

Order-preserving factor discovery from misaligned data

Arnaud Tibau Puig^{*†1}, Ami Wiesel^{‡2}, Aimee Zaas^{§¶}, Geoffrey S. Ginsburg[§],
Gilles Fleury[†] and Alfred O. Hero III^{*}

^{*}Dpt. of Electrical Engineering, University of Michigan, USA. Email: {atibaup, hero}@umich.edu

[†]E3S - SUPELEC Systems Sciences / Signal Processing and Electronic Systems Department, Supélec, France

[‡]School of Computer Science and Engineering, The Hebrew University of Jerusalem, Israel.

[§]Institute for Genome Sciences & Policy, Duke University, USA

[¶]Division of Infectious Diseases and International Health, Dpt. of Medicine, Duke University School of Medicine, USA

Abstract—We present a factor analysis method that accounts for possible temporal misalignment of the factor loadings across the population of samples. Our main hypothesis is that the data contains a subset of variables with similar but delayed profiles obeying a consistent precedence ordering relationship. Our model is motivated by the difficulty of gene expression analysis across subjects who have common patterns of immune response but show different onset times after a uniform inoculation time of a viral pathogen. The proposed method is based on a linear model with additional degrees of freedom that account for each subject's inherent delays. We present an algorithm to fit this model in a totally unsupervised manner and demonstrate its effectiveness on extracting gene expression factors affecting host response using a flu-virus human challenge study dataset.

Index Terms—Parallel Factor Analysis, Dictionary Learning, Low-rank Matrix Approximation

I. INTRODUCTION

With the advent of high-throughput data collection techniques, low-dimensional representations have become an essential tool for pre-processing, interpreting or compressing high-dimensional data. They are widely used in a variety of domains including electrocardiogram [1], image [2] or sound [3] processing. In traditional matrix factorization, the data is modeled as a linear combination of a number of factors. Thus, given an $n \times p$ data matrix \mathbf{X} , we have:

$$\mathbf{X} = \mathbf{M}\mathbf{A} + \epsilon,$$

where \mathbf{M} is a $n \times f$ matrix of factors, \mathbf{A} is a $f \times p$ matrix of scores and ϵ is a small residual. In order to obtain a low dimensional factorization, it is typical to assume that the matrix \mathbf{A} is sparse and the number of factors is small: $f \ll \text{rank}(\mathbf{X})$.

In many situations, we observe not one but several matrices \mathbf{X}_i and there are physical grounds for believing that the \mathbf{X}_i 's share an underlying model. This happens for instance when the observations consist of different time-blocks of sound from the same music piece [3] or when the \mathbf{X}_i 's contain gene expression data from different individuals inoculated with the

same viral entity [4]. One way to model this situation would be to assume that the factors are constant across observations while enforcing some form of consistency on the loading matrices:

$$\mathbf{X}_i = \mathbf{M}\mathbf{A}_i + \epsilon_i. \quad (1)$$

This corresponds to the usual linear factor analysis model. Depending on the constraints imposed on \mathbf{M} , \mathbf{A}_i , different methods arise such as Principal Components Analysis (PCA) [5], sparse PCA [1], k-SVD [6], structured PCA [2] or Non-Negative Matrix Factorization (NNMF) [7].

However, it is sometimes impossible to model the data with fixed factors, even though some sort of invariance exists. An example of this situation arises for instance when trying to analyze the immune system response to a viral entity through gene expression data. Figure 1 shows the gene expression levels for a single gene and different observations, each one corresponding to a different subject. In this real data example, all persons experience the same sort of expression response after viral inoculation (called "upregulation") but the moment when upregulation occurs is clearly not the same. If we attempt to train model (1) on this data, we will not be able to fit all subjects accurately. A more sensible approach for the data in

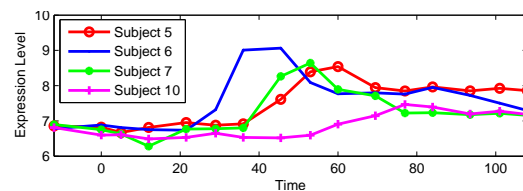


Fig. 1. Example of temporal misalignment of an upregulation motif across different subjects for gene *CCRL2*.

