

A Bayesian approach to expression network component analysis

Chiara Sabatti¹ and Lars Rohlin²

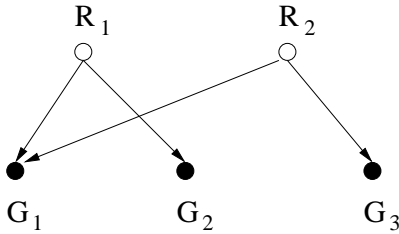
1. Departments of Statistics, and Human Genetics, UCLA, Los Angeles, CA

2. Department of Chemical Engineering, UCLA, Los Angeles, CA

Abstract—A semi-blind deconvolution method of analysis for gene expression data was proposed recently in a series of articles appeared in PNAS. We illustrate here how similar goals can be achieved in a Bayesian framework and how necessary information on the presence of binding sites can be obtained with Vocabulon, an algorithm based on a stochastic dictionary model.

I. INTRODUCTION

The following picture illustrate the type of network at the base of the analysis in this report.



The nodes named **R** in this graph, and drawn with an empty circle, represent concentration of active form of regulatory proteins. The empty circle indicate that we do not have measurements on their values. The nodes named **G** and represented with a full circle indicate genes, whose expression values we observe through a series of gene expression experiments. The arrows connecting nodes indicate which regulatory proteins have influence on which genes.

In the setting of Network Component Analysis [?], [3], the expression values of the different genes are linked to the values of the concentration of active form of regulatory proteins according to a simple model described in the following. Let e_{it} be the expression value of gene i at time-point t ; we are going to consider a total of N gene and M time-points. Let p_{jt} be the concentration of active form of protein j at time point t ; we are going to consider a total of L regulatory proteins. Then, we assume that

- $e_{it} \sim \mathcal{N}(\sum_{j=1}^L \alpha_{ij} p_{jt}, \sigma^2)$; Since for each gene we assume only a small number of the possible arrows in the graph above to exist, most of the α_{ij} will be zeroes. We will take care of making this clear in further formulations, but we leave here the full model for generality.
- that all the e_{it} are independent;
- and all e_{it} have identical variance.

Both the independence and the identical variance assumptions can be relaxed, modulo availability of sufficient data for

estimation. The normality assumption reflect the fact that a least square criteria appears as a satisfactory inferential principle. In the following we will use Σ to indicate be the variance covariance matrix of Γ . Note that the independence of the genes is *given* the values p_{jt} , that is really refers only to the random variation around their mean, that is determined by common regulatory proteins. In this model, both α_{ij} and p_{jt} are unknown. In this sense, it resembles a blind deconvolution problem. However, our setting is slightly different in that we know the number L of mixing components and we know which component participates in which observed signal e , that is we know the connectivity structure of the network described above. If we organize the expression values in a $N \times M$ matrix E , with each row representing a gene and each column representing an experiment, the formulation above can be written as:

$$E = AP + \Gamma,$$

with A and $N \times L$ matrix in which each column represent the action of a regulatory protein on all the genes in the system and P is a $L \times M$ matrix with each row representing the profile of each regulatory protein across experiments. The matrix Γ represent noise terms. Looking at the expression above, it is apparent that A and P represent a (noisy) decomposition of matrix E . Decomposing a matrix in the product of two is a problem that admits infinite solutions. Even if one considers as solutions the equivalent classes of all the matrices \tilde{P}, \tilde{A} obtained by pre or post-multiplication of a diagonal matrix X , ($\tilde{P} = XP, \tilde{A} = AX^{-1}$) uniqueness is not guaranteed. In order to identify a unique solution, one needs to impose further conditions on the problem. For example, the decomposition achieved in PCA requires orthogonality of the rows of P , the one in ICA require these to be realization of statistically independent random variables. The uniqueness of the NCA decomposition [?] is guaranteed by the presence of a sufficient number of zeroes in A . A more precise formulation of this problem is obtained if we look at the log-likelihood function derived from the model described:

$$\mathcal{L}(A, P|E) = \|E - AP\|,$$

where $\|X\| = \sum_{i=1, j=1}^{N, M} x_{ij}^2$, that is A and P enter the likelihood only in their product. If two pair of matrices A, P and \tilde{A}, \tilde{P} result in the same product F , then the parameters are non identifiable. In general, in this problem, we will be able to achieve at best quasi identifiability in the sense of

the equivalence classes described above. Requirements on the zero patterns of A to insure identifiability are given in [?] and sufficient conditions proved in Boscolo et al. (2004). In these references, we can also find a description of a likelihood maximizing algorithm, based on a two step least square iterative procedure. To describe this, it is convenient to introduce some notation. First, let $e_i = (e_{i1}, \dots, e_{iM})'$ represent the entire vector of observations on gene i . Let α_i^* be the vector formed by the non-zero components of $\alpha_i = (\alpha_{i1}, \dots, \alpha_{iL})'$; and let $P(i)$ be a matrix that has as column elements the vectors $p_j = (p_{j1}, \dots, p_{jM})'$ such that $\alpha_{ij} \neq 0$. Then, we can write:

$$\sum_{i=1}^N \sum_{t=1}^M (e_{it} - \sum_{j=1}^L \alpha_{ij} p_{jt})^2 = \sum_i^N (e_i - P(i)\alpha_i^*)'(e_i - P(i)\alpha_i^*). \quad (1)$$

