# STRUCTURED SPARSE PCA TO IDENTIFY MIRNA CO-REGULATORY MODULES

*Shaogang Ren and Xiaoning Qian*

Dept. of Electrical & Computer Engineering, Texas A&M University

## ABSTRACT

This paper presents a new mathematical formulation and the corresponding algorithms for structured sparse principal component analysis (PCA). We introduce a new concept of support matrices with structured prior based on Markov Random Field (MRF). Both the support matrices and principal components are regularized by the $L_1$ norm to be integrated in a coupled objective function to recover the structured sparsity from the given data. Block coordinate descent and subgradient-based optimization methods are utilized to search for proper local minima for the formulated non-convex optimization problem. We implement the proposed methods to jointly analyze micro-RNA (miRNA) and gene interaction data to identify miRNA-gene co-regulatory modules (co-modules). Our preliminary experiments demonstrate that our structured sparse PCA has the potential to identify meaningful co-regulatory modules with enriched cellular functionalities.

***Index Terms***— Sparse Learning, Structured Sparse PCA, Variable Integration, Feature Clustering

## 1. INTRODUCTION

Principal component analysis (PCA) is a classical method to summarize observed data for more concise representation and dimension reduction [1]. One limitation of traditional PCA is that the principal components typically involve contributions from all originally observed variables, which may not be able to provide an easy interpretation of the aggregated information in these principal components. In order to have learned components with interpretable functional meanings, alternative methods have recently been proposed, such as notably nonnegative matrix factorization (NMF [2]) as well as several versions of sparse PCA (SPCA) with different assumptions and applications [3–5].

In addition to the efforts of deriving sparse principal components, in real-world applications, we also expect that contributing variables to principal components may have some relationships with each other in physical world, explicitly or implicitly. These "structured" relationships among contributing variables can help us better interpret data and provide new insights into the underlying processes. For example, in neuroimaging, we are interested in localizing effective areas in Positron Emission Tomography (PET) and functional Magnetic Resonance Imaging (fMRI) signals so that we can identify regions that are associated with different activity states or patterns in brain that may be associated with certain diseases such as Alzheimer's disease [6–9]. In computation biology and bioinformatics, it is commonly believed that the cellular functions arise from elaborate coordination between multiple bio-molecules. Deriving relevant information that capture the structured signals, for example from biological functional pathways or modules, may help better understand the underlying cellular mechanisms and provide accurate prognosis and diagnosis of potential phenotypic changes that may represent different disease states [10, 11].

In this paper, we propose a novel way to solve the structured sparse PCA problem. Instead of directly imposing both sparsity and structure constraints to contributing variables in principal components, we achieve the structured sparsity with the help of "support matrices" that we introduce to our new optimization formulation. Unlike the structured sparsity-including norm (SSIN) in [8], which imposes convexity structure constraints and may not be necessarily realistic in different real-world applications, we would like to have more flexible and general structure constraints. By introducing binary support matrices as auxiliary variables, we can impose the structured sparsity constraints by having $L_1$ norm regularization together with a Markov Random Field (MRF) smoothness constraint on support matrices. Our new MRF Structured Sparse PCA (MS$^2$PCA) is formulated as a coupled optimization problem to simultaneously solve for principal components and support matrices. Due to the flexibility of the MRF structure constraints, we expect that our new MS$^2$PCA method can derive more general structured sparse components which may have adaptive and flexible structures based on inherent relationships of observed variables. We evaluate the performance of MS$^2$PCA for the identification of miRNA-gene co-regulatory modules based on an integrated data set of miRNA-gene and gene-gene interactions as detailed in the following section.

