Frontiers in Forecasting, Minneapolis
February 21-23, 2018

Sparse VAR-Models

Christophe Croux

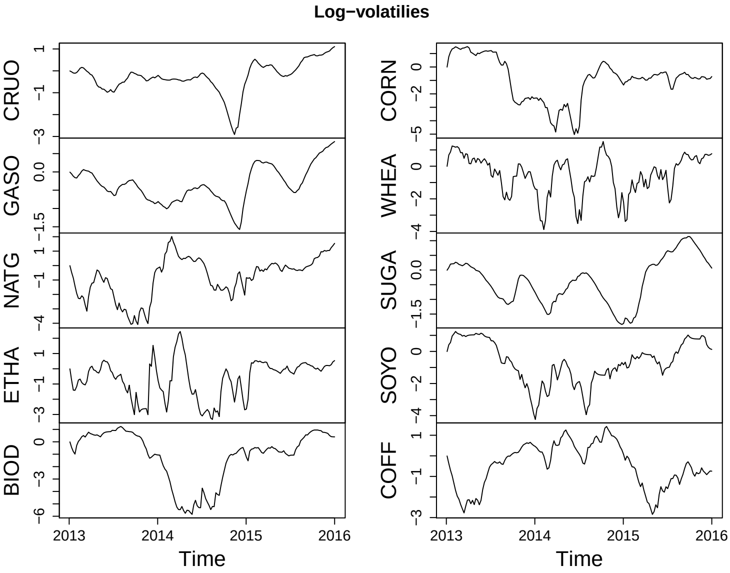
EDHEC Business School (France)

Joint Work with Ines Wilms (Cornell University), Luca Barbaglia (KU leuven), and Sarah Gelper (TU Eindhoven).

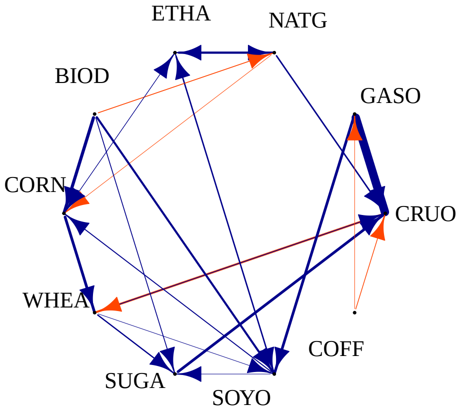
Section 1

VAR Models

10-dimensional time series



Network



Vector Autoregressive Model (VAR)

Two stationary time series $y_{1,t}$ and $y_{2,t}$.

VAR(1) in dimension $q = 2$:

$$\begin{cases} y_{1,t} &= \Gamma_{1,11}\, y_{1,t-1} + \Gamma_{1,12} y_{2,t-1} + e_{1t} \\ y_{2,t} &= \Gamma_{1,21}\, y_{1,t-1} + \Gamma_{1,22} y_{2,t-1} + e_{2t} \end{cases}$$

Covariance matrix of $(e_{1t}, e_{2t})'$ is Σ.
Vector notation: $\mathbf{y}_t = \Gamma_1 \mathbf{y}_{t-1} + \mathbf{e}_t$,

The VAR model

Let $\mathbf{y}_t$ be a q-dimensional stationary time series

Vector Autoregressive Model of order p:

$$\mathbf{y}_t = \Gamma_1 \mathbf{y}_{t-1} + \Gamma_2 \mathbf{y}_{t-2} + \ldots + \Gamma_p \mathbf{y}_{t-p} + \mathbf{e}_t \, ,$$

- Matrices Γ_j are autoregressive parameters
- $\mathbf{e}_t$ error with covariance matrix $\boldsymbol{\Sigma} = \boldsymbol{\Omega}^{-1}$.
- Standard estimation procedure: OLS equation by equation.