Figure 1 would be to fit each subject with a translated version of a common upregulation factor. This motivates the following class of models where the factors are allowed to vary across observations:

$$\mathbf{X}_i = \mathbf{M}_i\mathbf{A}_i + \epsilon_i. \quad (2)$$

By restricting our factors to be linear transformations of a common set of factors, we obtain a 3-way model which has been extensively considered in the signal processing and chemometrics literature [8], [9]. In this work we consider a restricted version of this model, where the constraints naturally arise from characteristic features of gene expression time

¹This work was supported in part by DARPA under the PHD program. The views, opinions, and findings contained in this article are those of the author and should not be interpreted as representing the official views or policies, either expressed or implied, of the Defense Advanced Research Projects Agency or the Department of Defense. Approved for Public Release, Distribution Unlimited.

²The work of A. Wiesel was supported by a Marie Curie Outgoing International Fellowship within the 7th European Community Framework Programme.

series. In particular, we restrict the columns of M_i to be circularly shifted versions of a common set of factors.

Our contributions are the following. First, we propose a constrained 3-way linear model that accounts for temporally misaligned factors. Second, we give a simple algorithm that allows us to fit this model in reasonable time. Finally, we demonstrate that our methodology is able to successfully extract the main features of a real dataset.

This paper is organized as follows. In Section 2 we first define our target structure and link it to a mathematical model. Second, we state the optimization problem associated to the fitting of our model and give a simple algorithm to find one of its local minima. In Section 4 we present the application of our methodology to a real gene expression dataset.

II. MODEL AND ALGORITHM

In this section, we first present our model hypotheses and translate them into a constrained factor model. Finally, we propose an efficient method for fitting it. Specifically, we consider the following hypotheses on the structure of gene expression data. Further motivation can be found in [10]:

- *H1: Motif consistency across subjects:* Gene expression patterns have consistent (though not-necessarily time aligned) motifs across all subjects.
- *H2: Motif sequence consistency across subjects:* If motif X precedes motif Y for subject i , the same precedence must hold for subject $j \neq i$.
- *H3: Motif consistency across groups of genes:* There are groups of genes that exhibit the same temporal expression patterns for a given subject.

Consider now the generative model in (2). Let F be a matrix whose columns are the f common *alignable* factors, and let $M(F, d)$ be a matrix valued function that applies a circular shift to each column of F according to the vector of parameters d , as depicted in Figure 2. Then, we have:

$$M_i = M(F, d^i). \quad (3)$$

In the context of gene expression response after viral inoculation, the columns in F are the set of signals emitted by the common immune system response and the vector $d^i \in \{0, \dots, n\}^f$ parameterizes each subject's incubation times. Using circular shifts introduces periodicity in our model. Some types of gene expression may display periodicity, e.g. circadian transcripts, while others, e.g. transient host response, may not. For transient gene expression profiles such as the ones we are interested in here, we use a truncated version of this periodic model (see [10]).

In order to enforce *H2*, the shifts d^i have to be such that the precedence order of motifs in a subject is preserved across all subjects. This can be achieved by ensuring that:

$$d_{j_1}^{s_1} \leq d_{j_2}^{s_1} \Leftrightarrow d_{j_1}^{s_2} \leq d_{j_2}^{s_2} \quad \forall s_2 \neq s_1, \quad (4)$$

that is, if factor j_1 precedes factor j_2 in subject s_1 , then the same ordering will hold in all other subjects. This can be simplified to constraining d^i to be an element of the set:

$$\mathcal{K} = \left\{ d \in \{0, \dots, n\}^f : d_{i+1} \geq d_i, \forall i \right\}. \quad (5)$$

On the other hand, each column of the matrix A_i describes the mixing weights of the factors that model a specific gene trajectory. By *H3*, we expect each factor to parsimoniously explain a large number of genes, the columns of A^i are to be sparse. By *H1*, the sparsity pattern is expected to be consistent across subjects, thus we will require A^i to have a group-sparse structure across different subjects.

