# Grouping of correlated feature vectors using treelets

**Jing Xiang**
Department of Machine Learning
Carnegie Mellon University
Pittsburgh, PA 15213
jingx@cs.cmu.edu

## Abstract

In many applications, features are highly correlated to one another and algorithms used to construct networks from the data such as Lasso are unable to distinguish between them. Specifically, many genes in microarray gene expression data are highly correlated and knowledge of the representative groups can be useful. By using treelets, a hierarchical tree and an orthonormal basis are constructed that reflect the internal structure of the data. Theoretically, a large sample consistency result is described, and for the situation where the covariance matrices have block structure, treelets perform better than standard principal component analysis with respect to convergence rates. For simulated data that consists of collinear variables, treelets is able to pick out the groups of variables. It is also able to distinguish between genes of different cell cycles from yeast data.

## 1 Introduction

A main challenge in systems biology is how to quantitatively model the topological and functional changes of biological networks over time. The objective is to understand the mechanisms of transcriptional regulation and signal transduction that control cell behaviour. Over the course of a cell's life cycle, the interactions between molecules in the cell are dynamic, and thus, molecular networks rewire over time instead of remaining static. Therefore, being able to infer multiple networks that change over time will allow us to better understand the evolution of important cellular mechanisms.

The availability of microarray data has benefitted the work of system biologists because they have access to gene expression data of a large number of genes simultaneously. Classical analysis of microarrays has focused on a single data set captured at one time point. However, in order to infer dynamic networks, we require multiple samples over time. This has been made possible by data sets such as the yeast expression data provided by Pramilla *et al.*[4] Currently time-varying dynamic Bayesian networks [5] use an $l_1$-regularized auto-regressive approach to learn a sequence of networks from time series gene expression data. However, it is often the case that groups of genes are correlated. Many of the guarantees for using the $l_1$ penalty, Lasso, are derived either under mutual incoherence or restricted eigenvalue conditions [1]. In addition, this causes problems in generating biologically sensible results. For instance, when considering picking the neighbours of a particular node, if many genes are strongly correlated amongst themselves, Lasso will pick one of them randomly. This means that the network will not capture the complete interaction.

This is the motivation for finding a method to group the genes that are strongly correlated. If these genes can be grouped effectively, then we can perform Lasso on the representative groups as opposed to the whole data set by constructing a new expression vector. Also, from a biological perspective, it is useful to know which genes are grouped together because some of them may be missing from the network generated by Lasso. In this study, we use treelets [3] to find the subsets of highly correlated genes. Treelets is an algorithm that produces not only groupings of the variables but also functions on the data. It constructs a multiscale orthonormal basis on a hierarchical tree. The authors describe it as a multi-resolution transform because it provides a set of functions that are defined on nested subspaces. The advantage of treelets for our purposes is that it returns a hierarchical structure for the variables. It should be noted however that treelets will not produce a vector that is representative of the groups. The construction of the vectors needs to be addressed as a separate problem.

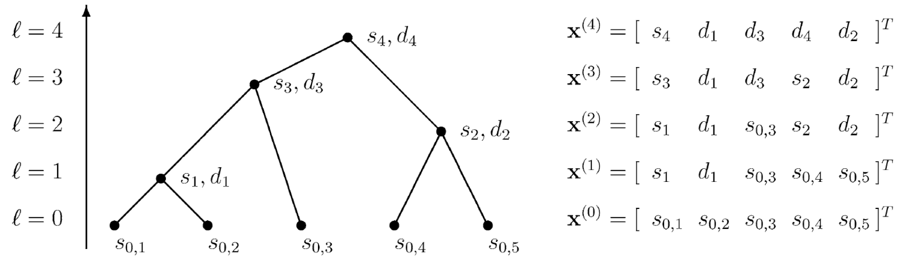
## 2 The Treelet Algorithm

The basic idea is that at each level of the tree, group together the most similar variables and substitute them by a coarser "sum variable" and a residual "difference variable". The new sum variables are computed by local principal component analysis (PCA) in two dimensions. The new variables then move on to further processing at higher levels of the tree while the remaining difference variables are stored. A brief outline of the treelet algorithm [3] is given below.

