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

Ann B Lee, Boaz Nadler, and Larry Wasserman

Department of Statistics (ABL, LW)
Carnegie Mellon University
Pittsburgh, PA

Department of Computer Science and Applied Mathematics (BN)
Weizmann Institute of Science
Rehovot, Israel

Running Title: Treelets

Address for correspondence and reprints:
Ann B Lee, Department of Statistics, Carnegie Mellon University, 5000 Forbes Avenue, Pittsburgh, PA 15213. E-mail: annlee@stat.cmu.edu
Abstract

In many modern applications, including analysis of gene expression and text documents, the data are noisy, high-dimensional, and unordered — with no particular meaning to the given order of the variables. Yet, successful learning is often possible due to sparsity; the fact that the data are typically redundant with underlying structures that can be represented by only a few features. In this paper we present treelets — a novel construction of multi-scale orthonormal bases that extends wavelets to non-smooth signals. Treelets capture the internal structure of the data and can as a dimensionality reduction tool significantly improve inference and prediction. We examine a variety of situations where treelets outperform principal component analysis and some common variable selection methods. The proposed method is illustrated on a linear mixture model, and on two real data sets: internet advertisements and DNA microarray data.

1 Introduction

For many modern data sets (e.g. DNA microarrays, financial and consumer data, text documents and internet web pages), the collected data are high-dimensional, noisy, and unordered, with no particular meaning to the given order of the variables. In this paper, we introduce a new methodology for the analysis of such data. We describe the theoretical properties of the method, and illustrate the proposed algorithm on a data set of internet advertisements, and a DNA microarray data set. These two data sets both contain structure in the form of complex groupings of highly correlated variables. The internet data include more than a thousand binary variables (various features of an image) and a couple of thousand observations (an image in an internet page). Some of the variables are exactly linearly related, while others are similar in more subtle ways. The DNA microarray data include the expression levels of several thousand genes but less than 100 samples (patients). Many sets of genes exhibit similar expression patterns across samples. The task in both cases is here classification. The results can therefore easily be compared with those of other clas-
sification algorithms. There is however a deeper underlying question which motivated our work: Is there a simple general methodology that, by construction, captures intrinsic localized structures, and that as a consequence improves inference and prediction of noisy, high-dimensional data when sample sizes are small? The method should be powerful enough to describe complex structures on multiple scales for unordered data, yet be simple enough to understand and analyze theoretically. Below we give some more background to this problem.

The key property that allows successful inference and prediction in mentioned high-dimensional settings is the notion of sparsity. Generally speaking, there are two main notions of sparsity. The first is sparsity of various quantities related either to the learning problem at hand or to the representation of the data in the original given variables. Examples include a sparse regression or classification vector (Tibshirani, 1996; Buckheit and Donoho, 1995), and a sparse structure to the covariance or inverse covariance matrix of the given variables (Bickel and Levina, 2007). The second notion is sparsity of the data itself. Here we are referring to a situation where the data, despite its apparent high dimensionality, is in fact highly redundant with underlying structures that can be represented by only a few features. Examples include data where many variables are approximately collinear or highly related, possibly in a complex multi-scale fashion, and data that lie on a non-linear manifold (Belkin and Niyogi, 2005; Coifman et al., 2005). While the two notions of sparsity are different, they are clearly related. In fact, a low intrinsic dimensionality of the data typically implies, for example, sparse regression or classification vectors as well as low-rank covariance matrices. However, this relation may not be directly transparent, as the sparsity of these quantities sometimes becomes evident only in a different basis representation of the data (Johnstone and Lu, 2004).

In either case, to take advantage of sparsity, one constrains the set of possible parameters of the problem. For the first kind of sparsity, two key tools are graphical models (Whittaker, 2001) that assume statistical dependence between specific variables, and regularization methods that penalize non-sparse solutions (Hastie et al., 2001). Examples of such regularization methods are...
the lasso (Tibshirani, 1996), regularized covariance estimation methods (Bickel and Levina, 2007; Levina and Zhu, 2007), and variable selection in high-dimensional graphs (Meinshausen and Bühlmann, 2006). For the second type of sparsity, where data is assumed to have a low intrinsic dimensionality, two standard dimensionality reduction tools are principal component analysis (Jolliffe, 2002) and wavelets (Ogden, 1997). Many of the latter “variable transformation” methods are not suitable for the analysis of noisy and unordered data in high dimensions. Such data is the interest of this work. Furthermore, collinearity of the variables is often a problem for existing dimensionality reduction techniques of all kinds — including least squares, and variable selection methods that do not take variable groupings into account. This form of sparsity is also of special interest here.

