

# A VAR-process approach to infer large-scale gene association networks from microarray time series data

Rainer Opgen-Rhein and Korbinian Strimmer  
Department of Statistics, Ludwig-Maximilians-Universität München

## Abstract

Graphical models provide a means to understand regulatory interactions among genes and gene products in a cell, and hence contribute to an enhanced understanding of systems biology. Here we investigate an approach that treats the observed gene expression over time as realizations of a vector autoregressive (VAR) process. We present a novel procedure that allows to infer VAR models from high dimensional small sample data. The new method is illustrated by analyzing highly replicated gene expression time series data.

## 1. Shrinkage regression

BEING in possession of efficient regression methods is crucial for the estimation of the VAR-model and therefore genetic network inference. The linear regression model is defined as

$$Y = A + XB + \epsilon.$$

with the ML-estimator for the matrix of the regression coefficients:

$$\hat{B} = (X'X)^{-1}X'Y$$

A drawback of this estimator in the “small  $n$ , large  $p$ ”-paradigm is, that  $X'X$  is often not invertible. Furthermore it can be shown that the empirical estimator is suboptimal in terms of MSE. Generally, regularization techniques allow to achieve a better estimator in terms of MSE by means of *shrinkage* (Stein-phenomenon).

To obtain a suitable estimator for the regression coefficients for large dimensional problems we introduce a novel procedure for *shrinkage regression*. The idea behind this regression method is that shrinkage is applied to the data to get “regularized” pseudo-data  $X^*$  and  $Y^*$  to which the ML-estimator can be applied.

## 2. Algorithm

APPLY shrinkage to combined centered observations  $X$  and response  $Y$

$$S = \Phi' \Phi = [XY]' [XY]$$

Construction of the shrinkage estimator  $S^*$ : convex combination of the unregularized estimator  $S$  and a suitable target  $S^{\text{Target}}$

$$S^* = \lambda S^{\text{Target}} + (1 - \lambda)S$$

The optimal shrinkage intensity  $\lambda$  minimizes the MSE risk function

$$R(\lambda) = E \left( \sum_{k=1}^p \sum_{l=1}^p (s_{kl}^* - s_{kl})^2 \right)$$

The minimum mean squared error  $R(\lambda^*)$  is achieved *exactly* and uniquely [4] for

$$\lambda^* = \frac{\sum_{k=1}^p \sum_{l=1}^p \text{Var}(s_{kl}) - \text{Cov}(s_{kl}, s_{kl}^{\text{Target}}) + \text{Bias}(s_{kl}) E(s_{kl} - s_{kl}^{\text{Target}})}{\sum_{k=1}^p \sum_{l=1}^p E[(s_{kl} - s_{kl}^{\text{Target}})^2]}$$

Decompose  $S^*$  via SVD and infer “regularized” pseudo-data:

$$S^* = VDV' = \Phi^{*'} \Phi^* \\ [X^* Y^*] = \Phi^* = \sqrt{D} V'$$

The matrix of shrunken regression coefficients can now be calculated

$$\hat{B}^* = (X^{*'} X^*)^{-1} X^{*'} Y^*$$

The algorithm explained above is the idea of the shrinkage regression. Nevertheless, in this form it has a drawback: the regularized pseudodata-matrix has dimension  $p \times p$  instead of  $n \times p$ , which means that it is computationally demanding. We use an efficient algorithm that makes use of the empirical *partial* variances  $\hat{S} = (\hat{s}_{kk})$  and empirical *partial* correlations  $\hat{r}_{kl}$ : this allows to apply shrinkage regression and to estimate the regression coefficients in an efficient way:

$$\hat{\beta}_k^* = \hat{r}_{kl}^* \cdot \sqrt{\frac{\hat{s}_{kk}^*}{\hat{s}_{ll}^*}}$$

## 3. Estimating Gene Association Networks

A VECTOR autoregressive process (here: VAR(1)-process) can be expressed by the following equation:

$$x_{t+1} = A + x_t B + \epsilon$$

We apply our method of shrinkage regression to efficiently estimate the parameters of the VAR-process. Note that in microarray experiments there are often repeated measurements of a short time series. To account for this special structure of the data, we define a data matrix which includes replication of longitudinal experiments and define “past” and “future” matrices that allow to regress all  $x_t$  to  $x_{t+1}$ . It is now possible to formulate a shrinkage regression estimator for the matrix of regression coefficients  $B$ :

$$\hat{B}^* = (X_p^{*'} X_p^*)^{-1} X_p^{*'} X_f^*$$

The regression coefficients reflect the influences among genes and can be used to infer gene association networks. Significant regression coefficients are included as edges in the gene association network. To identify these, statistical tests are needed. Here, we use “*local fd*” network search [3, 4]. We obtain a final network by visualizing all significant edges in a directed graph.

