Bayesian blind source separation applied to the lymphocyte pathway

Katrin Illner, *Helmholtz Zentrum München and Technische Universität München*,
katrin.illner@helmholtz-muenchen.de
Christiane Fuchs, *Helmholtz Zentrum München and Technische Universität München*,
christiane.fuchs@helmholtz-muenchen.de
Fabian J Theis, *Helmholtz Zentrum München and Technische Universität München*,
fabian.theis@helmholtz-muenchen.de

Abstract. In many biological applications one observes a multivariate mixture of signals, where both the mixing process and the signals are unknown. Blind source separation can extract such source signals. Often the data have additional structure, i. e. the variables (e. g. genes) are linked by an interaction network. Recently, we developed the probabilistic method `emGrade` that explicitly uses this network structure as a Bayesian network and thus performs a more appropriate separation of the data than standard methods. Here, we consider the application of `emGrade` to gene expression data together with a literature-derived pathway. Thanks to the probabilistic modeling, we can use model selection criteria and demonstrate the relevance of the pathway information for explaining the data. We further use estimates of missing observations to identify the most appropriate microarray probe sets for two genes that were not uniquely annotated after standard filtering. Finally, we identify genes relevant for the dynamics underlying the data; these genes were not detected without the network information.

Keywords. expectation maximization, model selection, gene expression data, gene regulatory networks

1 Introduction

Blind source separation (BSS) is a widely used method to extract informative signals from a multivariate observed mixture. In many applications the data have additional structure that can be exploited to achieve a more appropriate signal separation. Recently, our group developed two algorithms that explicitly include the network structure – `Grade` (graph-decorrelation algorithm) [5] and its probabilistic extension `emGrade` (expectation maximization graph-decorrelation algorithm) [4]. In the latter the network structure is modeled as a Bayesian network and parameters and source signals are estimated using expectation maximization. In this manuscript we demonstrate the application of `emGrade` to gene expression data where the genes are linked by a gene regulatory network. In Section 2, we briefly review the `emGrade` algorithm. As described in Section 3, we use publicly available microarray data for a lymphocyte pathway. In Section 4, we analyze the data using `emGrade`. The probabilistic framework enables us to use model selection criteria, and we find that the pathway information indeed improves our model. This has only been shown for synthetic data so far. Furthermore, we estimate missing observation values and determine the most appropriate microarray probe set for two genes that were not uniquely annotated after standard filtering. Finally, we characterize the estimated signals in terms of relevant genes and compare the gene sets from different observations. This leads the way to a biological interpretation of the estimated source signals. Section 5 concludes this paper.

Throughout the paper we use bold symbols to denote random variables and solid symbols to denote parameters and realizations of random variables, respectively.

2 The blind source separation method emGrade

In this section we shortly review the blind source separation method `emGrade` introduced in [4].

We assume observed Gaussian random variables $\boldsymbol{X} = \big(\boldsymbol{x}(i)\big)_{i=1}^{N}$ with state space $\mathbb{R}^m$ that are generated by the following linear mixing model:

$$\boldsymbol{x}(i) = A\boldsymbol{s}(i) + \mu + \boldsymbol{\varepsilon}(i)\ , \quad i = 1, \ldots, N\ . \tag{1}$$

Here, $A \in \mathbb{R}^{m\times q}$ denotes the mixing matrix, $\mu \in \mathbb{R}^m$ is a common mean vector for all i, and $\boldsymbol{\varepsilon}(i) \sim \mathcal{N}(0, \sigma^2 I)$ is additive measurement noise. The latent variables $\boldsymbol{S} = \big(\boldsymbol{s}(i)\big)_{i=1}^{N}$ are normally distributed with state space $\mathbb{R}^q$ ($q \leq m$). The components of these variables represent the source signals we are interested in, i. e. we have a source signal $\boldsymbol{s}_k = (\boldsymbol{s}_k(1), \ldots, \boldsymbol{s}_k(N))$ for $k = 1, \ldots, q$.