## 2. CO-REGULATORY MODULE IDENTIFICATION

It has been recently conjectured that miRNAs may play critical regulatory roles in gene regulatory networks. However, the specific functionalities of many miRNAs and their combinatorial effects in cellular processes are still unclear [12]. By integrating diverse genomic data to identify the co-regulatory modules of miRNAs and genes (co-modules), we may derive

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

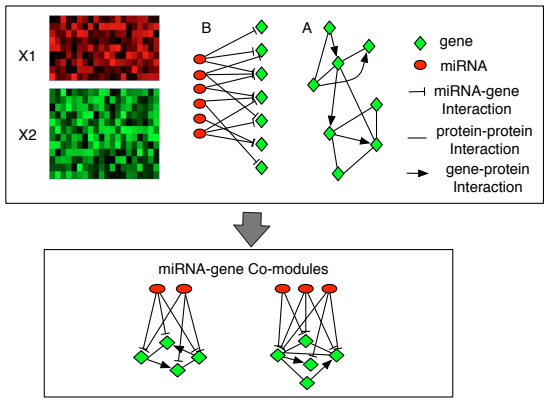


**Fig. 1**. Identification of miRNA-gene co-regulatory modules [12]. $X_1$ denotes miRNA expression data, $X_2$ is gene expression data; $A$ represents miRNA-gene interaction data, and $B$ is gene-gene interaction data.

better understanding on miRNAs and their functional roles in different cellular processes.

Figure 1 illustrates the co-module identification problem originally given in [12]. Assuming that we have both the miRNA ($X_1$) and gene expression data ($X_2$) together with the curated prior knowledge of the interactions among miRNAs and genes ($A$ and $B$), we would like to derive principal components that best explain the observed expression patterns. At the same time, we would like the derived principal components capture the functional relationships between miRNAs and genes, which can be enforced by imposing structured sparsity penalty based on the known interaction data in our MS$^2$PCA formulation. We note that our problem formulation and the optimization is different from the original work in [12], which relies on sparse network-regularized multiple nonnegative matrix factorization (SNMNMF) to figure out the underlying structure of the given data. The nonnegative requirement in SNMNMF may restrict the application of the method and introduce bias regarding the underlying structure. We will compare their performances using the same integrated data set.

## 3. MRF-BASED STRUCTURED SPARSE PCA

We now develop our MS$^2$PCA framework in a synthesis fashion [8] to find a set of factors or dictionary bases as principal components with the minimum reconstruction error for the given observed data. In this paper, the observed expression data is represented as an observation matrix $X = \begin{bmatrix} X_1 \\ X_2 \end{bmatrix} \in \mathbb{R}^{p\times n}$ with $n$ columns corresponding to $n$ samples. There are $p$ random variables denoting the total number of miRNAs and genes in the given data. We take $D$ to denote the "dictionary" or the set of $K$ principal components, which is a $p \times K$ matrix with $D^i_j$ denoting the matrix entry at the $ith$ row and $jth$ column. $D_k$ represents the $kth$ column vector of $D$, which corresponds to the $kth$ principal component. We introduce a binary support matrix $S \in \{0,1\}^{p\times K}$ and $P_S(D)$ represents the corresponding projection of the matrix $D$ onto the space of matrices supported by $S$:

$$P_S(D)^i_j = \begin{cases} 0 & \text{if } S^i_j = 0 \\ D^i_j & \text{if } S^i_j = 1 \end{cases}$$

$P_{S^\perp}$ is the complementary projection of $P_S$ with $P_S(D) + P_{S^\perp}(D) = D$ [13]. $||\cdot||_F$ is the matrix Frobenius norm, e.g. $||X||_F = \sqrt{\sum_{ij}(X^i_j)^2}$. We take $\|\cdot\|_2$ as the vector $L_2$ norm with $||X_k||_2 = \sqrt{\sum_i (X^i_k)^2}$ and $||X||_1 = \sum_{i,j}|X^i_j|$ denotes the $L_1$ norm. Finally, $sign(x)$ is the signum function, which takes on the sign of $x$ if $x$ is non-zero; Otherwise if $x = 0$, it takes any value in $[-1, 1]$. For a given matrix $X$, $sign(X)$ is a matrix in which each element takes the signum function value of the corresponding element in $X$. With the derived support matrices $S$ and principal components $D$, we can identify corresponding co-regulatory modules by assigning the co-module membership of miRNA (or gene) $i$ based on its corresponding z-scores in the derived $j$th principal components $D_j$ in $D$, which has been similarly done in [12].