As discussed in the paper, such low-dimensional structure naturally occurs in some linear mixture models, as well as in DNA microarray data where, for example, genes sharing the same pathway may exhibit highly correlated expression patterns.

Principal components (PCA) and wavelets — two common dimensionality reduction and variable transformation techniques which have inspired this work — have a number a strengths and weaknesses. PCA has gained much popularity due to its simplicity and unique property of providing a sequence of best linear approximations in a least squares sense. The method, however, has two main limitations. First, PCA computes a global representation, where each basis vector is a linear combination of all the original variables. This makes it difficult to interpret the results and detect internal localized structures in the data. For example, in gene expression data, it may be difficult to detect small sets of highly correlated genes with a global PCA analysis. The second limitation is that for noisy input data, PCA constructs an optimal representation of the noisy data but not necessarily of the (unknown) underlying noiseless data. When the number of variables $p$ is much larger than the number of observations $n$, the true underlying principal factors may be masked by the noise, yielding an inconsistent estimator in the joint limit $p, n \to \infty, p/n \to c$ (Johnstone and Lu, 2004). Even for a finite sample size $n$, this property of PCA and other global methods (e.g. partial least squares and ridge regression) can lead to large prediction errors in
regression and classification (Buckheit and Donoho, 1995; Nadler and Coifman, 2005b).

In contrast to PCA, wavelet methods describe the data in terms of localized basis functions. The representations are multi-scale, and for smooth data, also sparse (Donoho and Johnstone, 1995). Wavelets are often used in many non-parametric statistics tasks, including regression and density estimation. In recent years, wavelet expansions have also been combined with regularization methods to find regression vectors which are sparse in an a priori known wavelet basis (Candes and Tao, 2005; Donoho and Elad, 2003). The main limitation of wavelets is the implicit assumption of smoothness of the (noiseless) data as a function of its variables. In other words, wavelets are not suited for the analysis of unordered data.

In this paper, we propose a novel multi-scale representation of unstructured data, where variables are randomly ordered and do not necessarily satisfy any specific smoothness criteria. We call the construction treelets, as the method is inspired by both wavelets and hierarchical clustering trees. The motivation for the treelets is two-fold: One goal is to find a “natural” system of coordinates that reflects the underlying internal structures of the data, and that selects groups of similar variables. A second goal is to improve the performance of conventional regression and classification techniques in the “large p, small n” regime by compressing the data prior to learning.

The treelet algorithm starts from a pairwise similarity measure between the original variables and constructs, step by step, a data-driven multi-scale orthonormal basis. The basis functions are supported on nested clusters in a hierarchical tree. As in PCA, we explore the covariance structure of the data but — unlike PCA — the analysis is local and multi-scale. A theoretical analysis furthermore shows that treelets form a sparse basis when the data consist of groups of similar variables.

Other methods also relate to treelets. In recent years, hierarchical clustering algorithms have been widely used for identifying diseases and groups of co-expressed genes (Eisen et al., 1998; Tibshirani et al., 1999). Many researchers are also developing algorithms that combine gene selection and gene grouping; see e.g. Hastie et al. (2001); Dettling and Bühlmann (2004); Zou and Hastie.
The novelty and contribution of our approach to data representation and analysis is the simultaneous construction of a data-driven multi-scale basis and a hierarchical cluster tree. The introduction of a basis enables application of the well-developed machinery of orthonormal expansions, wavelets and wavelet packets for e.g. reconstruction, compression, and denoising of general unordered data. The treelet algorithm is similar to the “local Karhunen-Loeve Basis” for smooth ordered data by Coifman and Saito (1996), where the basis functions are data-driven but the tree structure is fixed. Our paper is also related to recent independent work on the “Haar wavelet transform of a dendrogram” by Murtagh (2007). The latter paper also suggests basis functions on a data-driven cluster tree but uses fixed wavelets on a pre-computed dendrogram. The treelet algorithm offers the advantages of both approaches as it incorporates adaptive basis functions as well as a data-driven tree structure. As shown in this paper, this unifying property turns out to be of key importance for statistical inference and prediction. In particular, when the data contain groups of highly correlated variables, the dominant (maximum-energy) basis functions are approximately constant on each such group.