To define the (joint) distribution of the latent variables we assume a weighted directed acyclic graph $G = (V, E, \Lambda)$ that is determined a priori. Let $V = (v_1, \ldots, v_N)$ be the set of nodes, $E \subset V \times V$ the set of edges, and let $\lambda_{ij} \in \mathbb{R}$ denote the weight assigned to the edge $(v_i, v_j) \in E$. We assume that the latent variables $\boldsymbol{S}$ form a Bayesian network with respect to G, i. e. the latent variables are associated to the nodes V and the joint distribution decomposes as

$$\text{(A0)} \quad \mathrm{p}(\boldsymbol{S}) = \prod_{i=n_0}^{N} \mathrm{p}(\boldsymbol{s}(i) \mid \mathbf{Pa}(i)) \prod_{i=1}^{n_0-1} \mathrm{p}(\boldsymbol{s}(i))\ .$$

Here, $\mathbf{Pa}(i)$ denotes the vector of all latent variables associated to the parent nodes of v_i, and we assume that $v_1, \ldots, v_{n_0}$ are root nodes. We then make the following stationarity and scaling assumptions where v_i and v_j are adjacent nodes:

$$\text{(A1)} \quad \mathbb{E}[\boldsymbol{s}(i)] = 0_q\ ,$$

$$\text{(A2)} \quad \mathrm{Cov}(\boldsymbol{s}(i), \boldsymbol{s}(i)) = I_q\ ,$$

$$\text{(A3)} \quad \mathrm{Cov}(\boldsymbol{s}(i), \boldsymbol{s}(j)) = \lambda_{ij}\, D\ .$$

We denote D as graph-delayed covariance, and we assume that it is diagonal. The assumptions (A0)-(A4) define a unique distribution of $\boldsymbol{S}$ which is characterized by the conditional distributions in (A0). The parameter D occurs in (A0) as different rational terms.

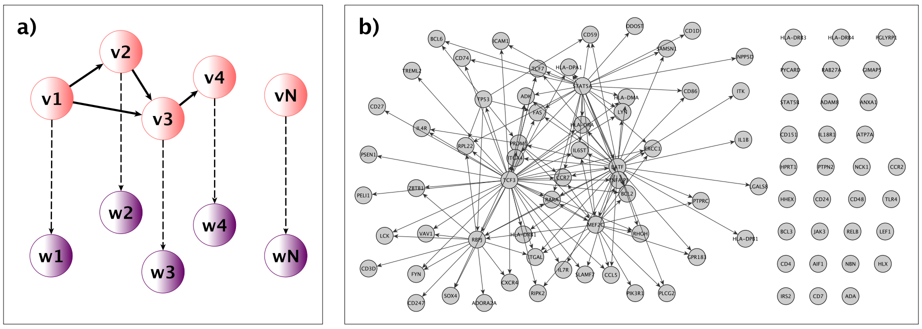

Figure 1: **Bayesian network for emGrade.** a) Graphical representation of the Bayesian network for `emGrade` with latent variables in red and observed variables in purple. The dependence between the latent variables is with respect to a known network structure, for instance a gene regulatory network. b) The pathway "lymphocyte activation" (net1) derived from the Genomatix database.

We now expand the Bayesian network and add nodes $w_1, \dots, w_N$ that represent the observed variables (Figure 1a). For all $i = 1, \dots, N$ we insert an edge (v_i, w_i), and the conditional distribution of the associated random variables is given as $\boldsymbol{x}(i) \mid \boldsymbol{s}(i) \sim \mathcal{N}(A\boldsymbol{s}(i) + \mu, \sigma^2 I)$. In the resulting Bayesian network we can estimate the latent variables $\boldsymbol{S}$ and the model parameter $\theta = (A, \mu, \sigma^2, D)$ using expectation maximization. For the expectation step we use the Bayes net toolbox [6] and estimate the latent variables from their posterior distribution. If data points are missing (i. e. some variables $\boldsymbol{x}(i)$ are unobserved) we can simply treat them as additional latent variables. For the maximization step we derive explicit updates rules for A, μ and σ^2 and use numerical maximization for D. All (diagonal) entries of D can be estimated separately, and the domain depends on the network structure. Both steps are repeated alternately until convergence. Here we assume convergence if the change for all parameter entries is less than 10^{-8}, or if a maximum number of 10 000 iterations is achieved. The resulting values for the parameters and source signals then define the `emGrade` estimate.