### 3.1. Mathematical Formulation

Given the observed data $X$, our MS$^2$PCA finds the matrix $D \in \mathbb{R}^{p\times K}$ as principal components and the matrix $A \in \mathbb{R}^{K\times n}$ as principal component scores so that $DA^T$ can approximate $X$ as closely as possible. We formulate our MS$^2$PCA problem as the following optimization problem:

$$\min_{D,A,S} \quad ||X - P_S(D)A^T||^2_F + \lambda_3 \textstyle\sum_k ||D_k||_1 \qquad (1)$$

$$+\lambda_1 \textstyle\sum_k ||S_k||_1 + \lambda_2 \sum_k \sum_{(a,b)\in\Phi} |S^a_k - S^b_k| \qquad (2)$$

$$s.t. \quad ||A_k||_2 \le 1 \ \ \forall k \qquad (3)$$

We emphasize that we introduce a support matrix $S$ to bring in a more general structure prior. As given in (2), we include in the objective function a smoothness term for $S$, which is similar as in the Ising model and Markov Random Field models [13–15]. At the same time, each column $S_k$ is regularized by the $L_1$ norm to make sure that only a limited number of variables are contributing to the corresponding principal components. For principal components $D$, in addition to the $L_1$ norm regularization, the coupling with $S$ will render structured sparsity to its corresponding principal components: When $S^i_k = 0$, the previous formula implicitly forces $D^i_k \to 0$ due to its $L_1$ norm regularization for sparsity and the corresponding $ith$ variable will not contribute to the $kth$ principal component as it will not help reduce the reconstruction error after the projection. In fact, we generally can model all $p$ observed variables by a graph $G = (V, \Phi)$, in which $V$ is the set of vertices corresponding to $p$ variables that may have potential contributions to each principal component $D_k$, and $\Phi$ denotes the edges connecting all inherently

related variables, which provide the potential structure prior. In our co-module identification problem, we can derive the edge set $\Phi$ from both the miRNA-gene interaction network $A$ and gene-gene interaction network $B$ as shown in Figure 1. The second part of the objective function (2) is essentially the summation of the energy functions of $S_k$ corresponding to each component $D_k$ in the form of the Ising model [13, 15]:

$$\lambda_1 \sum_{i \in V} S_k^i + \lambda_2 \sum_{(a,b) \in \Phi} |S_k^a - S_k^b|, \tag{4}$$

in which the first term regularizes for parsimony to penalize large support regions and the second term controls the dependency or smoothness between related variables. With this formulation, we search for the most parsimonious (sparse) principal components that adaptively capture the structural relationships among variables.

Finally, there exist ambiguous solutions as the reconstruction error depends on the matrix product $DA^T$ and the optimal solutions of $D$ and $A$ is not unique. To solve this problem, we add the constraints (3) to penalize the $L_2$ norm of each column in $A$, $A_k$, which corresponds to principal component scores for $D_k$. This avoids degenerated solutions with either too large or too small values in some $A_k$ during optimization.

### 3.2. Optimization for MS$^2$PCA

Our proposed formulation contains integer variables $S$, which make it a non-convex mixed nonlinear optimization problem. In our algorithm, we compute $A$, $D$ and $S$ in an alternative updating fashion as follows.

**1. Update $A$:** With fixed $D$ and $S$, we require to solve the following optimization problem to update $A$:

$$\min_A ||X - \widetilde{D}A^T||_F^2 \tag{5}$$

$$s.t.\ ||A_k||_2 \leq 1\ \ \forall k; \tag{6}$$

in which $\widetilde{D} = P_S(D)$ in (1). Similar to [8], we iteratively update each $A_k$ with fixed $\{A_j\}_{j \neq k}$ by Block Coordinate Descent (BCD).