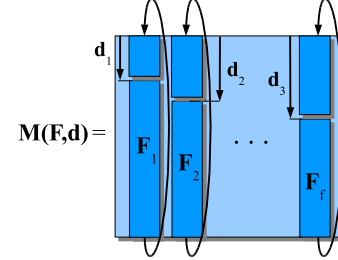


Fig. 2. Each subject's factor matrix M_i is obtained by applying a circular shift to a common set of factors F parameterized by a vector d^i .

We are now ready to formally state the factor analysis for misaligned time series. Given a number S (of possibly incomplete) $n \times p$ matrices X_i , our goal is to find S low dimensional approximations of the form:

$$X_i = M(F, d^i) A_i + \epsilon,$$

with d^i , F and A_i satisfying the assumptions stated above. In this paper we seek an interpretable model, that is, a model which effectively summarizes the features of the data while fitting it accurately. Our methodology can be also used as a dimensionality reduction pre-processing step previous to other analyses, such as clustering of gene expression time signatures [10].

We define the order-consistent dictionary learning problem as follows:

$$\begin{aligned} \min \quad & \sum_{i=1}^S \|X_i - M(F, d^i) A_i\|_F^2 + \\ & \lambda P_1(A_1, \dots, A_S) + \beta P_2(F) \quad (6) \\ \text{s.t.} \quad & \{d^i\} \in \mathcal{K}, F \in \mathcal{F}, A \in \{A_i\} \end{aligned}$$

where $\|\cdot\|_F$ is the Frobenius norm, $P_i(\cdot)$ are suitable penalty functions, λ, β are given tuning parameters. In our current implementation, the convex sets \mathcal{F}, \mathcal{A} restrict the variable space to the positive orthant (a key issue for interpretation, as in NNMF [7]) and \mathcal{K} is defined in (5). We use an l_1 Total Variation penalty on F to promote factors with few abrupt changes and a Group LASSO [11] penalty on the scores A_i in order to enforce consistency across subjects.

Problem (6) is a difficult, high dimensional, non convex optimization problem. A common approach for finding a local minima in such dictionary learning problems is to use block-coordinate descent, which alternates between the minimization with respect to the factors F and the scores A_i . Here we propose to solve (6) with the Block Coordinate Descent approach described in Algorithm 1, which iteratively minimizes (6) with respect to the shifts, the scores and the factors while keeping the other variables fixed. Since the objective in (6) is lower

bounded by 0 and its value decreases at each step, convergence of the algorithm is guaranteed.

Algorithm 1: BCD algorithm for finding a local minima of (6)

Input: Initial estimate of F and A_1, \dots, A_S .

Output: $F, A_1, \dots, A_S, d^1, \dots, d^S$

while Not Convergence do

$[d^1, \dots, d^S] \leftarrow \text{EstimateDelays}(F, A_1, \dots, A_S)$
 $[A_1, \dots, A_S] \leftarrow \text{EstimateScores}(F, d^1, \dots, d^S)$
 $[F] \leftarrow \text{EstimateFactors}(A, d^1, \dots, d^S)$

EstimateFactors and EstimateScores are convex penalized regression problems which can be efficiently solved. In our current implementation we use a constrained iterative thresholding algorithm [12], more details are given in [10]. EstimateDelays is trickier because the optimization domain is discrete. An efficient branch-and-bound approach is used to find the global solution to EstimateDelays, see Appendix A for details.

III. GENE EXPRESSION DATA ANALYSIS

In this section we apply our methodology to the study of an influenza A H3N2Wisconsin challenge study with multiple sampling time points of each subject for gene expression as part of the (DARPA) Predicting Health and Disease program [4]. This dataset consists of a collection of 272 microarray samples (of dimension 12023 genes) from 17 individuals. All of these subjects were inoculated with influenza A H3N2Wisconsin and $n = 16$ blood samples were extracted before and after inoculation at prespecified time points. Finally, the clinicians on the team established which of these subjects developed symptoms, assigning a binary label (Symptomatic (Sx) and Asymptomatic (Asx)) based on a standardized symptom scoring method. The time point when the strongest symptoms occurred (peak symptom onset time) was also recorded. For more details on the PHD challenge study experiment see [4].