3 The data

For the application of `emGrade` we consider gene expression data that are accessible online, and we use the Genomatix database [3] to derive a network structure for the gene interactions.

Gene expression data and pre-processing

In Calvano et al. [1] four healthy humans were treated with intravenous endotoxin, and gene expression measurements of blood leukocytes were taken at time points 0, 2, 4, 6, 9, and 24h after endotoxin administration. In a control study the leukocytes of four non-treated humans were taken at the same time points. After quantile normalization and filtering with the `limma` R-package [7] we get normalized expression values for 12 683 human genes. For source separation we consider a subset of $N \sim 100$ genes that are associated with a specific pathway. The derivation of the pathway is discussed in the next paragraph. We further divide the data into the measurements for each individual. We thus have observations LPS1-4 from the four treated persons and observations PT1-4 from the non-treated persons. Each selected gene corresponds to an observed random variable, and since the measurements are taken at six time points we have $m = 6$ as the dimension of the observed variables.

In simulations we found that the performance of `emGrade` increases if the variance of the observed variables has a similar range compared to the variance of the unobserved variables. Since we assume $\boldsymbol{s}(i) \sim \mathcal{N}(0, I)$ in (A2) we scale the variance of the components of $x(i)$ to 1, accordingly.

Literature-derived pathways

In our BSS method we assume an initially known network that describes the dependencies between the variables (genes). To derive such a network structure we use pathway information from the Genomatix Pathway System (GePS) [3]. Based on the expression data from [1] the database provides (amongst others) biological processes that are associated with changes between treatment and control group. One highly significant pathway is “lymphocyte activation” (net1) which consists of 91 genes and 138 edges (Figure 1b). For comparison we also investigate the less significant sub-pathway “cell proliferation” (net2). Since no further information about the strenght of interaction is available, we fix all egde weights at $\lambda_{ij} = 1/\#\{\text{parents of } v_j\}$.

4 Results

In the following, we apply `emGrade` to the gene expression data and pathways from the previous section and present our main findings.

Comparison of different networks

In a first investigation we apply `emGrade` to all patients LPS1-4 and PL1-4 separately. As network structures we consider net1 and net2 and for comparison a network without any edges (net0). If we fix one data set, we can compare the BIC values for different network structures and different numbers of source signals $q = 1, 2, 3, 4$. Figure 2 indicates that for all data sets the informative net1 is more appropriate compared to net0 (lower BIC values). Furthermore, we find that for the treatment groups LPS1-4 a higher number of source signals is preferred.

Missing observation values

We now investigate the predictive power of our model for missing observation values. As already stated in Section 2, we can easily treat missing observations as additional latent variables in our Bayesian network. We therefore leave out the observation value for one gene in the data LPS1 and compare the estimate $\hat{x}(i) = \hat{A}\hat{s}(i) + \hat{\mu}$ to the true observation $x(i)$. We assume net1 to be the underlying network and use genes that are highly connected as well as genes that are not connected. Figure 3 shows the Euclidean distances $\|x(i) - \hat{x}(i)\|_2$ for 10 different missing genes and for $q = 1, 2, 3, 4$ estimated source signals. For comparison we estimate parameters and source signals from the complete data set. As expected, we find a better agreement of $\hat{x}(i)$ with $x(i)$ in this case.