The organization of the paper is as follows: In Sec. 2 we describe the treelet algorithm. In Sec. 3 we analyze its theoretical properties. In Sec. 4 we discuss its performance on a linear mixture error-in-variable model and give a few illustrative examples of its use in representation and regression problems. Finally, in Sec. 5 we illustrate our method for classification of a data set of internet advertisements and a data set of gene expressions arrays.

A preliminary version of this paper was presented at AISTATS-07 (Lee and Nadler, 2007).

2 The Treelet Transform

In many modern data sets (e.g. DNA microarrays, word-document arrays, financial data, consumer databases, etc.), the data are not only high-dimensional but also redundant with many variables
(the genes, words, etc) that are related to each other. Hierarchical clustering algorithms (Jain et al., 1999; Xu and Wunsch, 2005) are often used for the organization and grouping of the variables of such data sets. These methods offer an easily interpretable description of the data structure in terms of a dendrogram, and only require the user to specify a measure of similarity between groups of observations or variables. So called agglomerative hierarchical methods start at the bottom of the tree and, at each level, merge the two groups with highest inter-group similarity into one larger cluster. The novelty of the proposed treelet algorithm is in constructing not only clusters or groupings of variables, but also a multi-scale orthonormal basis for the data. As in standard multi-resolution analysis (Mallat, 1998), the treelet algorithm provides a set of “scaling functions” defined on nested subspaces $V_0 \supset V_1 \supset \ldots \supset V_L$, and a set of orthogonal “detail functions” defined on residual spaces $\{W_\ell\}_{\ell=1}^L$ where $V_\ell \oplus W_\ell = V_{\ell-1}$. The details of the treelet algorithm are in Sec. 2.1.

In this paper, we measure the similarity $M_{ij}$ between two variables $s_i$ and $s_j$ with the correlation coefficient

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

(1)

where $\Sigma_{ij} = \mathbb{E}[(s_i - \mathbb{E}s_i)(s_j - \mathbb{E}s_j)]$ is the usual covariance. Other information-theoretic or graph-theoretic similarity measures are also possible and can potentially lead to better results. For some applications, one may want to use absolute values of correlation coefficients, or a weighted sum of covariances and correlations as in $M_{ij} = |\rho_{ij}| + \lambda|\Sigma_{ij}|$, where the parameter $\lambda$ is a non-negative number.

2.1 The Algorithm: Jacobi Rotations on Pairs of Similar Variables

The treelet algorithm is inspired by the classical Jacobi method for computing eigenvalues of a matrix (Golub and van Loan, 1996); there are also some similarities with the “Grand Tour” (Asimov, 1985), a visualization tool for viewing multidimensional data through a sequence of orthogonal projections. The main difference from Jacobi’s method — and the reason why the treelet trans-
form, in general, returns an orthonormal basis different from standard PCA — is that treelets are constructed on a hierarchical tree.

The idea is simple. At each level of the tree, we group together the most similar variables and replace them by a coarse-grained “sum variable” and a residual “difference variable”; the new variables are computed by a local PCA (or Jacobi rotation) in two dimensions. Unlike Jacobi’s original method, difference variables are stored, and only sum variables are processed at higher levels of the tree. Hence, the multi-resolution analysis interpretation. The details of the algorithm are as follows:

• At level \( \ell = 0 \), (the bottom of the tree), each observation or “signal” \( x \) is represented by the original variables \( x^{(0)} = [s_{0,1}, \ldots, s_{0,p}]^T \), where \( s_{0,k} = x_k \). Associate to these coordinates, the Dirac basis \( B_0 = [\phi_{0,1}, \phi_{0,2}, \ldots, \phi_{0,p}] \) where \( B_0 \) is the \( p \times p \) identity matrix. Compute the sample covariance and similarity matrices \( \hat{\Sigma}^{(0)} \) and \( \hat{M}^{(0)} \). Initialize the set of “sum variables”, \( S = \{1, 2, \ldots, p\} \).