Section 2

Sparse VAR

Log-Volatilies for commodities (weekly data)

VAR model for $q = 10$ time series (and $T = 156$ observations)

- One lag
  - $1 \times (q \times q) = 100$ regression parameters
  - 45 unique elements in Σ
- Two lags
  - $2 \times (q \times q) = 200$ regression parameters
  - 45 unique elements in Σ

$\rightarrow$ Explosion of number of parameters

The VAR model: Overparametrization

ML estimators will be

- Not computable
- Inaccurate

Sparse estimation $\equiv$ many estimated parameters equal to zero

- Suitable if T is small relative to the number of parameters
- Easier to interpret
- Automatic variable selection
- Better estimation and prediction performance

Sparse Estimation: Lasso

In the multiple linear regression model

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k + \varepsilon$$

Minimization problem

$$\hat{\beta} = \underset{\beta}{\operatorname{argmin}} \, (y - X\beta)'(y - X\beta) + \lambda \sum_{l=1}^{k} |\beta_l|.$$

Tibshirani (1996)

Lasso for the VAR model

- Multiple equations
  - Partial correlation structure of the error term
    $\rightarrow$ Glasso of Friedman et al. (2008)
- Dynamic nature of the model
  - Selecting a time series into one of the equations = selecting the variable and all its lags
    $\rightarrow$ Group lasso (Yuan and Lin, 2006)

Penalized ML estimation

Rewrite the VAR in matrix notation:

$$\mathbf{Y} = \mathbf{Y_L}\boldsymbol{\Gamma} + \mathbf{E},$$

where

- $\mathbf{Y} = (\mathbf{y}_{p+1}, \ldots, \mathbf{y}_T)'$
- $\mathbf{Y_L} = (\mathbf{X}_{p+1}, \ldots, \mathbf{X}_T)'$ with $\mathbf{X}_t = (\mathbf{y}'_{t-1}, \ldots, \mathbf{y}'_{t-p})'$
- $\boldsymbol{\Gamma} = (\boldsymbol{\Gamma}_1, \ldots, \boldsymbol{\Gamma}_p)'$
- $\mathbf{E} = (\boldsymbol{e}_{p+1}, \ldots, \boldsymbol{e}_T)'$.

Penalized ML estimation (cont.)

Penalized negative log likelihood:

$$(\widehat{\mathbf{\Gamma}}, \widehat{\mathbf{\Omega}}) = \underset{\mathbf{\Gamma},\mathbf{\Omega}}{\text{argmin}} \ \frac{1}{T}\text{tr}\Big((\mathbf{Y} - \mathbf{Y_L}\mathbf{\Gamma})\mathbf{\Omega}(\mathbf{Y} - \mathbf{Y_L}\mathbf{\Gamma})'\Big) - \log|\mathbf{\Omega}|$$

$$+\lambda_1 \sum_{g=1}^{G} ||\gamma_g||_2 + \lambda_2 \sum_{k \neq k'} |\Omega_{kk'}|,$$

with

- γ_g is subvector of $\mathbf{\Gamma}$
- $G = q^2$ total number of groups.

Algorithm

Solving for $\boldsymbol{\Gamma}|\boldsymbol{\Omega}$*:*

$$\widehat{\boldsymbol{\Gamma}}|\boldsymbol{\Omega} = \underset{\boldsymbol{\Gamma}}{\text{argmin}} \ \frac{1}{T}\text{tr}\Big((\mathbf{Y} - \mathbf{Y_L}\boldsymbol{\Gamma})\boldsymbol{\Omega}(\mathbf{Y} - \mathbf{Y_L}\boldsymbol{\Gamma})'\Big) + \lambda_1 \sum_{g=1}^{G} ||\gamma_g||_2.$$

$\rightarrow$ groupwise lasso

Algorithm (cont.)

Solving for $\mathbf{\Omega}|\mathbf{\Gamma}$:

$$\widehat{\mathbf{\Omega}}|\mathbf{\Gamma} = \underset{\mathbf{\Omega}}{\operatorname{argmin}} \ \frac{1}{T}\operatorname{tr}\Big((\mathbf{Y}-\mathbf{Y_L}\mathbf{\Gamma})\mathbf{\Omega}(\mathbf{Y}-\mathbf{Y_L}\mathbf{\Gamma})'\Big) - \log|\mathbf{\Omega}| + \lambda_2 \sum_{k\neq k'} |\Omega_{kk'}|.$$

→ penalized inverse covariance estimation (glasso)