In this study we demonstrate how our method is able to accurately reconstruct the observed gene expression trajectories with only 3 factors. We will also show that the estimated subject time delays are consistent with the recorded peak symptom times. Our analysis was performed over $p = 300$ significantly time-varying genes selected by Analysis of Variance. We apply our methodology to the $S = 9$ symptomatic subjects in the study. Choosing the number to be $f = 3$ reduces the chances of overfitting and is motivated by the fact that most genes show either a steady, an upregulation or a downregulation response. To avoid wrap-around effects, we work with a periodical model of dimension $n_F = 30$, which we truncate to fit the dimension $n = 16$ of the data (see [10] for more details). We compute our fit over a 5×5 grid of tuning parameters (λ, β) and use a heuristic to select the best pair.

To illustrate the goodness-of-fit of our model, we plot in Figure 3 the observed gene expression patterns of 9 strongly varying genes and compare them to the fitted response for three of the subjects, together with the relative approximation error. The average relative error is below 4% for all the subjects. It is clear that the gene trajectories have been smoothed while conserving their temporal alignment.

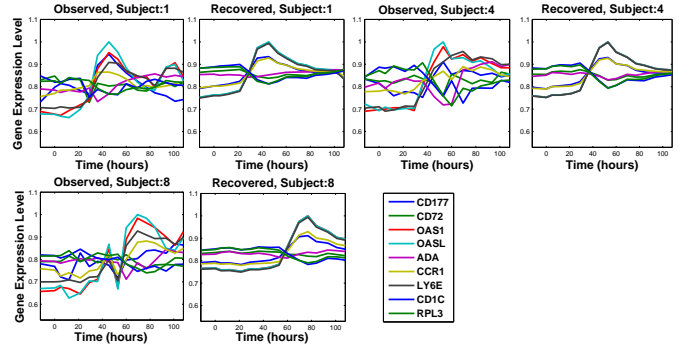


Fig. 3. Comparison of observed and fitted responses for three of the subjects and a subset of genes (named $CD177, CD72, OAS1, OASL, ADA, CCR1, LY6E, CD1C, RPL3$) which show strong temporal signatures. All subjects are reconstructed with a relative error below 4%.

The three factors obtained are shown in the middle plot of Figure 4. Factor 1 is clearly associated to sustained upregulation, factor 2 to sustained downregulation and factor 3 to a downregulation peak.

The top plot of Figure 4 shows the occurrence time of the chosen feature (e.g. for factor 1, the time when the gene expression changes from low to high) for each aligned factor together with the peak symptom onset time determined by clinical criteria. It is clear that the upregulation pattern of the first factor occurs a few hours before the onset peak time. This is consistent with the results reported in [13], where the upregulation peak was observed 36 hours before peak symptom time. Interestingly, the downregulation motifs associated with factor 2 and 3 consistently precede this upregulation motif.

In order to identify groups of genes showing similar expression signatures, we perform hierarchical clustering on the fitted scores $\{A_1, \dots, A_S\}$ using a standardized euclidean distance with the median as a linkage function. Four well separated clusters are obtained. From the bottom plot in Figure 4, it is clear that factor 1 is strongly associated to Cluster 4, whereas factor 3 corresponds to cluster 2 and 3. The expression signatures of this clusters are in very good agreement with the clusters previously found in [13] using different techniques. More details on these analyses are available in [10].

IV. CONCLUSIONS AND FUTURE WORK

We have proposed a method of precedence-order structured dictionary learning that accounts for possible temporal misalignments in a population of subjects undergoing a common treatment. We have described a simple model based on circular-shift translations of prototype motifs and have shown that the new factor analysis method can be a powerful tool for the analysis of a large gene expression temporal dataset.