Secondly, let us indicate with e the vector $(e_{11}, e_{21}, \dots, e_{N1}, e_{12}, e_{22}, \dots, e_{N2}, \dots, e_{1M}, \dots, e_{NM})'$, with p the similar vector $(p_{11}, \dots, p_{L1}, \dots, p_{iM}, \dots, p_{LM})'$; and with \mathcal{A} a block-diagonal matrix of dimensions $MN \times ML$ and with repeated diagonal blocks equal to the matrix obtained stacking as N rows, the vectors α_i' . Then, we can write:

$$\sum_{i=1}^N \sum_{t=1}^M (e_{it} - \sum_{j=1}^L \alpha_{ij} p_{jt})^2 = (e - \mathcal{A}p)'(e - \mathcal{A}p). \quad (2)$$

The two step-least square can be described as:

$$\begin{aligned} p_{\ell+1} &= (\mathcal{A}'_{\ell} \mathcal{A}_{\ell})^{-1} \mathcal{A}'_{\ell} e \\ \alpha_{i(\ell+1)}^* &= (P(i)^{\ell+1'} P(i)^{\ell+1})^{-1} P(i)^{\ell+1'} e_i \end{aligned}$$

Once estimates are obtained, their variability can be assessed using a bootstrap procedure. Note that given the fact that there is a scale and sign ambiguity, one has to be careful to use the same convention is normalization and sign assignment across samples. One relatively easy way around this problem, is to consider functions of P and A that are invariant with respect to the normalization of choice. We will call these ‘‘identifiable profiles’’. If we are interested in p_j , then the identifiable profile is $\sum_{i=1}^N \alpha_{ij} p_j$. If we are interested in α_{ij} , then the identifiable profile will be $\sum_t \alpha_{ij} p_{jt}$. Looking at these functions makes it evident, for example, that there is a confounding between the zeros of A and P . Clearly, if a transcription factor does not regulate any of the genes in the system, it cannot be estimated. Conversely, if a transcription factor is not activated in the experiments under study, its regulation strength on the genes in the set cannot be differentiated from zero. For completeness, we give a description of the bootstrap procedure that can be used. The parametric bootstrap is best suited to our problem, where the number of parameters to be estimated is tied to the number of observations. Hence, we use $E - \hat{A}\hat{P}$ to define a matrix of errors and we create a series of bootstrap datasets by adding to $\hat{A}\hat{P}$ a resampled version of the errors. Estimating the parameters in the bootstrap datasets and recording their values we can obtain confidence intervals for the parameters on the model (or their identifiable profiles). Bootstrap ‘‘with surgery’’ can also be used to conduct test

of hypothesis. In such case, bootstrap samples from the null hypothesis of $p_{jt} = 0$ are obtained setting $\hat{p}_{jt} = 0$ in the \hat{P} matrix and resampling again from the same error distribution as above. While the bootstrap seems to perform adequately in this problem, a theoretical proof of its validity is hard to achieve given the dependence of the parameter space on the number of observations. This represents one limitation of the current approach.

An other limitation can be identified in the requirements for identifiability of the model: while the conditions on A and P make intuitive sense in addition to guarantee identifiability, they are dictated by mathematical reasons and not by biology. Indeed, it is quite possible that the true regulatory network does not satisfy the identifiability conditions.

An other difficulty of the present implementation of NCA is that the distinction between zero and non zero elements of A is rigid: perfect information on the topology of the regulatory network is assumed. Such information is, however, typically available only on a small number of genes. There are, on the contrary, a number of methods that one can use to impute the topology on the base of sequence information. These methods tend to produce a higher number of false positive than false negative. It would be of interest to attempt NCA with these type of prior information on the network structure.

II. A BAYESIAN APPROACH

Taking a Bayesian approach allows to address the two issues raised above (identifiability and error estimation) in an easy and efficient way. Indeed, the assumption of any prior distribution introduces further constraints on the parameters that help overcoming the issues of identifiability. Moreover, the presence of a proper posterior distribution leads to a simple evaluation of the variability of the estimates. Given this general framework, we review what prior assumptions may be reasonable and computationally efficient and how it is possible to evaluate the posterior distribution appropriately.