Selection of tuning parameters

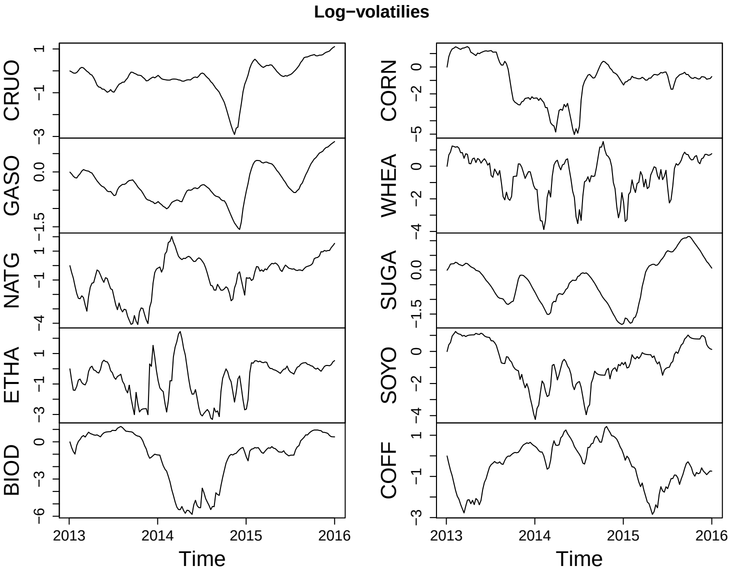
In the iteration step $\mathbf{\Gamma}|\mathbf{\Omega}$, select λ_1 to minimize

$$BIC_{\lambda_1} = -2 \log L_{\lambda_1} + k_{\lambda_1} \log(T),$$

- L_{λ_1} is the estimated likelihood using λ_1
- k_{λ_1} is the number of non-zero estimated regression coefficients.

In the iteration step $\mathbf{\Omega}|\mathbf{\Gamma}$, select λ_2 analogously.

Commodity prices: log volatilities (weekly data)



Networks from the VAR coefficients $\widehat{\Gamma}$.

Network with q nodes. Each node corresponds with a time series.

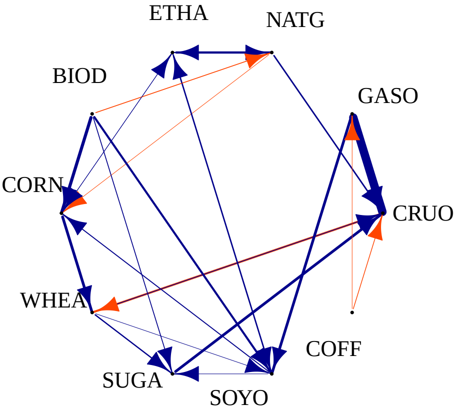
- draw an **edge** from node i to node j if

$$\sum_{p=1}^{P} |\widehat{\Gamma}_{p,ji}| \neq 0$$

Additionally (if $p = 1$)

- the edge **width** is the size of the effect
- the edge **color** is the sign of the effect (blue if positive, red if negative)

Network on the AutoRegressive coefficients



Diebold and Yilmaz (2015)

Granger Causality

Time series i is Granger Causing time series j

$$\updownarrow$$

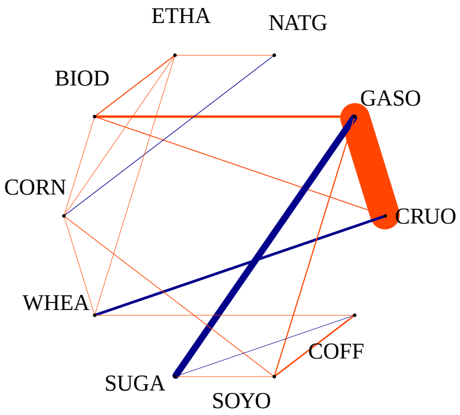
Time series i it has *incremental* predictive power in forecasting series j

$$\updownarrow$$

In the network there is an arrow going from node i to node j

Granger Causality test in high dimensions: Wilms, Gelper, Croux, 2016

Network on the precision matrix



References

- Rothman, Levina, and Zhu (2010), "Sparse multivariate regression with covariance estimation," *Journal of Computational and Graphical Statistics*.
- Hsu, Hung, and Chang (2008), "Subset selection for vector autoregressive processes using lasso," *Computational Statistics and Data Analysis*.
- Davis, Zang, and Zheng (2016), "Sparse vector autoregressive modeling," *Journal of Computational and Graphical Statistics*.
- Basu and Michailidis (2015), "Regularized estimation in sparse high-dimensional time series models," *Annals of Statistics*.
- Nicholson, Matteson, and Bien, J. (2017), VARX-L: Structured regularization for large vector autoregressions with exogenous variables, International Journal of Forecasting.
- Wilms, Basu, Bien, and Matteson (2017), bigtime: Sparse Estimation of Large Time Series Models, R package.
- ...
- Gelper S., Wilms I. and Croux C. (2016), "Identifying demand effects in a large network of product categories," *Journal of Retailing*.