- Starting from the bottom of the tree, level *l* = 0, each observation **x** is represented by the original variables $\mathbf{x}^0$ = $[s_{0,1}$, ...., $s_{0,p}]^T$. Initialize the set of sum variables, S = 1,2,....,p. Note that at the bottom level, the sum variables are exactly the same as the original variables. No differencing has occurred yet. This is assigned the Dirac basis $B_0$ = [$\phi_{0,1}$,$\phi_{0,2}$... $\phi_{0,p}$] where $B_0$ is a p x p identity matrix.
- Calculate the sample covariance and similarity matrices $\hat{\Sigma}^0$ and $\hat{M}^0$ where the similarity between the two variables is simply the correlation coefficient.
- For *l* = 1...L, repeat:
  1. Find the two most similar sum variables $(\alpha, \beta)$ with respect to the similarity matrix $\hat{M}^0$.
  2. Do local PCA on the pair. This amounts to finding the Jacobi rotation matrix [2] that decorrelates $\mathbf{x}_\alpha$ and $\mathbf{x}_\beta$. Specifically, find J($\alpha$, $\beta$, $\theta_l$) such that $|\theta_l| \leq \pi/4$ and $\hat{\Sigma}^l_{\alpha\beta}$ = $\hat{\Sigma}^l_{\beta\alpha}$ = 0 where $\hat{\Sigma}^l$ = $J^T\hat{\Sigma}^{l-1}$J. The transformation means a change of basis $B_l = B_{l-1}$J and the new coordinates are $\mathbf{x}^l$ = $J^T\mathbf{x}^{l-1}$.
  3. If $\hat{\Sigma}^l_{\alpha\alpha} \geq \hat{\Sigma}^l_{\beta\beta}$ where $\alpha$ and $\beta$ are the first and second principal components, define the sum and difference variables as $s_l = \mathbf{x}^l_\alpha$ and $d_l = \mathbf{x}^l_\beta$. Define the scaling and detail functions $\psi_l$ and $\phi_l$ as columns $\alpha$ and $\beta$ of the basis matrix $B_l$.

     Thus for a given level *l*, we have the orthonormal treelet decomposition. This is illustrated in Figure 1.

$$\mathbf{x} = \sum_{i=1}^{p-l} s_{l,i}\phi_{l,i} + \sum_{i=1}^{l} d_i\psi_i \tag{1}$$



$\ell = 4$
$\ell = 3$
$\ell = 2$
$\ell = 1$
$\ell = 0$
$s_4, d_4$
$s_3, d_3$
$s_2, d_2$
$s_1, d_1$
$s_{0,1}$
$s_{0,2}$
$s_{0,3}$
$s_{0,4}$
$s_{0,5}$
$\mathbf{x}^{(4)} = [\ s_4\ \ d_1\ \ d_3\ \ d_4\ \ d_2\ ]^T$
$\mathbf{x}^{(3)} = [\ s_3\ \ d_1\ \ d_3\ \ s_2\ \ d_2\ ]^T$
$\mathbf{x}^{(2)} = [\ s_1\ \ d_1\ \ s_{0,3}\ \ s_2\ \ d_2\ ]^T$
$\mathbf{x}^{(1)} = [\ s_1\ \ d_1\ \ s_{0,3}\ \ s_{0,4}\ \ s_{0,5}\ ]^T$
$\mathbf{x}^{(0)} = [\ s_{0,1}\ \ s_{0,2}\ \ s_{0,3}\ \ s_{0,4}\ \ s_{0,5}\ ]^T$

Figure 1: This shows the treelet decomposition for data where dimension p = 5. Note that the bottom level l=0 is populated by the original p variables. At each successive level, the pair of most similar sum variables are combined and replaced by the sum and difference variables. Figure was taken from Lee *et al.*, 2008 [3].

# 3 Methods

To group the genes, I will be using treelets which produce a multi-scale orthogonal basis as well as a hierarchical cluster tree that reveals the internal structure of the data [3]. This will be tested on simulated data and selected genes from the yeast data set described below.