## 4. Results

### Simulation

SHRINKAGE regression was used to reconstruct a network with 50 nodes and 100 edges. For this, a VAR-process was simulated with 100 nonzero regression coefficients drawn from a uniform distribution from  $(-1)$  to  $(-0.2)$  and  $0.2$  to  $1$ . The network was reconstructed for a length of the time series between 5 and 200. Our method is compared with the OLS-estimator and Ridge regression.

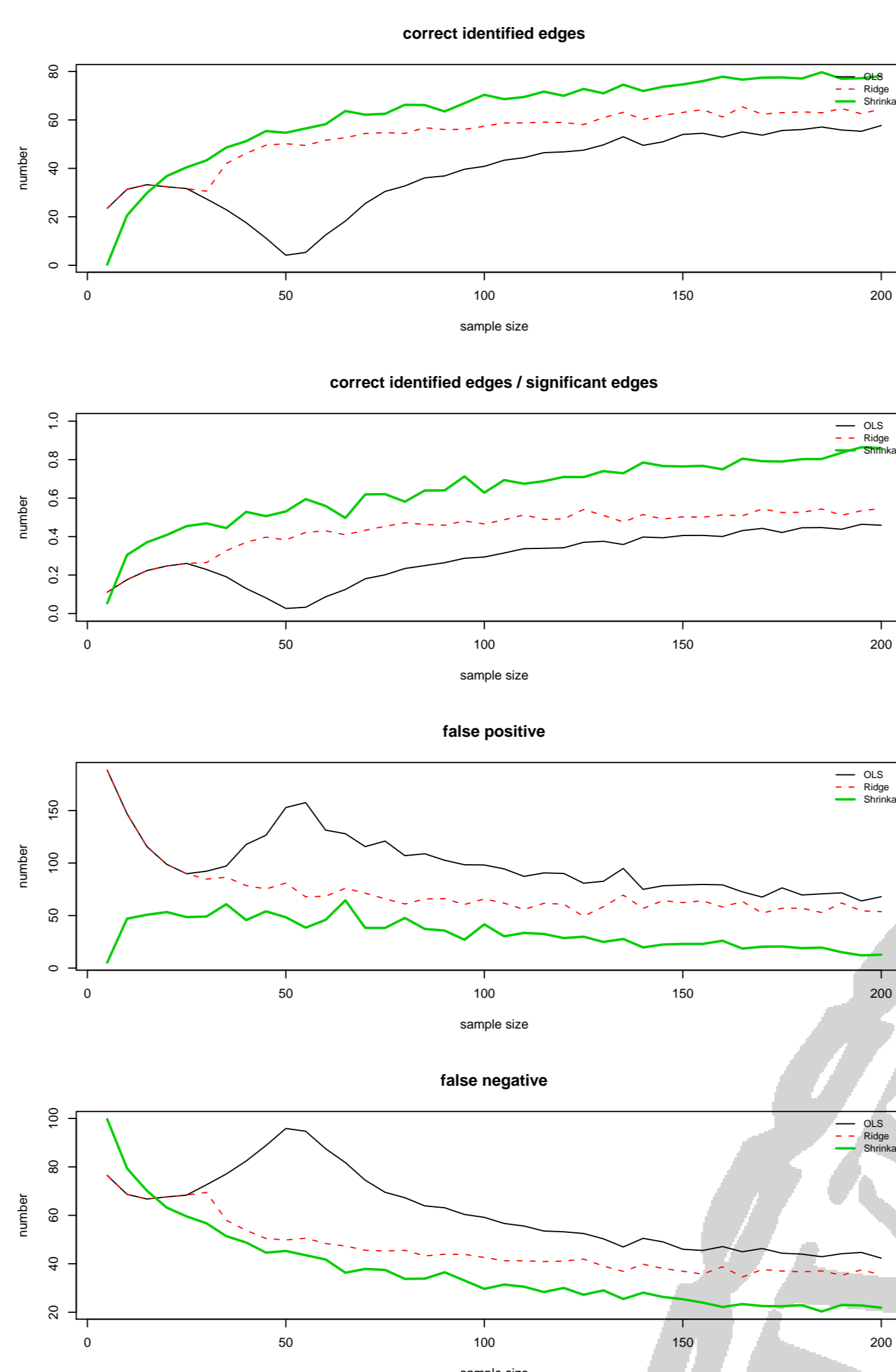


Figure 1: Reconstructed nodes of a VAR-process

A strength of our method also lies in its discriminatory power: we repeated the simulation for white noise (that means, all entries of the regression matrix are zero). Our shrinkage estimator strongly reduces the amount of false positive edges compared to OLS and Ridge regression if there is no structure in the data.

