Treelets - An Adaptive Multi-Scale Basis for Sparse Unordered Data

Author: Ann B. Lee, Boaz Nadler and Larry Wasserman
Group member: Rao Fu, Peisi Yan, Ye Shen, Sabayasachi Chatterjee, Yang Liu

1 Introduction

Treelet method is inspired by both hierarchical clustering trees and wavelets. Treelet method is kind of a step-wise local PCA method, which has an advantage over PCA, for PCA only captures global structure of data. There are two goals of treelet method: One goal is to develop sets of coordinate systems which could interpret internal structure of the data given. The second goal is to reduce data dimensions when we deal with situations like “small n, large p”.

2 Method

Treelet method focuses on doing local PCAs, which could also be described as “Jacobi rotations on pairs of similar variables”. The pairs of similar variables are chosen based on similarity measure $M_{ij}$, which is defined as the correlation between variables $s_i$ and $s_j$.

$$M_{ij} = \rho_{ij} = \frac{\Sigma_{ij}}{\sqrt{\Sigma_{ii} \Sigma_{jj}}},$$

where $\Sigma_{ij}$ is the covariance. In fact, $M_{ij}$ is not constrained as being coefficient. It could be any values which are reasonable to capture the similarity between variables.

Assume we have a given data matrix $x^{(0)} = [s_{0,1}, ..., s_{0,p}]^T$, where $s_{0,k} = x_k$. And the associated Dirac basis is $B_0 = [\phi_{0,1}, ..., \phi_{0,p}]$, where $B_0$ is the $p \times p$ identity matrix. Now based on the given data matrix, we can compute covariance and similarity matrices $\hat{\Sigma}^{(0)}$ and $\hat{M}^{(0)}$. Also, initialize the set of ”sum variables" $s = \{1, 2, ..., p\}$.

Repeat the following procedure for $l = 1, ..., L$, where $L \leq p - 1$. 

1
1. Find the pair of variables \((\alpha, \beta)\) that are most similar according to similarity matrix \(\hat{M}^{(l-1)}\).
\[
(\alpha, \beta) = \text{argmax}_{i,j \in \mathbf{s}} \hat{M}_{ij}^{(l-1)},
\]
where \(i < j\) and maximization is over variables that belong to "sum variables" \(s\).

2. Perform a local PCA on this pair. Find a Jacobi rotation matrix
\[
J(\alpha, \beta, \theta_l) = \begin{bmatrix}
1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\
0 & \cdots & c & \cdots & -s & \cdots & 0 \\
\vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\
0 & \cdots & \cdots & c & \cdots & 0 \\
\vdots & \vdots & \vdots & \ddots & \vdots & \vdots & \vdots \\
0 & \cdots & \cdots & 0 & \cdots & \cdots & 1
\end{bmatrix},
\]
where \(c = \cos(\theta_l)\) and \(s = \sin(\theta_l)\) are chosen to decorrelates \(x_\alpha\) and \(x_\beta\). So that \(\hat{\Sigma}_{\alpha\beta}^{(l)} = \hat{\Sigma}_{\beta\alpha}^{(l)} = 0\). This transformation is equal to a basis transformation \(B_l = B_{l-1}J\) and new coordinates \(x^{(l)} = J^T x^{(l-1)}\).

3. Assume after transformation, \(\hat{\Sigma}_{\alpha\alpha}^{(l)} \geq \hat{\Sigma}_{\beta\beta}^{(l)}\). Then update "sum variables" \(s = s \setminus \{\beta\}\).
Now at each level \(l\), we have an orthonormal treelet decomposition
\[
x = \sum_{i \in \mathbf{s}} s_{l,i} \phi_{l,i} + \sum_{1 \leq j \leq p, j \notin \mathbf{s}} d_{l,j} \psi_{l,j},
\]
where \(\phi_{l,i}\) and \(\psi_{l,j}\) are the \(i\)th and \(j\)th columns of basis \(B_l\), and \(s_{l,i}\) and \(d_{l,j}\) are the projections of \(x\) onto \(\phi_{l,i}\) and \(\psi_{l,j}\).

An illustrated graph is shown above. In this example, we have \(p = 5\). The left graph shows the building process of the tree. Each level, a sum variable and corresponding difference
variable are generated. In the right graph, corresponding coordinates transformation is shown, which is associated with an orthonormal basis in $R^p$ that captures the local structure of data.

3 “Best K-basis” method to decide tree height

For a given orthonormal basis $B = (w_1, ..., w_p)$, assign a normalized energy score $\xi$ to each vector $w_i$ to be

$$\xi(w_i) = \frac{E\{|w_i \cdot x|^2\}}{E\{|x|^2\}}.$$ 

Sample estimate is $\hat{\xi}(w_i) = \frac{\sum_{j=1}^{n}|w_i \cdot x_j|^2}{\sum_{j=1}^{n}|x_j|^2}$. We can sort the $p$ estimates from high to low. Define $\Gamma_K$ to be the sum of the largest $K$ estimates. Then the best $K$-basis is the treelet basis with the highest score

$$B_L = \arg\max_{0 \leq l \leq p-1} \Gamma_K(B_l).$$

Here, we assume $K$ is some value pre-determined. For real data, cross-validation method is used to decide the tree height, which means part of the given data is used to build the tree and the rest data are used to estimate scores in case of overfitting.

4 One Theory

Let $T(\Sigma) = J^T \Sigma J$ denote the covariance matrix after one step of treelet algorithm when starting with covariance matrix $\Sigma$. Let $T^l(\Sigma)$ denote the covariance matrix after $l$ steps of the treelet algorithm. Define $||A||_\infty = \max_{j,k} |A_{j,k}|$ and let

$$T_n(\Sigma, \delta_n) = \bigcup_{||\Lambda - \Sigma||_\infty \leq \delta_n} T(\Lambda).$$

Define $T^1_n(\Sigma, \delta_n) = T_n(\Sigma, \delta_n)$, and

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

Let $\hat{\Sigma}_n$ denote the sample covariance matrix.

Assumption(A1): Assume that $x$ finite variance and satisfies one of the following three assumptions: (a) each $x_j$ is bounded or (b) $x$ is multivariate normal or (c) there exist $M$ and $s$ such that $E(|x_j x_k|^q) \leq q! M^{q-2}s/2$ for all $q \geq 2$.

Assumption(A2): The dimension $p_n$ satisfies $p_n \leq n^c$ for some $c \geq 0$. 

3
Theorem Suppose that (A1) and (A2) hold. Let \( \delta_n = K \sqrt{\log n / n} \), where \( K \geq 2c \). Then, as \( n, p_n \to \infty \),

\[
P(T^l(\hat{\Sigma}_n) \in T^l_n(\Sigma, \delta_n), l = 1, \ldots, p_n) \to 1.
\]