

# INFERRING GENETIC NETWORKS VIA SUBSPACE SYSTEM IDENTIFICATION

Rainer Opgen-Rhein

Department of Statistics, University of Munich,  
D-80539 Munich, GERMANY, rainer.opgen-rhein@stat.uni-muenchen.de

## ABSTRACT

Gene expression time series data are promising to infer genetic regulatory networks. State space formulation can be used to model dynamic influence between genes. Subspace algorithms are generally an efficient way to identify this system. Due to the sparsity of genomic data standard subspace algorithms cannot be used. I will discuss these problems and present approaches to overcome this issue.

## 1. INTRODUCTION

The emergence of microarray technology made it possible to study thousands of genes simultaneously. One important aim of system biology is to identify possible interactions between these genes. Standard static techniques like clustering (see e.g. [1]), which are often used in biology, only allow to model co-regulation of genes. To infer networks based on interaction of genes dynamical methods are necessary. State space models are a useful tool to model dynamic systems and can be applied to gene expression time series data [2].

## 2. USING STATE SPACE MODELS TO INFER GENETIC NETWORKS

State space models assume, that observations  $y_t, t = 1, 2, \dots$  were generated from unobserved “states”  $x_t$ . The states define a first-order Markov process. The basic formulation is [3]:

$$x_{t+1} = Ax_t + w_t \quad (1)$$

$$y_t = Cx_t + v_t \quad (2)$$

This model can be used and expanded to describe gene expression data. The observations are seen as the gene expression levels measured with microarrays. The states capture effects which were not included in the microarray experiment (e.g. genes not measured on the microarray). To achieve this the gene expression levels  $g$  are fed back to the observation equation as well as to the state equation [4]:

$$x_{t+1} = Ax_t + Bg_{t-1} + w_t \quad (3)$$

$$g_t = Cx_t + Dg_{t-1} + v_t \quad (4)$$

- $A$ : effects between the latent variables (state dynamic matrix)

- $B$ : influence of gene expression from previous time point on hidden variables
- $C$ : influence of hidden variables on observed gene expression levels
- $D$ : gene-gene expression level influences at consecutive time points

The main interest focuses on the matrix  $CB + D$ , because it not only captures the direct influence from one gene to another ( $D$ ) but also indirect interaction via the hidden states. In addition, it can be shown if the gene expression model is stable, controllable and observable, the matrix  $CB + D$  remains invariant to any coordinate transformations of the state and is, therefore *identifiable* [2].

[4] estimated the system by using the EM-algorithm, a concept transferred in the Bayesian framework by [5] using the variational Bayes algorithm [6]. These concepts are in theory statistically efficient but suffer from some major drawbacks: they are iterative and might end up in a local minima. Furthermore they are very time-consuming and need mathematical operations which are not reliable with this complex multivariate datasets.

An alternative to this technique are subspace algorithms which will be introduced in the next chapter.

## 3. SUBSPACE METHODS

*Subspace algorithms* try to identify the matrices and the hidden state vectors of the state space model (the system). In contrast to the *EM-algorithm*, which assumes given matrices and infers then the hidden state vector (and afterwards tries to maximise the likelihood by changing the matrices and fixing the hidden state vector and repeats these steps iteratively), subspace algorithms directly infer the hidden state vector *without* using the system matrices. This is done by using linear algebra tools which only need the input–output data. Once the hidden states are known, the problem of system identification reduces to a linear least squares problem [7].

It can be shown that subspace algorithms are able to reconstruct the system matrices ( $A, B, C, D$ ) (and therefore the matrix  $CB + D$  which represents the gene interaction network) if certain assumptions are fulfilled. The specific data structure obtained from microarray experiments clearly violates several of them. The aim is to modify

the common stochastic subspace algorithms to be nevertheless able to make meaningful inferences from microarray data. In the following sections the main problems and means of overcoming them are explicated.

## 4. PROBLEMS

### 4.1. Closed-loop data

Common subspace algorithms assume a connection between input and output only via the state space model equations. In the gene interaction model the output data are fed back into the state space equations, therefore forming a closed loop system. [8] showed that this leads to inconsistent system identifications because the estimates of the states are biased. These authors also discuss a mathematical way to artificially avoid the closed loop.