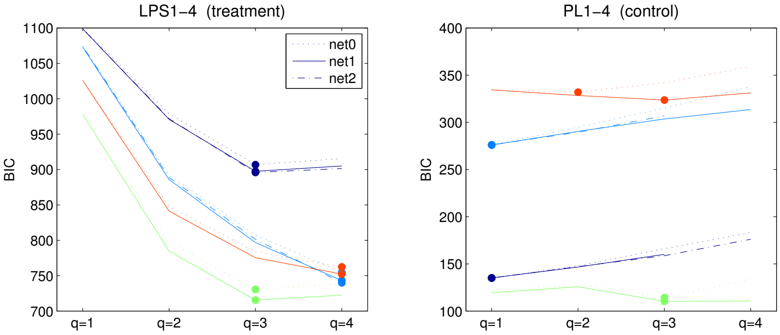

Figure 2: **Comparison of different networks.** The plots show the BIC values for patients LPS1-4 (left) and PL1-4 (right) in case of $q = 1, 2, 3$ and 4 source signals. As network structures we consider net0 and net1 and for LPS1-2 also net2. The different patients are coded in different colors and the dots indicate the value for q with the lowest BIC value.

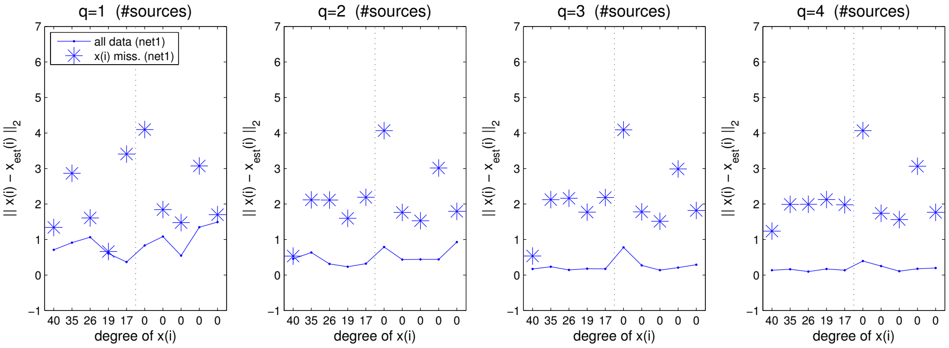

Figure 3: **Reconstruction of missing observation values.** We consider data LPS1 where we leave out the measurements of one gene, and we use net1 as network structures. We then apply `emGrade` with $q = 1, 2, 3, 4$ sources (left to right). The stars illustrate the Euclidean distances between the estimates $\hat{x}(i) = \hat{A}\hat{s}(i) + \hat{\mu}$ and the true observation values $x(i)$ where we consider different missing genes with high and low connectivity (degree). The solid lines show the corresponding distances when $\hat{x}(i)$ is estimated from the complete data set.

The estimation of missing observations provides a useful feature in the present data situation: The genes HLA-DRB1 and HLA-DRB3 from net1 are annotated to 5 and 2 probe sets of the microarray chip. Gene filtering performed with the `limma` R-package omits these genes and one does not know which probe set provides the most appropriate expression values. We therefore treat both genes as missing observations and compare our estimates to the measurements of the different probe sets. Table 1 shows the microarray measurements of all probe sets together with our estimates. The comparison suggests to use the 4th probe set as observation for HLA-DRB1 and the 2nd probe set as observation for HLA-DRB3.

	HLA-DRB 1						HLA-DRB 3		
time	obs.1	obs.2	obs.3	obs.4	obs.5	est.	obs.1	obs.2	est.
0h	4.99	5.23	5.26	**5.20**	2.46	5.44	5.20	**2.46**	2.52
2h	4.24	4.26	4.38	**4.11**	2.76	3.74	4.11	**2.76**	2.29
4h	4.26	4.65	4.47	**4.43**	2.54	4.66	4.43	**2.54**	2.81
6h	4.52	4.81	4.58	**4.66**	2.76	4.60	4.66	**2.76**	2.38
9h	4.66	5.21	5.22	**5.04**	2.62	5.15	5.04	**2.62**	2.61
24h	4.86	5.28	5.12	**5.23**	2.08	5.11	5.23	**2.08**	2.21

Table 1: **Identification of the most appropriate microarray probe set.** The table shows the microarray measurements from LPS1 at all probe sets that are linked to the genes HLA-DRB1 and HLA-DRB3. If we treat both genes as missing observations we get estimated observations (red). The comparison to the measurements identifies the most appropriate annotated probe set for both genes (bold symbols).