**2. Update $D$:** With $A$ and $S$ , we now solve the following optimization to update $D$:

$$\min_D ||X - P_S(D)A^T||_F^2 + \lambda_3 ||D||_1. \tag{7}$$

As the first term complicates the optimization by combinatorial contributions from $D$, there is no direct method to solve this problem. We solve it approximately by rewriting the objective function as

$$\min_D \{L = ||\widetilde{X} - DA^T||_F^2 + \lambda_3 ||D||_1\}, \tag{8}$$

in which $\widetilde{X} = X + P_{S^\perp}(D^{t-1})A^T$, and $D^{t-1}$ is from previous step $t-1$. This indeed resembles the formulation of the LASSO regression problem [16] and we adopt the "shooting algorithm" [16] to solve this subproblem.

**3. Update $S$:** When we fix $A$ and $D$, we expand the objective formulation for each column $S_k$ of the support matrix $S$ to get the following optimization problem:

$$\begin{aligned}\min_{S_k} \sum_i \{&(D_k^i \sum_j 2X_j^i A_k^j)(1 - S_k^i) + S_k^i(\lambda_1 \\ &+ (D_k^i)^2 \sum_j (A_k^j)^2 + D_k^i \sum_j 2A_k^j \sum_{m:m \neq k} S_m^i D_m^i A_m^j)\} \\ &+ \lambda_2 \sum_{(a,b) \in \Phi} |S_k^a - S_k^b| + C,\end{aligned} \tag{9}$$

where $C$ is a constant. The energy function (9) is a standard form of solving the first-order MRF with binary variables, which can be solved using graph-cut algorithms [13, 17, 18]. At each iteration, we update $S_k$ one by one until $S$ does not change.

Our MS$^2$PCA algorithm is summarized in Algorithm 1.

**Algorithm 1** MRF Structured Sparse PCA

**Input:** Number of principal components $K$; data matrix $X$
**Initialization:** Initialize $D$, and $A$ as random matrices and S as full matrix.
**repeat**
  **repeat**
    **for** k= 1 to K **do**
      Update $A_k$
    **end for**
  **until** $A$ does not change
  **repeat**
    Update $D$ with subgradient-based optimization
    Update $S$ with MRF labeling algorithm
  **until** $D$ does not change
**until** $D$ does not change

## 4. EXPERIMENTAL RESULTS

We implement our MS$^2$PCA to identify miRNA-gene co-regulatory modules based on the data set provide by [12], which has 385 ovarian cancer samples from the TCGA data portal (http:// canergenome.nih.gov/). After pre-processing, there are 559 miRNAs and 12,456 genes in the expression data corresponding to $X_1$ and $X_2$ in Figure 1. For the interaction prior, there are 31,949 gene-gene interaction pairs by combining the protein-protein interaction data and DNA-protein interaction data ($B$ in Figure 1) and 243,331 miRNA-gene interactions from MicroCosm website($A$ in Figure 1). To fairly compare MS$^2$PCA and SNMNMF, we set the co-module number to 50 to reproduce the results for SNMNMF with all of the parameters set to the best performing values reported in [12]. [1] For MS$^2$PCA, we set the miRNA-gene

[1] We note that due to the updated online data sets and miRNA label conversion, our obtained results from SNMNMF are different from the reported results in the original paper.

**Table 1**. Performance comparison on module identification

| Method | SNMNMF | MS$^2$PCA |
|---|---|---|
| Fitting Error | 0.4080 | 0.3653 |
| Time Cost (Sec.) | 151000 | 1977 |
| MiRNA Clusters (q-value $< 0.05$) | 7 | 10 |
| Gene Func. Sets (q-value $< 0.05$) | 14 | 22 |
| Key co-modules | 2 | 4 |