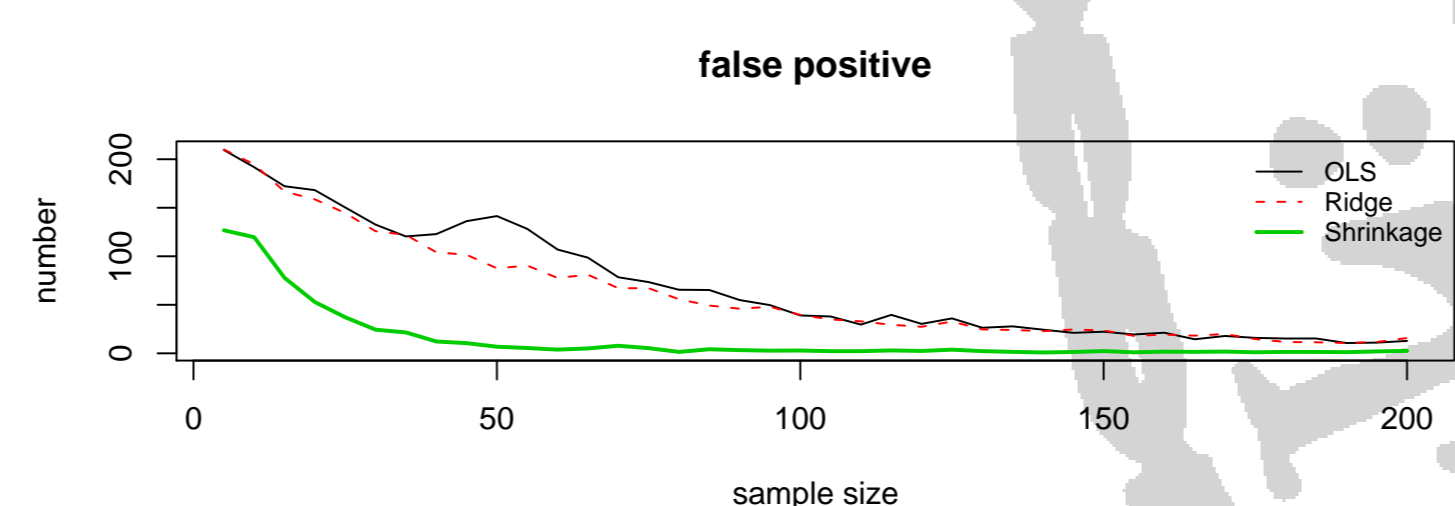


Figure 2: Reconstructed nodes of a VAR-process

### Microarray time series

WE employed shrinkage estimation of the VAR-process to a real world example. Specifically, we reanalyzed a microarray time series data set [2]. These data characterize the response of a human T-cell line (Jirkat) to a treatment with PMA and ioconomin, and consist of 10 time points with 44 replications each. We estimate the regularized regression coefficients for the VAR-process as described above and use the *locfd* algorithm to identify significant edges. The resulting network exhibits 53 significant edges and is displayed in figure 3.