Section 3

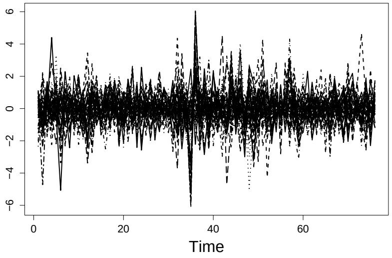
Application: a Market Response Model

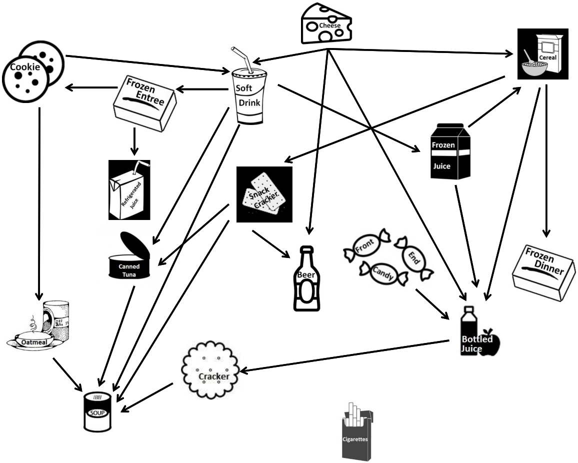
Data

Sales, promotion and prices for 17 product categories.

$T = 77$ weekly observations

VAR model for $q = 3 \times 17 = 51$ time series




Cheese
Cereal
Cookie
Frozen Entree
Soft Drink
Refrigerated Juice
Snack Cracker
Frozen Juice
Canned Tuna
Front
End
Candy
Beer
Frozen Dinner
Oatmeal
Bottled Juice
Cracker
SOUP
Cigarettes

Sales Forecasting

One-step ahead, using rolling window.
Averaged over the 17 product categories, *and the 15 stores*.

Method	Mean Absolute Forecast Error
Sparse	736
Bayesian: Minnesota	875
Bayesian: Normal-Inverse Wishart	1078
Least Squares	1298
Restricted LS	784

Sparse method significantly better ($P < 0.01$, Diebold-Mariano test)

Simulation Study

- Bayesian methods
  - Minnesota prior (Koop and Korobilis, 2009)
  - Normal-Inverse Wishart prior (Banbura et al, 2010)

Simulation Design: Sparse high-dimensional : $q = 10, p = 2, T = 50$

Method	Mean Absolute Estimation Error
Sparse	0.041
Bayesian: Minnesota	0.044
Bayesian: Normal-Inverse Wishart	0.077
Least Squares	0.157
Restricted LS	0.121

Section 4

Extensions

Multi-Class VAR

We have a chain of 17 stores.

- Estimate a VAR for each store separately.
- Pool the data for the 17 stores in one single class , and estimate one single VAR (or panel-VAR.)
- Multi-class estimation, where the data decide what the classes are.

Danaher, Wang, and Witten, D. (2014): joint graphical lasso.
Tibshirani, Saunders, Rosset, Zhu, and Knight (2005): Fused Lasso

Multi-class sparse VAR: K classes

$\mathbf{\Gamma}^{(k)}$, $k = 1, \ldots, K$ are minimizing

$$\sum_{k=1}^{K}\sum_{t=1}^{T}(\mathbf{y}_t^{(k)} - \mathbf{\Gamma}^{(k)}\mathbf{Y_L}^{(k)})'\widetilde{\mathbf{\Omega}}^{(k)}(\mathbf{y}_t^{(k)} - \mathbf{\Gamma}^{(k)}\mathbf{Y_L}^{(k)}) - TJ\log|\mathbf{\Omega}^{(k)}|$$
$$+\sum_{k=1}^{K}\sum_{i,j=1}^{J}\sum_{p=1}^{P}|\Gamma_{p,ij}^{(k)}| + \sum_{k\neq k'}^{K}\sum_{i,j=1}^{J}\sum_{p=1}^{P}|\Gamma_{p,ij}^{(k)} - \Gamma_{p,ij}^{(k')}| + \ldots$$

Wilms, I.; Barbaglia, L.; Croux, C. (2018), JRSSS-C

Furtermore

- t-Lasso for estimating VAR
- cointegration