• Repeat for \( \ell = 1, \ldots, L \)

1. **Find the two most similar sum variables according to the similarity matrix \( \hat{M}^{(\ell-1)} \).**

   Let
   \[
   (\alpha, \beta) = \arg \max_{i,j \in S} \hat{M}_{ij}^{(\ell-1)} .
   \] (2)

   where \( i < j \), and maximization is only over pairs of sum variables that belong to the set \( S \). As in standard wavelet analysis, difference variables (defined in step 3) are not processed.

2. **Perform a local PCA on this pair.** Find a Jacobi rotation matrix

   \[
   J(\alpha, \beta, \theta_\ell) = \begin{bmatrix}
   1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\
   \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\
   0 & \cdots & c & \cdots & -s & \cdots & 0 \\
   \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\
   0 & \cdots & s & \cdots & c & \cdots & 0 \\
   \vdots & \ddots & \vdots & \ddots & \ddots & \ddots & \vdots \\
   0 & \cdots & 0 & \cdots & 0 & \cdots & 1
   \end{bmatrix}
   \] (3)
where \( c = \cos(\theta) \) and \( s = \sin(\theta) \), that decorrelates \( x_\alpha \) and \( x_\beta \); more specifically, find a rotation angle \( \theta \) such that \( |\theta| \leq \pi/4 \) and \( \hat{\Sigma}_{\alpha\beta}^{(l)} = \hat{\Sigma}_{\beta\alpha}^{(l)} = 0 \), where \( \hat{\Sigma}^{(l)} = J^T \hat{\Sigma}^{(l-1)} J \). This transformation corresponds to a change of basis \( B_\ell = J^T B_{l-1} \), and new coordinates \( x^{(\ell)} = J^T x^{(\ell-1)} \). Update the similarity matrix \( \hat{M}^{(\ell)} \) accordingly.

3. **Multi-resolution analysis.** For ease of notation, assume that \( \hat{\Sigma}_{\alpha\alpha}^{(l)} \geq \hat{\Sigma}_{\beta\beta}^{(l)} \) after the Jacobi rotation, where the indices \( \alpha \) and \( \beta \) correspond to the first and second principal components, respectively. Define the sum and difference variables at level \( \ell \) as \( s_\ell = x_\alpha^{(\ell)} \) and \( d_\ell = x_\beta^{(\ell)} \). Similarly, define the scaling and detail functions \( \phi_\ell \) and \( \psi_\ell \) as columns \( \alpha \) and \( \beta \) of the basis matrix \( B_\ell \). Remove the difference variable from the set of sum variables, \( S = S \setminus \{\beta\} \). At level \( \ell \), we have the orthonormal treelet decomposition

\[
    x = \sum_{i=1}^{p-\ell} s_{\ell,i} \phi_{\ell,i} + \sum_{i=1}^{\ell} d_i \psi_i. \tag{4}
\]

where the new set of scaling vectors \( \{\phi_{\ell,i}\}_{i=1}^{p-\ell} \) is the union of the vector \( \phi_\ell \) and the scaling vectors \( \{\phi_{\ell-1,j}\}_{j \neq \alpha,\beta} \) from the previous level, and the new coarse-grained sum variables \( \{s_{\ell,i}\}_{i=1}^{p-\ell} \) are the projections of the original data onto these vectors. As in standard multi-resolution analysis, the first sum is the coarse-grained representation of the signal, while the second sum captures the residuals at different scales.

The output of the algorithm can be summarized in terms of a cluster tree with a height \( L \leq p-1 \) and an ordered set of rotations and pairs of indices, \( \{(\theta_\ell, \alpha_\ell, \beta_\ell)\}_{\ell=1}^{L} \). Fig. 1 (left) shows an example of a treelet construction for a signal of length \( p = 5 \), with the signal representations \( x^{(\ell)} \) at the different levels of the tree shown on the right.

We now briefly consider the complexity of the treelet algorithm on a general data set with \( n \) observations and \( p \) variables. For a naive implementation with an exhaustive search for the optimal pair \( (\alpha, \beta) \) in Eq. 2, the overall complexity is \( m + O(Lp^2) \) operations, where \( m = O(\min(np^2, pn^2)) \) is the cost of computing the sample covariance matrix by singular value de-
composition, and $L$ is the height of the tree. However, by storing the similarity matrices $\hat{\Sigma}(0)$ and $\hat{M}(0)$ and keeping track of their local changes, the complexity can be further reduced to $m + O(Lp)$.