# 4 Description of Data

## 4.1 Simulated Data

I plan on simulating two data sets that contain groups of feature vectors that are correlated in different ways. Both procedures for simulation are outlined below.

First I generate groups of vectors. Within each group, the vectors are correlated, more specifically they are collinear. For simplicity and ease of visualization, I created 3 groups of 10 vectors. For each group g, a vector of size 1 x p was generated using the following algorithm. In this setting, both $a^g$ and $x_0$ are scalars.

For each group g:

1. Generate $a^g \sim$ Unif(-1,1)
2. Generate $x_0^g \sim \mathcal{N}(0,1)$
3. Burn-in: Repeat 1000 times, $x_0^g = a^g x_0^g + \epsilon$, $\epsilon \sim \mathcal{N}(0,1)$
4. Generate time series: $x_t^g = a^g x_{t-1}^g + \epsilon_t$

Then each representative vector from group g is then scaled by a constant to generate 10 different vectors. These vectors are linearly dependent so treelets should be able to group them without difficulty.

Another way of generating correlated data is by using the algorithm below. Here, A is a matrix and $x_0$ is a vector.

For each group g:

1. Generate $A^g \sim \mathcal{N}(0,1)$. In order to have a stationary time series, the eigenvalues of the matrix $A^g$ must be between -1 and 1. Thus we divide by the maximum eigenvalue to ensure that this holds.
2. Generate $x_0^g \sim \mathcal{N}(0,1)$
3. Burn-in: Repeat 1000 times, $x_0^g = A^g x_0^g + \epsilon$, $\epsilon \sim \mathcal{N}(0,1)$
4. Generate time series: $x_t^g = A^g x_{t-1}^g + \epsilon_t$

For each group, this produces a set of 10 vectors that are correlated but not collinear.

## 4.2 Yeast Data

The data set that will be used is microarray data collected across the cell cycle of budding yeast [4]. For testing, 15 correlated genes that are specific to each cell cycle were selected. The genes selected are specific to the G1 (growth 1), S (synthesis) and G2/M (growth 2 and mitosis) phases of the yeast cell cycle. Thus, the genes within each group are highly correlated and tend to evolve similarly over time. However, note that while the expression over time of genes are correlated, they are not collinear. The variance in the expression values over time is different for within-group genes as can be observed in Figure 2.

# 5 Summary of Theoretical Results

## 5.1 Large Sample Properties of the Treelet Transform

If certain conditions hold, then we can discuss the large sample properties of treelets. We'll denote $\Sigma$ as the covariance matrix and $\hat{\Sigma}$ as the sample covariance matrix. First, let's define $T(\Sigma) = J^T \Sigma J$ as