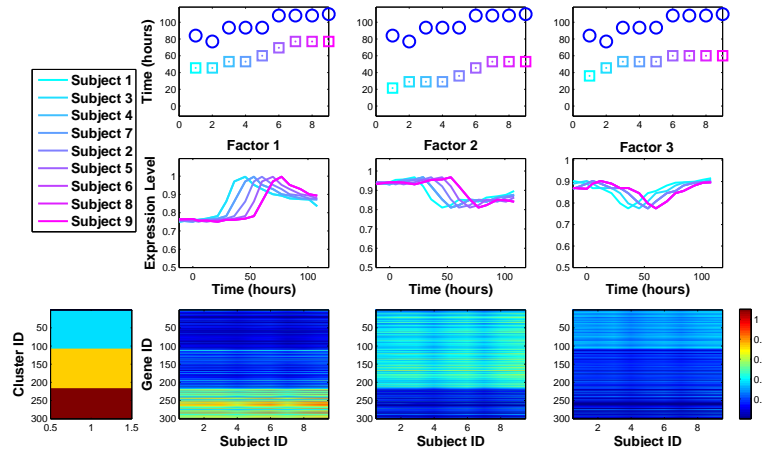


Fig. 4. Top Plot: Motif occurrence time for each factor (\square) and peak symptom time reported by clinicians (\circ). Middle Plot: Aligned factors for each subject. Factor 1 can be interpreted as an upregulation pattern, factor 2 as a persistent downregulation motif and factor 3 as a small downregulation peak. Bottom: scores corresponding to each gene for each of the 3 factors. Clearly, the first 2 factors account for most of the variance and the majority of the genes are strongly associated to only one of them.

APPENDIX

A. Solving EstimateDelays

EstimateDelays requires solving S uncoupled problems of the form:

$$\min_{\mathbf{d} \in \mathcal{K}} \|\mathbf{X} - \mathbf{M}(\mathbf{F}, \mathbf{d})\mathbf{A}\|_F^2, \quad (7)$$

where $\|\cdot\|_F$ is the Frobenius norm and the set \mathcal{K} is defined in (5). We use a branch-and-bound [14] approach to solve (7). For this purpose, we recursively branch (split) the set \mathcal{K} into two subsets. Consider the following standard decomposition of the set \mathcal{K} into two subsets $\mathcal{K} = \mathcal{I}_1 \cup \mathcal{I}_2$, where:

$$\mathcal{I}_1 := \{\mathbf{d} \in \mathcal{K} : \mathbf{d}_j \leq \gamma\} \quad \mathcal{I}_2 := \cup \{\mathbf{d} \in \mathcal{K} : \mathbf{d}_j \geq \gamma\} \quad (8)$$

which holds for any $1 \leq j \leq f$, $0 \leq \gamma \leq n-1$. The same splitting procedure can be subsequently applied to \mathcal{I}_1 , \mathcal{I}_2 and its resulting subsets. Upon application of this decomposition k times, the resulting subsets will be of the form:

$$\mathcal{I}_t := \{\mathbf{d} \in \mathcal{K} : \cap_{i=1}^k \mathbf{d}_{j_i} \leq (\text{or } \geq) \gamma_i\}.$$

Next, each subproblem (7) constrained to a set \mathcal{I}_t can be bounded as follows. First, we denote by $g(\mathbf{d})$ the objective function in (7) and define $g_{\min}(\mathcal{I}_t) := \min_{\mathbf{d} \in \mathcal{I}_t} g(\mathbf{d})$. We can relax the coupling induced by \mathcal{K} to obtain the following relaxation:

$$\mathcal{R}_t := \left\{ \mathbf{d} \in \{0, \dots, n\}^f : \cap_{i=1}^k \mathbf{d}_{j_i} \leq (\text{or } \geq) \gamma_i \right\} \supseteq \mathcal{I}_t \quad (9)$$

