $K(Kp+m)$. The significant number of regression parameters is proportional to the number of lags, hence interpretation may be difficult and may not be good when fitted to small data.

Several tools have been proposed to aid in the interpretation of VAR models, most notably Impulse Response Functions (IRF) and Granger Causality tests (GC).

Granger Causality

The general idea behind **Granger causality** is that a variable X Granger causes Y if past values of X can help explain Y.

Of course, if Granger causality holds this does not guarantee that X causes Y. Nevertheless, if past values of X have explanatory power for current values of Y, it at least suggests that X improves Y's predictability.

In the context of a VAR model, testing for Granger causality simply implies to test the significance of the cross-coefficients (π_{12} and π_{21} in the VAR(1) example).

2. Graphical VARs

Graphical Model

Graphical models can be informally defined as statistical models represented in the form of a graph, where the **nodes** (vertices) represents the variables and the presence of an **edge** between a pair of vertices means that the variables are in some sense related.

Some advantages of a graphical model are:

- represents graphically the logical implication of the relationships
- suitable representation of the causal relationships using directed edges
- clarity of interpretation when analyzing complex interactions.

DAG Model

In cross-sectional data, the graph usually represents the conditional independence structure of the system and (by symmetry) the edges are undirected. In time-series data, the time dimension of a process makes it more feasible to consider directed flow.

A **Directed Acyclic Graph** (DAG) is a collection $G = \{V, E\}$, where V is the set of vertices and E is the set of edges.

Connecting DAG and VAR

Given the VAR model in (1), there is a one-to-one relationship between the Π_i matrices and DAGs, that is

$$x_{a,t-i} \to x_{b,t} \iff \Pi_i(a,b) \neq 0$$

Where $x_{a,t-i} \to x_{b,t}$ means that $x_{a,t-i}$ "causes" somehow $x_{b,t}$

Hence, **directed edges** represent the **Granger causality** relations.

Characterization of Graphical VARs

For a Gaussian VAR process x_t with a Granger causality graph $G(V, E)$, the following two conditions hold:

1. $(a, b) \notin E_1 \iff \Pi_i(a, b) = 0, \quad \forall i = 1, ..., k$
2. $(a, b) \notin E_2 \iff \Sigma^{-1}(a, b) = 0$

where E_1 and E_2 are the sets of directed and undirected edges, and $\Pi_i(a, b)$ is the (a, b)th element of Π_i.

A Gaussian VAR process satisfying these conditions is said to belong to a **graphical Granger causal VAR model**, $VAR(G, k)$.

3. Bayesian Graphical VARs

The $VAR(G, k)$ Problem

The unknown quantities of a $VAR(G, k)$ process are:

- the underlying Granger causality graph G
- the number of lags k
- the modelling parameters $\Pi_1, ..., \Pi_k$ and Σ (from now θ)

Inference on G and k is essentially a model determination problem in a very large space and the authors proposed a solution within a Bayesian framework.

Posterior of $VAR(G, k)$

Following a **Bayesian approach**, the authors proposed to model jointly Granger causality and lag length. The **joint posterior** distribution of (G, k) conditional on the observed time series X can be written as

$$\pi(G, k|X) = \frac{m(G, k|X)\ p(G, k)}{\sum_{G \in \mathcal{G}} \sum_{k=0}^{K} m(G, k|X)\ p(G, k)}$$

where $p(G, k)$ is the joint prior of G and k, and $m(G, k|X)$ is the marginal likelihood of the observed time series X.

Prior of G and k

The **joint prior** of G and k is over a discrete set and can be chosen in many ways. For instance, if there is no reason for favoring any particular graph in $\mathcal{G}$ a priori, the following prior can be used

$$p(G, k) = \frac{p(k)}{|\mathcal{G}|}$$

where $p(k)$ is some discrete distribution over the integers $k = 0, 1, ..., K$.

Marginal Likelihood of X

The **marginal likelihood** of the observed time series X is given by

$$m(G, k|X) = \int L(X|\theta, G, k)\ p(\theta|G, k)\ d\theta$$

where $L(X|\theta, G, k)$ is the likelihood function under model (G, k) with parameters θ, and $p(\theta|G, k)$ is the prior distribution of θ.

For the $VAR(G, k)$ family of models, a common choice for $p(\theta|G, k)$ is the inverse Wishart distribution, which, however, is an **improper prior** and it is not directly usable for deriving the joint posterior of G and k.

The Fractional Bayes Approach

A solution to this problem is to use a **Fractional Bayes approach**: first, a small part of the sample is sacrificed in updating the improper prior to a proper posterior, then this posterior is used as a new prior for the remaining observations.

Because of the **Fractional Marginal Likelihood** does not factorizes under a $VAR(G, k)$ model, the authors derived an **approximated** version that produced favorable results in model inference.

Finally, the estimate for θ is obtained iterating between two conditional estimators until convergence (given in Lutkepohl 1993 and Lauritzen 1996).

4. Conclusions

Conclusions

- DAG approach can be used to study the Granger causality relationships in VAR models
- Since the $VAR(G, k)$ usually has a very large number of parameters, a fractional Bayes approach can be used to approximate the joint posterior of the process
- Empirically valid results

Bibliografy

Corander, J. & Villani, M. (2006), 'A Bayesian Approach to Modelling Graphical Vector Autoregressions', Journal of Time Series Analysis 27(1), 141–156.

Thank you!