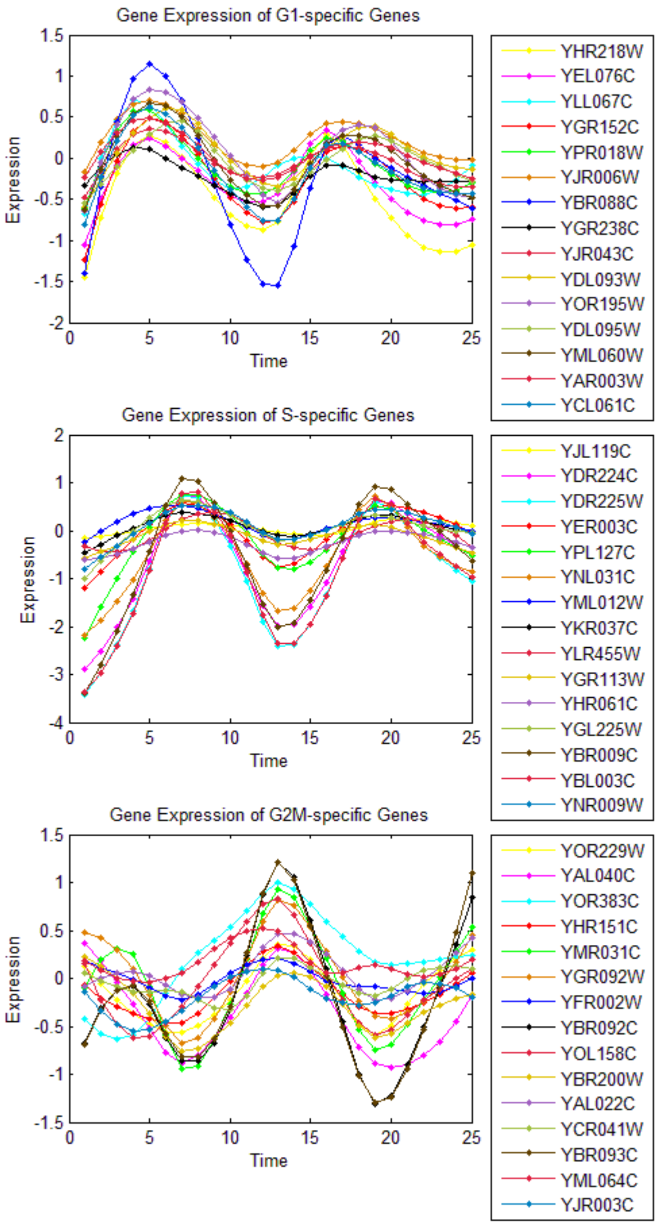


Figure 2: These figures show the expression data of three groups of cell-specific genes over time. From top to bottom, are G1-phase genes, S-phase genes, and G2/M-phase genes. The legend to the right of each plot show the open reading frame (ORF) IDs of each gene. It is important to notice that while the genes within the same group are correlated, they are not collinear.

the covariance matrix after one step of the treelet algorithm. Then it follows that $T^l(\Sigma)$ is the covariance matrix after $l$ steps. We'll define the infinity norm of some matrix A as $\|A\|_\infty = max_{j,k}|A_{jk}|$. Let

$$\mathcal{T}_n(\Sigma, \delta_n) = \bigcup_{\|\Lambda - \Sigma\|_\infty \leq \delta_n} T(\Lambda) \tag{2}$$

where $\Lambda$ is some covariance matrix close to $\Sigma$. In addition, $\mathcal{T}_n^1(\Sigma, \delta_n) = \mathcal{T}_n(\Sigma, \delta_n)$, and

$$\mathcal{T}_n^l(\Sigma, \delta_n) = \bigcup_{\Lambda \in \mathcal{T}_n^{l-1}} T(\Lambda), l \geq 2. \tag{3}$$

The equation above describes the set of covariance matrices at step $l$ of the treelet algorithm.

The following assumptions must hold for the large sample result:

A1) It must be the case the **x** has finite variance and satisfies one of the following conditions:

1. each $x_j$ is bounded,
2. x is multivariate normal, or
3. there exists $M$ and $s$ such that $E(|x_j x_k|^q) \leq q! M^{q-2} s/2$ for all $q \geq 2$.

A2) The dimension $p_n$ satisfies $p_n \leq n^c$ for some $c > 0$.

This brings us to Theorem 1 which states that if the conditions A1 and A2 hold, and we let $\delta_n = K\sqrt{log\ n/n}$ where $K > 2c$. Then as $n, p_n \to \infty$,

$$P(T^l(\hat{\Sigma}_n) \in \mathcal{T}_n^l(\Sigma, \delta_n), l = 1, ..., p_n) \to 1. \tag{4}$$

This result shows that $T^l(\hat{\Sigma})$ is not far from $T^l(\Lambda)$ for some $\Lambda$ close to $\Sigma$. It should be noted that while producing a result that says $\|T^l(\Sigma) - T^l(\hat{\Sigma})\|_\infty$ would be better, this is not possible since correlation coefficients are used to measure similarity.

### 5.2 Treelets for Covariance Matrices with Block Structures

The results in this section apply specifically to covariance matrices with block structures. This analysis is useful since many real life data sets including gene expression data exhibit approximate block structures. In these situations, treelets provide a sparse representation when there are inherent block structures.