Genes associated to source signals

Next, we determine key genes associated with the estimated source signals, and we compare source signals that are estimated from different observations. Let $s_k = (s_k(1), \ldots, s_k(N))$ be an estimated source signal. Based on a cut-off $c > 0$ we select all genes with absolute value larger than c. This yields a set of key genes that characterize s_k. With this we can compare key genes of source signals that are estimated from different observations. We compare the treatment groups LPS1-3 and the control groups PL1-3. Since the estimated source signals are unique only up to sign and permutation we first align the source signals and minimize

$$\min_{P_1,P_2} \{ \| P_1S_1 - P_2S_2 \|_2 + \| P_1S_1 - S_3 \|_2 + \| P_2S_2 - S_3 \|_2 \} .$$

Here, P_1 and P_2 are matrices with one entry ±1 per row and column and all other entries equal to zero. Figure 4 illustrates the alignment of source signals and Figure 5 indicates that we have a higher key gene agreement for the treatment groups LPS1-3 compared to the control groups PL1-3. For the control groups only the network without edges (net0) yields a source with high agreement of key genes.

5 Discussion and conclusion

In this manuscript we applied the recently developed blind source separation method `emGrade` to gene expression data. The method aims to separate multivariate data with known network structure into informative source signals. We discussed the pre-processing of publicly available microarray data consisting of treatment data LPS1-4 and control data PL1-4. From the Genomatix database we derived the "lymphocyte activation" pathway which reflects differences between the control and treatment group. We then applied `emGrade` to this data set.

In comparison to a network without egdes, the pathway information improved our estimates and we found lower BIC values – this was true for the treatment and the control group. Nevertheless, the pathway information played a major role particularly for the treatment group where more source signals were preferred. We further investigated the estimation of missing observation values. For two genes from the lymphocyte pathway standard annotation to one unique

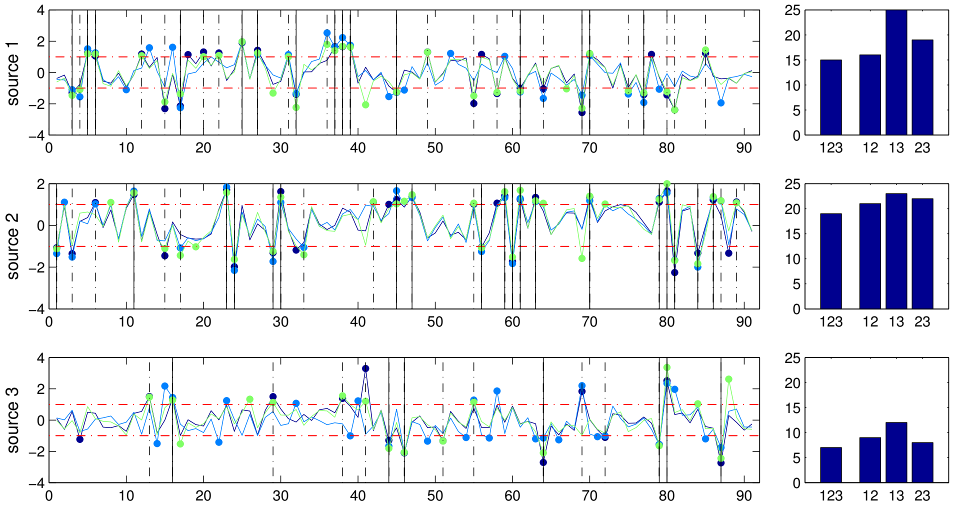

Figure 4: **Alignment of source signals and intersection of key genes for LPS.** For patients LPS1, LPS2 and LPS3 and network structure net1 we determine the `emGrade` source estimates. The plots on the left show the aligned source signals (different patients in different colors) together with the selected key genes (dots). The horizontal red lines show the cut-off for key gene selection. Solid vertical lines indicate genes that are key genes in the aligned sources from all patients, dashed vertical lines indicate genes that are key genes in the aligned sources from at least two patients. The bars on the right provide the counts of key genes in all estimates (123) and the counts of key genes in two estimates (12), (13) and (23).