### 2.2 Selecting the Height $L$ of the Tree and a “Best K-Basis”

There are many different methods one could use for selecting the appropriate height $L$ of the tree. In hierarchical (agglomerative) clustering, it is common to stop merging clusters when the similarity measure falls below a certain preset threshold value. Other adaptive methods perform a hypothesis test at each level of the tree to determine whether the difference between two clusters is significant. Any of the model selection methods commonly used in hierarchical clustering can also be applied to treelets. In this work, we propose a different approach to model selection that is inspired by the “best basis” paradigm in wavelet signal processing (Coifman and Wickerhauser, 1992; Saito and Coifman, 1995).

Consider IID data $x^1, \ldots, x^n$, where $x^i \in \mathbb{R}^p$ is a $p$-dimensional random vector. As in standard multi-resolution analysis, one can construct $p$ different orthonormal bases with the treelet algo-
algorithm. Denote these by $B_0, \ldots, B_{p-1}$, where the orthonormal basis $B_\ell$ corresponds to the basis at level $\ell$ in the tree. Suppose now that we are interested in finding the “best” $K$-dimensional treelet representation, where the dimension $K < p$ is assumed to be given, and where we measure how good a model is in terms of a suitably defined score. In this paper, we use a measure based on normalized energies, though many other alternative measures exist (e.g. Fisher’s discriminant score, classification error rates, entropy and other sparsity measures etc.). Furthermore, we propose the following greedy scoring and selection scheme:

For a given orthonormal basis $B = (w_1, \ldots, w_p)$, assign a normalized energy score $\mathcal{E}$ to each basis vector $w_i$ according to

$$\mathcal{E}(w_i) = \frac{\mathbb{E}\{|w_i \cdot x|^2\}}{\mathbb{E}\{||x||^2\}}. \quad (5)$$

The corresponding sample estimate is $\hat{\mathcal{E}}(w) = \frac{\sum_{j=1}^{n} |w_i \cdot x_j|^2}{\sum_{j=1}^{n} ||x_j||^2}$. Sort the vectors according to decreasing energy, $w_{(1)}, \ldots, w_{(p)}$, and define the score $\Gamma_K$ of the basis $B$ by summing the $K$ largest terms, i.e. let $\Gamma_K(B) = \sum_{i=1}^{K} \mathcal{E}(w_i)$. The best $K$-basis is the treelet basis with the highest score

$$B_L = \arg \max_{B_\ell: 0 \leq \ell \leq p-1} \Gamma_K(B_\ell), \quad (6)$$

i.e. the basis that best compresses data with only $K$ components. In case of degeneracies, we choose the basis with the smallest $\ell$. Furthermore, to estimate the score $\Gamma_K$ given a particular data set, we use cross-validation to avoid overfitting. In other words, the treelets are constructed using subsets of the original data set and the score is computed on independent test sets. Both theoretical calculations (Sec. 3.2) and simulations (see e.g. Fig. 3) indicate that an energy-based measure is useful for detecting natural groupings of variables in data.

### 3 Theory

#### 3.1 Large Sample Properties of the Treelet Transform

In this section we examine some of the large sample properties of treelets, when one uses correlation coefficients (Eq. 1) to measure similarities of variables. We start by introducing some notation
and definitions: Let $T(\Sigma) = J^T \Sigma J$ denote the covariance matrix after one step of the treelet algorithm when starting with covariance matrix $\Sigma$. Let $T^\ell(\Sigma)$ denote the covariance matrix after $\ell$ steps of the treelet algorithm. Thus, $T^\ell = T \circ \cdots \circ T$ corresponds to $T$ applied $\ell$ times. Define $||A||_\infty = \max_{j,k} |A_{jk}|$ and let

$$T_n(\Sigma, \delta_n) = \bigcup_{||\Lambda - \Sigma||_\infty \leq \delta_n} T(\Lambda).$$  \hspace{1cm} (7)

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

$$T_n^\ell(\Sigma, \delta_n) = \bigcup_{\Lambda \in T_n^{\ell-1}} T(\Lambda), \hspace{0.5cm} \ell \geq 2. \hspace{1cm} (8)$$

Let $\Sigma_n$ denote the sample covariance matrix. We make the following assumptions:

(A1) Assume that $x$ has 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 $\mathbb{E}(|x_j x_k|^q) \leq q! M^{q-2}s/2$ for all $q \geq 2$.