### 4.2. Multiple observations

Standard subspace algorithms assume a single time series, while microarray experiments often provide several measurements at a specific time point. Hence, the algorithm needs to be modified to make efficient and meaningful statistical use of these multiple observations. Nevertheless this fact is conceptually not a problem, because the chance of greatly improving the estimation is provided.

### 4.3. Irregular sampling

Analysis of time series assumes fixed intervals between the measurements, which is often violated in biological experiments: normally gene expression is measured with increasing intervals after the first observations. If we do not assume that the genetic interactions slow down after an initial treatment, the gaps between measurements have to be filled. Filtering methods provide a way to estimate the gene expression at unobserved time points. Irregular sampling problems can also occur if there are different numbers of replications at certain time points or if not all genes are measured at a certain time point.

### 4.4. Short time series

Subspace algorithms are asymptotically exact when the length of the time series goes to infinity [7], whereas time series obtained from microarray experiments are often very short. This problem is strongly connected with the next, the small  $n$  large  $p$  problem.

### 4.5. Small $n$ large $p$

Although microarray experiments provide large data sets, the actual sample size ( $n$ ) is very small: this is due to the extremely high dimension ( $p$ ), the number of genes measured in an experiment. In static models this “only” means that the number of replications is very small compared to the number of variables. In dynamical models the length of the time series is also very small, as described in the previous section. Although conceptually different, these two problems are strongly connected.

This small  $n$  large  $p$  problem is widely recognized in the biosciences, but also in astronomy and econometrics.

A number of techniques like regularization have been developed to address this issues. These strategies have to be modified and adopted that they can be used in subspace algorithms and system identification.

## 5. CONCLUSIONS

Exact reconstruction of the genetic system seems impossible given the sparsity of the available data. State space models are nevertheless a promising tool to analyse the possible existence of interactions between genes. These interactions can be visualized in genetic networks which may then be used to generate hypothesis for future research. Subspace algorithms can help to identify these possible genetic interactions.

## 6. ACKNOWLEDGMENTS

This work in progress joint with Korbinian Strimmer, Department of Statistics, University of Munich.

## 7. REFERENCES

- [1] J. Dopazo, E. Zanders, I. Dragoni, G. Amphlett, and F. Falciani, “Methods and approaches in the analysis of gene expression data,” *Journal of Immunological Methods*, vol. 250, pp. 93–112, 2001.
- [2] C. Rangel, J. Angus, Z. Ghahramani, and D. L. Wild, “Modeling genetic regulatory networks using gene expression profiling and state space models,” in *Applications of Probabilistic Modelling in Medical Informatics and Bioinformatics*, D. Husmeier, S. Roberts, and R. Dybowski, Eds., pp. 269–293. Springer Verlag, 2004.
- [3] Z. Ghahramani and G. E. Hinton, “Parameter estimation for linear dynamical systems,” Technical report, Department of Copputer Science, University of Toronto, 1996.
- [4] C. Rangel, J. Angus, Z. Ghahramani, M. Lioumi, E. Sotheran, A. Gaiba, D. L. Wild, and F. Falciani, “Modeling T-cell activation using gene expression profiling and state space models,” *Bioinformatics*, vol. 20, pp. 1361–1372, 2004.
- [5] M. J. Beal, F. Falciani, Z. Ghahramani, C. Rangel, and D. L. Wild, “A Bayesian approach to reconstructing genetic regulatory networks with hidden factors,” *Bioinformatics Advance Access*, 2004.
- [6] M. J. Beal, *Variational Algorithms for Approximate Bayesian Inference*, Ph.D. thesis, University of Cambridge, 2003.
- [7] P. Van Overschee and B. De Moor, *Subspace Identification for Linear Systems*, Kluwer Academic Publishers, Boston, London, Dordrecht, 1996.
- [8] L. Ljung and T. McKelvey, “Subspace identification from closed loop data,” *Signal Processing*, vol. 52, pp. 209–215, 1996.