microarray probe set failed. When treated as missing observations `emGrade` could identify the most appropriate annotated probe set in both cases. Finally, we characterized the estimated source signals in terms of key genes, i. e. genes with high absolute value in the respective source signals. We found a high number of key genes (per source) that were in agreement with LPS1-3. For PL1-3 these numbers were lower. This might again indicate that the "lymphocyte activation" pathway better explains the dynamics in the treatment group.

In our ongoing work we extend the proposed BSS model and consider different pathway structures for each source signal. We aim to separate the data into a pre-defined set of pathways and determine the impact of the estimated source signals in terms of the graph-delayed covariance.

Acknowledgements

The authors very much thank Steffen Sass and Nikola Müller for helping with the data preprocessing and the pathway extraction from the Genomatix database. We also thank Justin Feigelman for careful proofreading. Furthermore, the work was financially supported by the German Federal Ministry of Education and Research (BMBF) within the GerontoSys project

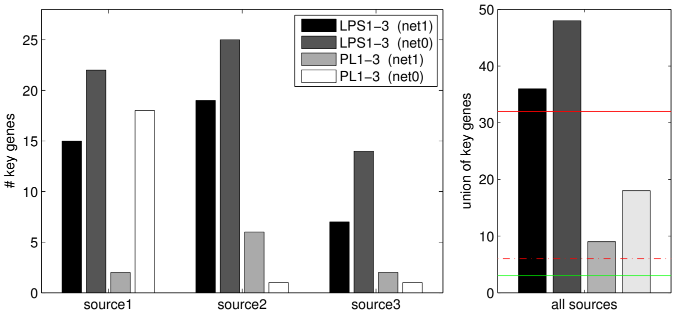

Figure 5: **Intersection of key genes for LPS and PL.** For patients LPS1-3 and PL1-3 and network structures net1 and net0 we determine the `emGrade` source estimates. The left figure shows the counts of genes that are key genes in all aligned source estimates from LPS1-3 or PL1-3, respectively. The right figure gives the total number of key genes in all sources. The solid red line indicates the count of identical key genes in LPS1-3 (net1) and LPS1-3 (net0), the dashed red line indicates the corresponding count for PL1-3. The green line is the count of identical key genes in all four groups.

“Stromal Aging” (Grant No. FKZ 0315576C), the German Research Foundation (DFG) within the project “InKoMBio” (Grant No. BO 3834/1-1), and the European Union within the ERC grant “LatentCauses”.

Bibliography

[1] Calvano, S. E., et al. (2005). *A network-based analysis of systemic inflammation in humans.* Nature, **437**(7061), 1032–1037.

[2] Friedman, N., Linial, M., Nachman, I., and Pe’er, D. (2000). *Using Bayesian networks to analyze expression data.* Journal of Computational Biology, **7**(3-4), 601–620.

[3] Genomatix Pathway system (GePS), http://www.genomatix.de/

[4] Illner, K., Fuchs, C. and Theis, F. J. (2012). *Blind source separation using latent Gaussian graphical models.* Ninth International Workshop on Computational Systems Biology, WCSB 2012, 34–46.

[5] Kowarsch, A., Blöchl, F., Bohl, S., Saile, M., Gretz, N., Klingmüller, U., and Theis, F.J. (2010). *Knowledge-based matrix factorization temporally resolves the cellular responses to IL-6 stimulation.* BMC Bioinformatics, **11**, 585–598.

[6] Murphy, K., et al. (2001). *The Bayes net toolbox for Matlab.* Computing Science and Statistics, **33**(2), 1024–1034.

[7] Smyth, G. K. (2005). *Limma: linear models for microarray data.* In: Bioinformatics and Computational Biology Solutions using R and Bioconductor, Springer, 397–420.