Available online at www.sciencedirect.com

ScienceDirect

IFAC Papers Online CONFERENCE PAPER ARCHIVE

IFAC-PapersOnLine 49-7 (2016) 472–477

New iterative approach (ISNCA) for constrained matrix factorization methods

Nadav Bar * **Naresh D. Jayavelu** **

* *Dept. of Chemical Engineering, Norwegian University of Science and Technology, Trondheim NO7491, Norway (e-mail: nadi.bar@ntnu.no).*
** *Turku Centre for Biotechnology, University of Turku and Åbo Akademi University, 20521 Turku, Finland (e-mail: njayavel@btk.fi)*

Abstract:
Gene regulation networks are complex, often involve thousands of genes, regulators and the connections between them. To understand the complex interactions between these genes and regulators with time, large empirical data is used the so called time-series gene expression data. Many statistical tools are used to analyze this data but they often impose restrictions that reduce the size of the network and make the solution less feasible from a biological perspective. We developed the iterative subnetwork component analysis (ISNCA), a method that decomposes the empirical data of two or more overlapping subnetworks with joint components at one iteration, and updates the solution at the next iteration by subtracting the contribution of each of the subnetworks. This predict - update method managed to relax the restrictions and solve larger networks. We generalized the method in this paper to include both regulators and genes in the joint partition, and demonstrated its accuracy using a synthetic network with a known matrix decomposition. We also applied the ISNCA on large biological data taken from mice cells and obtained larger and more accurate solutions than achieved by previous methods.

Keywords: Gene expression data, Network analysis, Data analysis, principle component analysis, Network component analysis, Iterative methods, ISNCA, matrix decomposition, big data

1. INTRODUCTION

One of the main drivers of cellular processes is the gene expression, the mechanism that produce proteins through a complex regulatory network. These networks involve thousands of target genes (TGs), transcription factors (TFs) and the interactions between these. Data obtained from measurements such as microarray and RNA sequencing are broadly used to asses the gene expression levels. Many researchers attempted to make sense of these complex networks and analyzed the data. As a consequence, several matrix decomposition methods were developed in the recent decades in order to extract meaningful biological information. We can mention for instance the known principle component analysis (PCA, see Raychaudhuri et al., 2000; Wall et al., 2003), singular value decomposition (SVD, see Wall et al., 2001, 2003), independent component analysis (ICA, Liebermeister, 2002), and partial least squares regression, (PLSR, Boulesteix and Strimmer, 2005). To further understand our measurements, we wish to decompose the data matrix E into a component matrix A (also called *scores*) and a coefficients matrix P (also called *loadings*), or

$$E = AP \tag{1}$$

In PCA for instance, the scores A are given by the left singular vectors of E, multiplied with the corresponding singular values, and the loadings matrix P are the right singular vectors of E. To exploit available biological knowledge in the literature and databases, several groups (the first by Liao et al., 2003) developed the network component analysis (NCA), a method that predicts the activity of the TFs on the TGs by matrix decomposition provided with a-priori biological knowledge and gene expression data. This development led to a large number of publications involving NCA. However, several restrictions on the decomposed matrices ensure a unique solution up to a scaling factor (Liao et al., 2003). These restrictions prevent often the incorporation of known biological phenomenons, such as redundancy of regulation and co-regulation. Without these, the solution is less feasible from biological perspective (Jayavelu et al., 2015) and is simply reduced to a theoretical and meaningless solution. Furthermore, these restrictions reduce the size of the network significantly, potentially losing TFs and TGs that may be important for the system under study.

To address this problem, we developed a novel method called Iterative Sub-Network Component Analysis (ISNCA, Jayavelu et al., 2015). Our method iterates between two or more smaller subnetworks, each satisfies the restrictions of the NCA, and provides a solution to the entire (larger) network (that usually cannot be solved by the NCA). The ISNCA employs a predict-update strategy, that incorporates the information of the joint partitions of the networks at one iteration, to update the following one. In this manner, our method relaxes the NCA restrictions on the entire network, and significantly enlarge the size

⋆ Corresponding author: nadi.bar@ntnu.no

10.1016/j.ifacol.2016.07.387

of the solution. Additionally, it relaxes the restrictions of redundancy and cooperativity, and thus render the solution more biologically feasible. We demonstrated the ISNCA previously on a small network (5-10 components) and on large biological measurements taken from breast cancer cells (Jayavelu et al., 2015), but it was not tested on data with a pre-known solution to evaluate its accuracy and performance. Here, we extended the work to a more general case, where the joint partition can contain TFs and their TGs, and studied the accuracy of its solution using a synthetic network with 15 genes and 8 TFs with pre-known activity patterns. The ISNCA yielded predictions with high fidelity, where traditional NCA failed to apply. We also tested the ISNCA on a large expression data with several replicates, taken from mouse cells, and showed that our method enlarged the size of these networks by 15%.

The remaining of the paper is organized as follows: Section 2 describes the ISNCA method, and presents the concept with an example. Section 3 presents simulations of a synthetic case study, and compares the true temporal activities of the TFs to the activities predicted by the ISNCA. We demonstrate the ISNCA on real gene expression measurements, taken from mice T-cells. We discuss some aspects of the ISNCA, and conclude in section 4.