#### 5.2.1 Exact Analysis in the limit as n $\to \infty$

Here, we consider an ideal situation where the variables within the same group are collinear and those from different groups are only weakly correlated. The calculations are exact and computed in the limit. The main results state that if you have a K × K block covariance matrix with added white noise, and the variables from different blocks are weakly correlated, then the K maximum variance scaling functions are constant on each block. This is true under certain conditions that describe the noise level and within-block and between-block correlations of the data. This is specified formally as Theorem 2 below [3].

THEOREM 2. Assume that **x** = $(x_1, x_2, ...., x_p)^T$ is a random vector with distribution F, mean 0 and covariance matrix $\Sigma = C + \sigma^2 I_p$ where $\sigma^2$ represents the variance of white noise in each variable. $I_p$ is a $p \times p$ identity matrix and $1_{p_i \times p_j}$ is a $p_i \times p_j$ matrix with all entries equal to 1.

$$\begin{pmatrix} C_{11} & C_{12} & \dots & C_{1K} \\ C_{12} & C_{22} & \dots & C_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ C_{1K} & C_{2K} & \dots & C_{KK} \end{pmatrix}$$

$C$ is a K × K block matrix with within-block covariance matrices $C_{kk} = \sigma_k^2 1_{p_k \times p_k} (k = 1 \dots K)$ and between-block covariance matrices $C_{ij} = \sigma_{ij} 1_{p_i \times p_j}$ $(i, j = 1 \dots K; i \neq j)$. If

$$\max_{1 \leq i,j \leq K} \left( \frac{\sigma_{ij}}{\sigma_i \sigma_j} \right) < \frac{1}{\sqrt{1 + 3\max(\delta^2, \delta^4)}} \tag{5}$$

where $\delta = \frac{\sigma}{min_k \sigma_k}$, then the treelet decomposiition at level $l = p - K$ has the form

$$T^{p-K}(\Sigma) = \sum_{k=1}^{K} s_k \phi_k + \sum_{i=1}^{p-K} d_i \psi_i \tag{6}$$

where $s_k = \frac{1}{\sqrt{p_k}} \sum_{j \in \mathcal{B}_k} x_j$, $\phi_k = \frac{1}{\sqrt{p_k}} I_{\mathcal{B}_k}$, and $\mathcal{B}_k$ represents the set of indices of variables in block k $(1 \ldots K)$. The expansion coefficients have means $E[s_k] = E[d_i] = 0$ and variances $V[s_k] = p_k \sigma_k^2 + \sigma^2$ and $V[d_i] = O(\sigma^2)$ for $i = 1 \ldots p - K$.

Thus, if the conditions of the theorem are met, then all treelets associated with levels $l > p - K$ are constant on groups of similar variables. The key results for the full decomposition at maximum level $l = p - 1$ of the tree follows directly from Theorem 2.

If the conditions in Theorem 2 are satisfied, then a full treelet decomposition gives $T^{p-1}(\Sigma) = s\phi + \sum_{i=1}^{p-1} d_i \psi_i$ where the scaling function $\phi$ and the K-1 detail functions $\psi_{p-K+1} \ldots \psi_{p-1}$ are constant on each of the K blocks. The coefficients $s$ and $d_{p-K+1} \ldots d_{p-1}$ reflect between-block structures as opposed to the coefficents $d_1 \ldots d_{p-K}$ which only reflect the noise in the data with variances $V[d_i] = O(\sigma^2)$ for $i = 1 \ldots p - K$.

This means that K can be found, parameter-free, by finding the energy distribution of a full treelet decomposition. In addition, treelets can reveal the block structure even if it's hidden because of background noise variables.