**Table 2**. Key co-modules from MS$^2$PCA

| Module ID | GO Term | q-value | q-value (miRNA) |
|---|---|---|---|
| 1 | GO:0030593 | 2.88e-07 | 2.77e-04 |
| 18 | GO:0022617 | 7.77-06 | 0.0105 |
| | GO:0030574 | 5.57e-05 | |
| 34 | GO:0016339 | 3.47e-06 | 0.0095 |
| 39 | GO:0007389 | 0.0123 | 0.0189 |

interaction pair weight with 50 and gene-gene interaction with 1. For tuning parameters, we empirically set $\lambda_1 = 5$, $\lambda_2 = 5$, and $\lambda_3 = 30$ so that 50 co-modules are identified with 4.8 miRNAs and 77.4 genes on average in each co-module, similar as SNMNMF (3.8 miRNA and 78 genes). The fitting accuracy, co-module properties and time cost for both methods are given in Table 1. As shown in the table, our method can achieve better fitting accuracy, better co-module properties and less computation cost as well.

MS$^2$PCA has identified more statistically significantly enriched co-modules with respect to both miRNA clusters and Gene Ontology (GO) biological process (BP) terms (http://www.geneontology.org/). We first evaluate whether co-modules can shed light on the combinational regulation from miRNAs by evaluating miRNAs in each identified co-module based on the genomic distance miRNA clusters obtained from the miRBase database as similarly done in [12]. In total, miRBase provides 216 such miRNA clusters with the size ranging from 2 to 51 miRNAs. We check each identified co-module to see whether it is significantly enriched with miRNAs within this set of miRNA clusters by Fisher's exact test with the false discovery rate (FDR) correction [19]. The co-modules with computed q-values less than $0.05$ are considered as significantly enriched miRNA clusters. Table 1 shows that MS$^2$PCA can discover more enriched miRNA clusters than SNMNMF. To further investigate whether the identified co-modules are enriched with cellular functionalities, we filter out 4,360 GO BP terms with more than 300 genes and fewer than 5 genes as in [12], which cover 12,700 genes in total. With the similar gene set enrichment analysis based on Fisher's exact test and FDR correction, MS$^2$PCA again identifies more enriched gene functional sets than SNMNMF with q-value threshold at 0.05. Finally, we find co-modules with both GO-term and miRNA-cluster enrichment q-values smaller than 0.05 and consider these modules as key regulatory co-modules that regulates cell functions. We again see the same trend that MS$^2$PCA performs better than SNMNMF. For example, Table 2 lists four key co-modules discovered by MS$^2$PCA. The identified co-modules 1 (GO:0030593) and 39 (GO:0007389) are statistically significantly enriched with corresponding miRNAs and genes that participate in cell chemotaxis bio-process and BMP signaling pathway, both of which are associated to ovarian cancer development. The identified miRNA cluster for co-module 18 includes miRNAs $\{$mir-143,mir-145$\}$ and both of them are related to ovarian cancer regulation [12]. Further study of these key co-modules may help better understand the regulatory roles of corresponding molecules in ovarian caner. Due to the page limit, we will provide more detailed functional analysis of identified co-modules in our corresponding journal manuscript.

In summary, compared to SNMNMF, our MS$^2$PCA can reach a better fitting accuracy, indicating that the derived principal components better explain the expression data. MS$^2$PCA is more flexible as it does not force the non-negativeness of the expression data as done in SNMNMF so that it can derive more biologically meaningful co-modules evaluated by miRBase and GO terms. Furthermore, the flexibility of MRF structure prior also contributes to better recovering of the cellular regulatory structure. In addition, due to the non-negativeness constraints imposed in SNMNMF, the adopted multiplicative updating algorithm for optimization has a slow convergence rate and thus is computationally less efficient than MS$^2$PCA.

## 5. CONCLUSIONS AND FUTURE WORK

In this paper, we introduce a novel approach MS$^2$PCA for recovered structured sparse structure in the observed data by flexible modeling of arbitrary pair-wise relationships among variables. The proposed method is implemented to identify miRNA-gene co-regulatory modules and the preliminary performance comparison with the state-of-the-art method SNMNMF [12] has demonstrated that MS$^2$PCA can better explain the observed data and recovers more meaningful co-regulatory modules with lower computational cost. With the flexibly of the MRF structure prior, there are many potential applications for our MS$^2$PCA. Future work includes testing MS$^2$PCA model on other biomedical data in addition to -omic data, for example analyzing functional brain images such as fMRI. We aim to recover the functional or active regions for different mental activities. By further analysis and thereafter better understanding of functional structures, we hope that we may derive deeper insights into the underlying disease mechanisms for better disease prognosis and prevention.