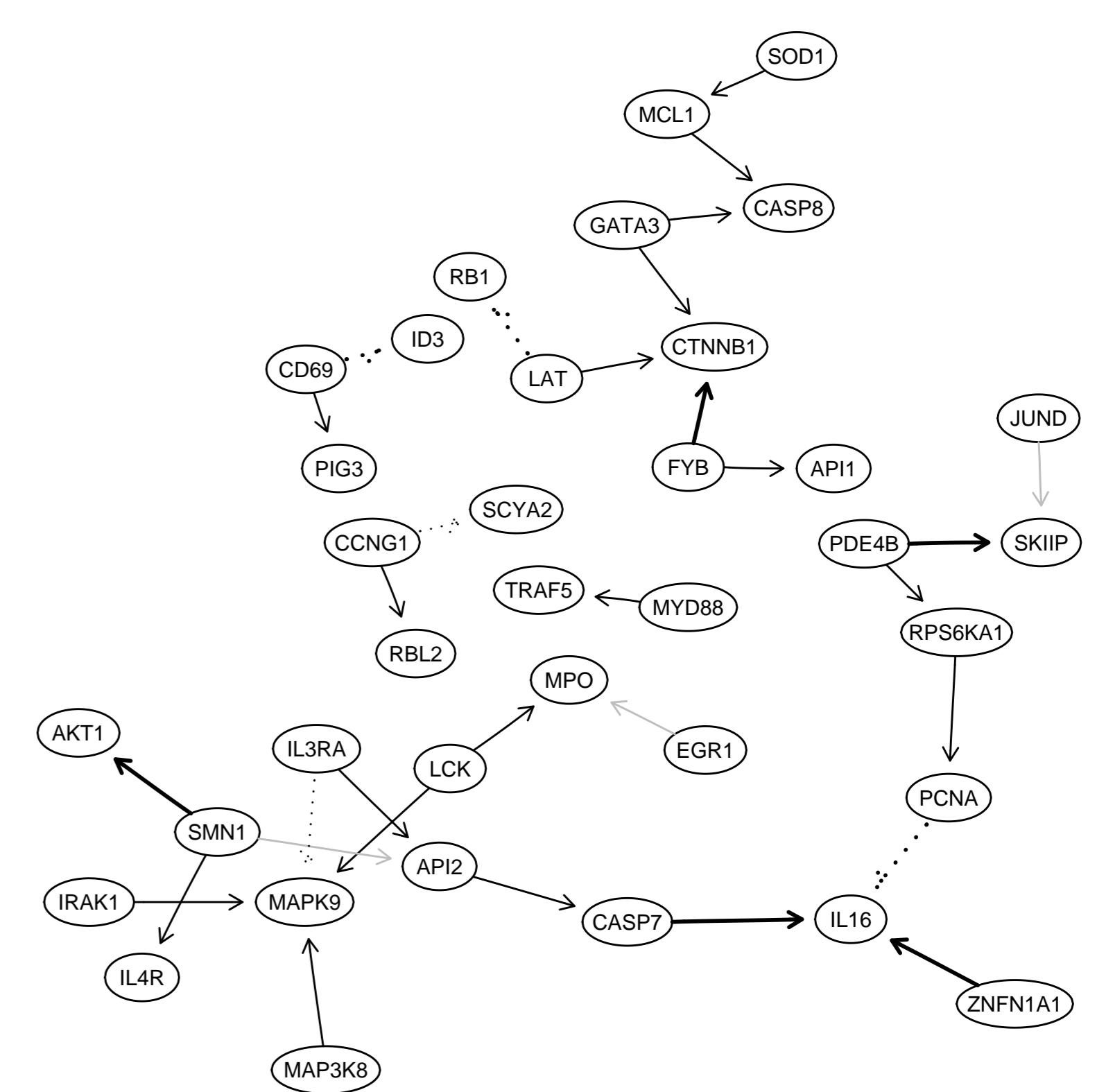


Figure 3: Gene dependency network inferred from human T-cell data [2]. We see, e.g., the activation of MAPK9 (gene alias: JNK2) which mediates immediate-early gene expression in response to various cell stimuli [1].

## 5. Summary

WE introduced a method to infer a gene dependency network from longitudinal data from the perspective of a VAR-process. For this, we suggest a regularized regression algorithm that uses Stein-shrinkage to be applicable to large-dimensional problems and employ this method to vector autoregressive processes.

## References

- [1] Mingo-Sion, A., Marietta, P., Koller, E., Wolf, D., Van Den Berg, C.: Inhibition of JNK reduces G2/M transit independent of p53, leading to endoreduplication, decreased proliferation, and apoptosis in breast cancer cells. *Oncogene* **23**(2) (2004) 596–604
- [2] Rangel, C., Angus, J., Ghahramani, Z., Lioumi, M., Sotharan, E., Gaiba, A., Wild, D.L., Falciani, F.: Modeling T-cell activation using gene expression profiling and state space modeling. *Bioinformatics* **20** (2004) 1361–1372
- [3] Schäfer, J., Strimmer, K.: An empirical Bayes approach to inferring large-scale gene association networks. *Bioinformatics* **21** (2005) 754–764
- [4] Schäfer, J., Strimmer, K.: A shrinkage approach to large-scale covariance matrix estimation and implications for functional genomics. *Statist. Appl. Genet. Mol. Biol.* **4** (2005) 32
- [5] Stein, C.: Inadmissibility of the usual estimator for the mean of a multivariate normal distribution. In Neyman, J., ed.: *Proc. Third Berkeley Symp. Math. Statist. Probab.* Volume 1., Berkeley, Univ. California Press (1956) 197–206
- [6] Whittaker, J.: *Graphical Models in Applied Multivariate Statistics.* Wiley, New York (1990)