2. METHODS

2.1 Mathematical formulation

One of the matrix decomposition methods mentioned in the previous section is the network component analysis (NCA). The objective is to decompose the matrix E into the matrices A and P, where A has a predefined structure. More specifically,

$$E = AP + \epsilon \tag{2}$$

where the measurement matrix $E \in \mathbb{R}^{n\times m}$ contains the m samples (time points or conditions) of the n components, the matrix $A \in \mathbb{R}^{n\times l}$ is the topology matrix that defines the sign and size of the connections of n components to their l regulators. $P \in \mathbb{R}^{l\times m}$ represents the temporal activity of the regulators, or how each of the l regulator's pattern propagates with time. The term ϵ is associated with the measurement noise. The decomposition (2) of E (given) into A and P (unknown) can be achieved by solving the following optimization problem

$$\min \|E - AP\| \tag{3a}$$
$$s.t.\ A \in Z_0 \tag{3b}$$

Here Z_0 (given) represents the a priori known structure for A where certain elements are fixed to be zero. A zero entry occurs when no data or knowledge exists that proves any association (connection) between the regulators and their corresponding components (Liao et al., 2003).

Three conservative, but essential restrictions to ensure a unique solution to (3) exist (discussed previously in Liao et al., 2003). Briefly, in addition to the Z_0 constraint, (i) the matrix A must have a full column rank, (ii) each column of A must have at least $(l-1)$ zeros, and (iii) the predicted matrix P must have a full row rank. To satisfy these conditions, a pruning procedure removes (usually randomly) columns and rows of A until the first two conditions are satisfied. This procedure significantly reduces the size of the network, and often removes important components that may be needed for the study. We wish to relax or eliminate this procedure, in order to analyze larger networks.

2.2 Our iterative subnetwork component analysis

Recall the matrices $E \in \mathbb{R}^{n\times m}$,$A \in \mathbb{R}^{n\times l}$, and $P \in \mathbb{R}^{l\times m}$ from (2). We first divide the initial network into two subnetworks i, $(i \in 1,2)$, sharing common genes (TGs) and/or regulators (TFs). Let the subscripts u and c represent the exclusive (unique) and joint (common) partition components of each subnetwork i, respectively.

We distinguish between genes that are exclusively in partition i and are regulated by TFs in the exclusive partition j $(j \in 1,2,\ A_{u_iu_j},$ Fig. 1), genes that are exclusively in subnetwork i and are regulated by TFs in the joint partition (A_{u_ic}), genes that are in the joint partition c, but regulated by TFs in the exclusive partition i (A_{cu_i}), and genes that are in the joint partition, and regulated by the TFs in the joint partition (A_{cc}).

Then the matrices E and A in equation 2 for each subnetwork i can be decomposed by the following:

$$E_1 = A_1P_1 = \begin{bmatrix} E_{u1} \\ E_c \end{bmatrix} = \begin{bmatrix} A_{u_1u_1} & A_{u_1c} \\ A_{cu_1} & A_{cc} \end{bmatrix} \begin{bmatrix} P_{u1} \\ P_c \end{bmatrix} \tag{4a}$$

$$E_2 = A_2P_2 = \begin{bmatrix} E_{u2} \\ E_c \end{bmatrix} = \begin{bmatrix} A_{u_2u_2} & A_{u_2c} \\ A_{cu_2} & A_{cc} \end{bmatrix} \begin{bmatrix} P_{u2} \\ P_c \end{bmatrix} \tag{4b}$$

where $E_{u_i} \in \mathbb{R}^{n_{u_i}\times m}$ and $E_c \in \mathbb{R}^{n_c\times m}$ denote the expression matrices of sub-networks $i = 1,2$, and $n_{u1} + n_{u2} + n_c = n$. $A_{u_iu_i} \in \mathbb{R}^{n_{u_i}\times l_{u_i}}$, A_{u_ic} and $A_{cu_i} \in \mathbb{R}^{n_c\times l_{u_i}}$, and $A_{cc} \in \mathbb{R}^{n_c\times l_c}$ denote the partition matrices of A of subnetwork i, and $l_{u_1} + l_{u_2} + l_c = l$. In all the following, when we write A_i, E_i or P_i, we refer to matrices of the entire subnetwork i, including both their unique and joint partitions.

To simplify the problem, we assume that genes exclusively in subnetwork i are not regulated by any TFs that is exclusively in the other subnetwork. The two subnetworks can be combined to one overall network presentation:

$$E = AP = \begin{bmatrix} E_{u_1} \\ E_c \\ E_{u_2} \end{bmatrix} = \begin{bmatrix} A_{u_1u_1} & A_{u_1c} & \mathbf{0}_2 \\ A_{cu_1} & A_{cc} & A_{cu_2} \\ \mathbf{0}_1 & A_{u_2c} & A_{u_2u_2} \end{bmatrix} \begin{bmatrix} P_{u_1} \\ P_c \\ P_{u_2} \end{bmatrix} \tag{5}$$

with the zero matrices $\mathbf{0}_1 \in \mathbb{R}^{n_{u_2}\times l_{u_1}}$ and $\mathbf{0}_2 \in \mathbb{R}^{n_{u_1}\times l_{u_2}}$. The simplification can be relaxed by replacing the zero matrices $\mathbf{0}_1$ and $\mathbf{0}_2$ in A with $A_{u_1u_2}$ and $A_{u_2u_1}$, respectively.

Example 1. ISNCA topology network matrix: Consider a network with 6 TGs and 3 TFs (Fig. 1). We can decompose A to its unique and joint partitions as:

$$A_{u_1u_1} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad A_{u_1c} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad A_{cu_1} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, A_{cc} = \begin{bmatrix} 0 \\ 1 \\ 1 \end{bmatrix}, \tag{6}$$

$$A_{cu_2} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}, \quad A_{u_2c} = \begin{bmatrix} 0 \end{bmatrix}, \quad A_{u_2u_2} = \begin{bmatrix} 1 \end{bmatrix}, \tag{7}$$

$$\mathbf{O}_1 = \begin{bmatrix} 0 \end{bmatrix}, \quad \mathbf{O}_2 = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{8}$$

and the overall network equation is