**Acknowledgements** The project is partially supported by NSF Award MCB-1244068 and NIH/NIDDK 1R21DK092845.

# References


[1] I. T. Jolli, *Principal Component Analysis*, Springer, 2002.

[2] 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, October 1999.

[3] H. Zou, T. Hastie, and R. Tibshirani, "Sparse principal component analysis," *J. Comput. Graph. Stat*, vol. 15, no. 2, pp. 265–286, 2006.

[4] Ron Zass and Amnon Shashua, "Nonnegative sparse pca," in *In Neural Information Processing Systems*, 2007.

[5] D. M. Witten, R. Tibshirani, and T. Hastiei, "A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis," *Biostatistics*, vol. 10, no. 3, pp. 515, 2009.

[6] A. Gramfort and M. Kowalski, "Improvingm/eeg source localization with an inter-condition sparse prior," in *In IEEE International Symposium on Biomedical Imaging*, 2009.

[7] Z. J. Xiang, Y. T. Xi, U. Hasson, and P. J. Ramadge, "Boosting with spatial regularization," in *In Advances in Neural Information Processing Systems*, 2009.

[8] R. Jenatton, G. Obozinski, and F Bach, "Structured sparse principal component analysis," in *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2010.

[9] Shuai Huang, Jing Li a, Liang Sun, Jieping Ye, Adam Fleisher, Teresa Wu, Kewei Chen, and Eric Reiman, "Learning brain connectivity of alzheimer's disease by sparse inverse covariance estimation," *NeuroImage*, vol. 50, pp. 935–949, 2010.

[10] Amin Ahmadi Adl, Xiaoning Qian, Ping Xu, Kendra Vehik, and Jeffrey P. Krischer, "Feature ranking based on synergy networks to identify prognostic markers in dpt-1," in *GENSiPS proceeding*, 2012.

[11] Seyed Javad Sajjadi, Xiaoning Qian, and Bo Zeng, "Network-based methods to identify highly discriminating subsets of biomarkers," in *GENSiPS proceeding*, 2012.

[12] Shihua Zhang, Qingjiao Li, Juan Liu, Xianghong, and Jasmine Zhou, "A novel computational framework for simultaneous integration of multiple types of genomic data to identify microrna-gene regulatory modules," *Bioinformatics*, vol. 27, no. 13, pp. i401–i409, 2011.

[13] X. Zhou, C. Yang, and W. Yu, "Moving object detection by detecting contiguous outliers in the low-rank representation.," *IEEE Trans Pattern Anal Mach Intell*, 2012.

[14] S. Geman and D. Geman, "Stochastic relaxation, gibbs distributions, and the bayesian restoration of images," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 6, pp. 721–741, 1984.

[15] Stan Z. Li, *Markov Random Field Modeling in Image Analysis*, Springer, 2009.

[16] Wenjiang J. FU, "Penalized regressions: The bridge versus the lasso," *Journal of Computational and Graphical Statistics*, vol. 7, no. 3, pp. 397416, 1998.

[17] V. Kolmogorov and R. Zabih, "What energy functions can be minimizedvia graph cuts?," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 26, no. 2, pp. 147159, 2004.

[18] Y. Boykov, O. Veksler, and R. Zabih, "Fast approximate energy minimization via graph cuts," *IEEE Trans. Pattern Anal. Mach. Intell*, vol. 23, no. 11, pp. 12221239, 2001.

[19] J.D. Storey and R Tibshirani, "Statistical significance for genome-wide studies," *Proc. Natl Acad. Sci.*, vol. 100, pp. 94409445, 2003.