Letting $\mathbf{X}^{\parallel} = \mathbf{X}\mathbf{A}^{\dagger}\mathbf{A}$ and $\mathbf{X}^{\perp} = \mathbf{X}(\mathbf{I} - \mathbf{A}^{\dagger}\mathbf{A})$, it can be shown [10] that:

$$\underline{g}(\mathbf{d}) := \underline{\lambda}(\mathbf{A}\mathbf{A}') \|\mathbf{X}\mathbf{A}^{\dagger} - \mathbf{M}(\mathbf{F}, \mathbf{d})\|_F^2 + \|\mathbf{X}^{\perp}\|_F^2 \leq g(\mathbf{d}),$$

where $\underline{\lambda}(\mathbf{X})$ denotes the smallest eigenvalue of a symmetric matrix \mathbf{X} . Combining the relaxation in (9) with the last inequality, we obtain a lower bound on $g_{\min}(\mathcal{I}_t)$:

$$\Phi_{lb}(\mathcal{I}_t) := \min_{\mathbf{d} \in \mathcal{R}_t} \underline{g}(\mathbf{d}) \leq g_{\min}(\mathcal{I}_t),$$

which can be evaluated by performing f decoupled discrete grid searches. On the other hand, evaluating the objective at any point in the feasible subset \mathcal{I}_t gives an upper bound for $g_{\min}(\mathcal{I}_t)$:

$$g_{\min}(\mathcal{I}_t) \leq \Phi_{ub}(\mathcal{I}_t) := g(\mathbf{d}) \text{ for } \forall \mathbf{d} \in \mathcal{I}_t. \quad (10)$$

REFERENCES

- [1] I.M. Johnstone and A.Y. Lu, "On consistency and sparsity for principal components analysis in high dimensions," *Journal of the American Statistical Association*, vol. 104, no. 486, pp. 682–693, 2009.
- [2] R. Jenatton, G. Obozinski, and F. Bach, "Structured Sparse Principal Component Analysis," in *Proceedings of the 13th International Conference on Artificial Intelligence and Statistics (AISTATS) 2010*, 2010, vol. 9.
- [3] T. Blumensath and M. Davies, "Sparse and shift-invariant representations of music," *IEEE Transactions on Speech and Audio Processing*, vol. 14, no. 1, pp. 50, 2006.
- [4] A. K. Zaas et al., "Gene Expression Signatures Diagnose Influenza and Other Symptomatic Respiratory Viral Infections in Humans," *Cell Host and Microbe*, vol. 6, pp. 207–217, 2009.
- [5] K. Pearson, "LIII. On lines and planes of closest fit to systems of points in space," *Philosophical Magazine Series 6*, vol. 2, no. 11, pp. 559–572, 1901.
- [6] M. Aharon, M. Elad, and A. Bruckstein, "K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation," *IEEE Transactions on signal processing*, vol. 54, no. 11, pp. 4311, 2006.
- [7] D.D. Lee and H.S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature*, vol. 401, no. 6755, pp. 788–791, 1999.
- [8] J.D. Carroll and J.J. Chang, "Analysis of individual differences in multidimensional scaling via an N-way generalization of Eckart-Young decomposition," *Psychometrika*, vol. 35, no. 3, pp. 283–319, 1970.
- [9] T.G. Kolda and B.W. Bader, "Tensor decompositions and applications," *SIAM Review*, vol. 51, no. 3, pp. 455–500, 2009.
- [10] A. Tibau Puig, A. Wiesel, and A.O. Hero, "Order-preserving factor analysis," Tech. Rep., University of Michigan, June 2010.
- [11] M. Yuan and Y. Lin, "Model selection and estimation in regression with grouped variables," *Journal of the Royal Statistical Society Series B Statistical Methodology*, vol. 68, no. 1, pp. 49, 2006.
- [12] N. Pustelnik, C. Chaux, and J.C. Pesquet, "A constrained forward-backward algorithm for image recovery problems," in *Proc. EUSIPCO*, 2008, pp. 25–29.
- [13] Y. Huang et al., "Temporal Dynamics of Host Molecular Responses Differentiate Symptomatic and Asymptomatic Influenza A Infection," *Submitted*, 2010.
- [14] S. Boyd, A. Ghosh, and A. Magnani, "Branch and bound methods," *Notes for EE392o, Stanford University*, 2003.