### 5.2.2 Convergence Rates

An estimate of the sample size required for treelets to find the inherent structures of data can be given under certain assumptions. If we assume block covariance matrices, it can be shown that treelets discovers the correct groupings of variables if the sample size n $\gg$ O(log $p$) where $p$ is the dimension of the data. This is better than standard PCA which is consistent when n $\gg$ O($p$). Define a covariance matrix $\Sigma$ with K blocks. $A_{L,n}$ is the event that K maximum variance treelets at level $L = p - K$, for a data set with n observations are supported only on variables form the same block. This is the ideal case where the treelet algorithm finds the exact grouping of variables. Briefly stated, the result says that if the assumption A1 holds, then

$$P(A_{L,n}^C) \leq L c_1 p^2 e^{n c_2 t^2} \tag{7}$$

for some positive constants $c_1$ and $c_2$. If we require that $P(A_{L,n}^C) < \alpha$, then the sample size must satisfy the following:

$$n \geq \frac{1}{c_2 t^2} log \left( \frac{L c_1 p^2}{\alpha} \right) \tag{8}$$

# 6 Results

## 6.1 Simulation Experiments

If the data is simulated such that groups of collinear vectors are produced, treelets perform fairly well. As you can see from Figure 3, the hierarchical tree plot on the left splits into 3 groups, each containing 10 components. This is exactly how the simulated data was constructed. In addition, the coloured treelets on the right also contain groups of 10.

However, if the data is simulated such that the groups that are correlated but not collinear, then treelets have a difficult time finding the groups of correlated components. From the hierarchical tree on the left (Fig. 4), it shows that the components have been divided into groups of 8, 4 and 18. In addition, from the plot on the right, it seems that group T3 is composed of data samples from all three groups.

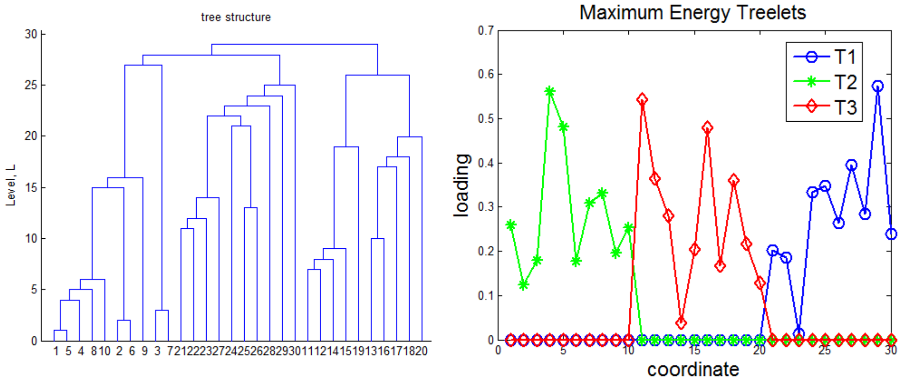


Figure 3: Treelets were used to group vectors that belonged to three different groups. Vectors within the same group were collinear. As you can see from the hierarchical tree plot (left) and the treelets produced from the simulation (right), the algorithm finds the 3 groups of 10 vectors correctly. Note that the x-axis on the right hand plot corresponds to the numbering of the vectors. The first 10 belong to the first group, the second 10 to the second group and so on.

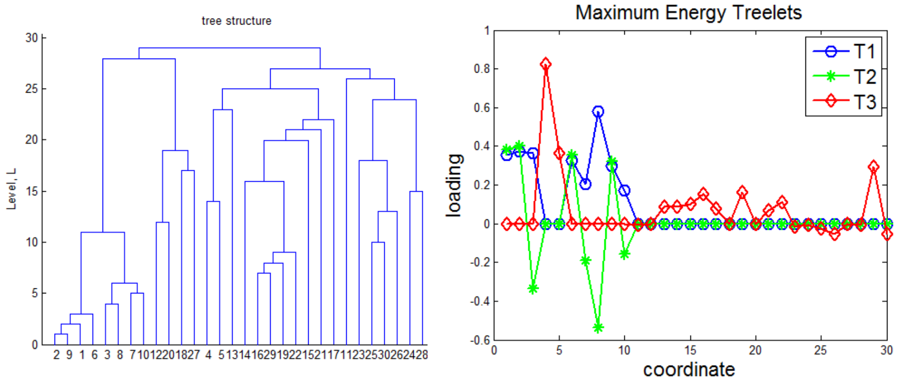


Figure 4: Again, treelets were used to group vectors that belonged to three different groups. However, the vectors within the same group were correlated but not collinear. Neither the hierarchical tree plot (left) or the treelets produced from the simulation (right) show the correct groupings.