It is again convenient to resort to the parameterization of the problem used to describe the maximization algorithm. Perhaps the easiest form of prior distribution assumes independence between all of α_i^* and p and chooses a gaussian form for their distribution:

$$\begin{aligned} p &\sim \mathcal{N}(0, \Pi) \\ \alpha_i^* &\sim \mathcal{N}(0, \Lambda) \end{aligned}$$

In both cases we choose a mean of zero, because the parameters in question are equally likely to be positive or negative. The covariance matrices Π and Λ can be diagonal or incorporate some prior information on dependence. In particular, Π can be chosen such that p_j is independent from p_k , but there is dependence across the values of the same regulatory protein across experiments. This is particularly useful when the experiments come from a time series situation, so that one can expect smooth variation of the values of p_{jt} . Notice that when the variance covariance matrices are diagonal and we

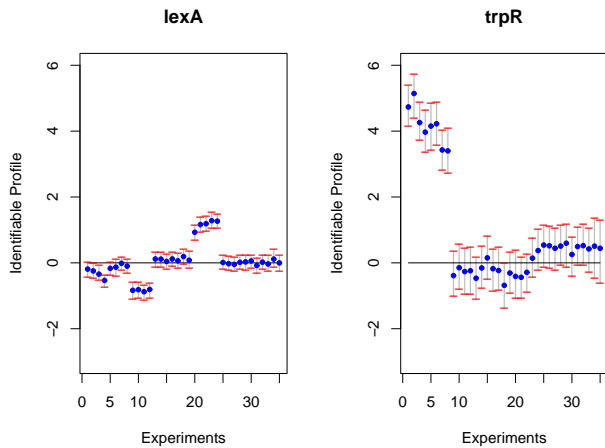


Fig. 2. Profiles of activation for the LexA and trpR regulons in the 35 analyzed experiments. Vertical bars indicate confidence sets.

Using the prediction of binding sites obtained with Vocabulon, and the binding sites which are known from literature, we were able to identify a set of 1448 genes regulated by 38 proteins. The breakdown of how many genes are regulated by each Transcription Factor (TF) is given in Figure 1. There were 6240 missing values in E (around 12%) of the data. Considering only the genes for which complete observations were available, would have reduced the number of genes to 302. Using our missing value imputation procedure we were able to use information coming from all the 1448 genes.

Biological knowledge on the nature of the microarrays experiments suggests that the LexA regulon should be activated in the UV experiments and the TrpR regulon should be activated in the Tryptophan timecourse. Indeed, as shown in Figure 2, the TFA profiles of these two regulators appear different from zero and positives in correspondence to those experiments. This reassures us both that the prediction of the Vocabulon algorithm for the zeros in the A matrix are correct and that the framework of Bayesian NCA is useful to capture the intracellular mechanisms.

ACKNOWLEDGMENTS

We thank James Liao for introducing us to the problem and allowing us to analyze unpublished data from his laboratory. Discussions with Riccardo Boscolo and Garreth James were instrumental to our understanding of the problem.

C. Sabatti thankfully acknowledges support from NSF (grant DMS0239427) and from NASA/Ames (grant NCC2-1364).

REFERENCES

[1] Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B, Shao Y. The complete genome sequence of *Escherichia coli* K-12. *Science*. 1997 Sep 5;277(5331):1453-74.

[2] Courcelle, J., Khodursky, A., Peter, B., Brown, P.O. and Hanawalt, P.C. (2001) Comparative gene expression profiles following UV exposure in wild-type and SOS-deficient *Escherichia coli*. *Genetics*, 158, 41-64.

[3] K Kao, Y Yang, R Boscolo, C Sabatti, V Roychowdhury, and J Liao Transcriptome-based determination of multiple transcription regulator activities in *Escherichia coli* by using network component analysis *PNAS* 2004 101: 641-646; bibitemliaoPNAS Liao, J., R. Boscolo, Y. Yang, L. Tran, C. Sabatti, and V. Roychowdhury (2003) "Network component analysis: Reconstruction of regulatory signals in biological systems" to appear in *PNAS*

[4] Khodursky AB, Peter BJ, Cozzarelli NR, Botstein D, Brown PO, Yanofsky C. DNA microarray analysis of gene expression in response to physiological and genetic changes that affect tryptophan metabolism in *Escherichia coli*. *Proc Natl Acad Sci U S A*. 2000 Oct 24;97(22):12170-5.

[5] K. Robison, A. M. McGuire, and G. M. Church, "A comprehensive library of DNA-binding site matrices for 55 proteins applied to the complete *Escherichia coli* K12 genome," *Journal of Molecular Biology*, vol. 284, pp. 241-254, 1998.

[6] C. Sabatti and K. Lange, "Genomewide motif identification using a dictionary model," *IEEE Proceedings*, vol. 90, pp. 1803-1810, 2002.

[7] C. Sabatti, L. Rohlin, K. Lange, and J. Liao (2004) "Vocabulon: a dictionary model approach for reconstruction and localization of transcription factor binding sites." UCLA Statistics department preprint # 369.

[8] C. Sabatti, L. Rohlin, M. Oh, J. Liao, "Co-expression pattern from DNA microarray experiments as a tool for operon prediction," *Nucleic Acid Research*, vol. 30, pp. 2886-2893, 2002.