### 6.2 Experiments on Yeast Data

Recall that a set of time series gene expression data of 45 genes that are specific to different cell cycles shown in Figure 2 was used for the testing of treelets. The G1-phase genes are numbered from 1-15, S from 16-30 and G2M from 31 to 45. From Figure 5, it is shown that the top three basis vectors are able to distinguish the different groups of genes even though the within-group genes are not collinear.

From the hierarchical tree (Fig. 6), it can be observed that the tree divides at the top-level into 4 groups. Ideally, it would divide into 3 groups. However, given that some of the gene expression curves of G1 and S are similar, it is conceivable that the hierarchical tree may not be perfect.

## 7 Conclusion

Treelets is a method that constructs a multiscale orthogonal basis and hierarchical clustering tree in order to uncover the internal structure of data. Because many real world data sets such as gene expression data are highly correlated, it can be used to group the correlated feature vectors. It has been demonstrated that treelets obtain the correct results when the feature vectors are collinear with some noise. However, as the simulation results indicate, if the data is correlated but not collinear,

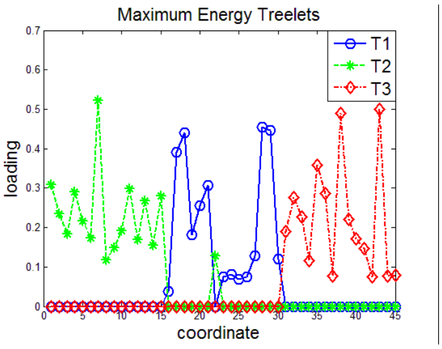


Figure 5: The maximum energy basis vectors produced by treelets on the yeast data. The three cell cycle specific groups of genes are clearly separated into their respective groups.

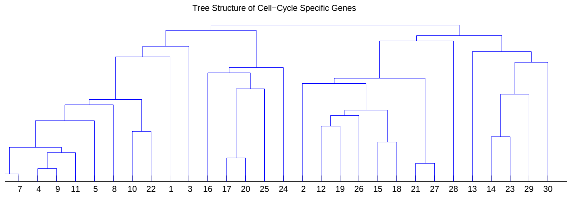


Figure 6: The hierarchical tree produced by treelets for the yeast data.

treelets is not always successful. But when using real gene expression data from yeast, treelets was able to group genes from different cell cycles successfully without them being collinear.

Currently, it is unclear what is required for variables within the same group to be classified by treelets as such. The theoretical results that discuss treelets on block structure covariance matrices assume the ideal case where variables within the same group are collinear. An interesting future direction is to more formally define what it means to be correlated in such a way that guarantees that treelets will be able to separate the various groups. This would be useful in determining whether treelets would be helpful for particular data sets.

## References

[1] P.J. Bickel, Y. Ritov and A. Tsybakov. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, **37**(4): 1705-1732.

[2] G. Golub and C.F. Van Loan. (1996). Matrix Computations, 3rd Edition. Baltimore: Johns Hopkins University Press.

[3] A.B. Lee, B. Nadler, and L. Wasserman. Treelets-An Adaptive Multi-scale basis for sparse unordered data. (2008). *The Annals of Applied Statistics*, **2**(2): 435-471.

[4] T. Pramila, W. Wu, S. Miles, W.S. Noble, and L.L. Breed. (2006). The Forkhead transcription factor Hcm1 regulates chromosome segregation genes and fills the S-phase gap in the transcriptional circuitry of the cell cycle. *Genes & Development*, **20**: 2266-2278.

[5] L. Song, M. Kolar and E.P. Xing. (2009). Time-Varying Dynamic Bayesian Networks. *Proceeding of the 23rd Neural Information Processing Systems.*

[6] A. Subramanian, P. Tamayo, V.K. Mootha, S. Mukherjee, B.L. Ebert, M.A. Gillette, A. Paulovich, S.L. Pomeroy, T.R. Golub, E.S. Lander, and J.P. Mesirov. Gene set enrichment analysis: A knowledge-based approach for interpreting genome-wide expression profiles. *Proceedings of the National Academy of Sciences*, **102**(43): 15545-15550.