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

**Theorem 1** Suppose that (A1) and (A2) hold. Let $\delta_n = K \sqrt{\log n/n}$ where $K > 2c$. Then, as $n, p_n \to \infty$,

$$\mathbb{P}(T_n^\ell(\Sigma_n) \in T_n^\ell(\Sigma, \delta_n), \ell = 1, \ldots, p_n) \to 1. \hspace{1cm} (9)$$

Some discussion is in order. The result says that $T_n^\ell(\Sigma_n)$ is not too far from $T_n^\ell(\Lambda)$ for some $\Lambda$ close to $\Sigma$. It would perhaps be more satisfying to have a result that says that $||T^\ell(\Sigma) - T^\ell(\Sigma_n)||_\infty$ converges to 0. This would be possible if one used covariances to measure similarity, but not in the case of correlation coefficients.

For example, it is easy to construct a covariance matrix $\Sigma$ with following properties:

1. $\rho_{12}$ is the largest off-diagonal correlation,

2. $\rho_{34}$ is nearly equal to $\rho_{12}$,
3. the $2 \times 2$ submatrix of $\Sigma$ corresponding to $x_1$ and $x_2$ is very different than the $2 \times 2$ submatrix of $\Sigma$ corresponding to $x_3$ and $x_4$.

In this case, there is nontrivial probability that $\hat{\rho}_{34} > \hat{\rho}_{12}$ due to sample fluctuations. Therefore $T(\Sigma)$ performs a rotation on the first two coordinates while $T(\hat{\Sigma})$ performs a rotation on the third and fourth coordinates. Since the two corresponding submatrices are quite different, the two rotations will be quite different. Hence, $T(\Sigma)$ can be quite different from $T(\hat{\Sigma})$. This does not pose any problem since inferring $T(\Sigma)$ is not the goal. Under the stated conditions, we would consider both $T(\Sigma)$ and $T(\hat{\Sigma})$ to be reasonable transformations. We examine the details and include the proof of Theorem 1 in Appendix A.

Because $T(\Sigma_1)$ and $T(\Sigma_2)$ can be quite different even when the matrices $\Sigma_1$ and $\Sigma_2$ are close, it might be of interest to study the stability of $T(\hat{\Sigma}_n)$. This can be done using the bootstrap. Construct $B$ bootstrap replications of the data and corresponding sample covariance matrices $\hat{\Sigma}_{n,1}, \ldots, \hat{\Sigma}_{n,B}$. Let $\delta_n = J_n^{-1}(1 - \alpha)$, where $J_n$ is the empirical distribution function of $\{||\hat{\Sigma}_{n,b}^* - \hat{\Sigma}_n||_{\infty}, b = 1, \ldots, B\}$ and $\alpha$ is the confidence level. If $F$ has finite fourth moments and $p$ is fixed, then it follows from Corollary 1 of Beran and Srivastava (1985) that

$$\lim_{n \to \infty} P_F(\Sigma \in C_n) = 1 - \alpha$$

where $C_n = \{\Lambda : ||\Lambda - \hat{\Sigma}_n||_{\infty} \leq \delta_n\}$. Let

$$A_n = \left\{ T(\Lambda) : \Lambda \in C_n \right\}.$$  

It follows that $P(T(\Sigma) \in A_n) \to 1 - \alpha$. The set $A_n$ can be approximated by applying $T$ to all $\hat{\Sigma}_{n,b}^*$ for which $||\hat{\Sigma}_{n,b}^* - \hat{\Sigma}_n||_{\infty} < \delta_n$.

### 3.2 Treelets on Covariance Matrices with Block Structures

#### 3.2.1 An Exact Analysis in the Limit $n \to \infty.$

Many real life data sets, including gene arrays, consumer data sets and word-documents, display covariance matrices with approximate block structures. The treelet transform is especially well
suited for representing and analyzing such data — even in cases when the clusters are not revealed until after an a priori unknown permutation of the original variables. Here we show that treelets provide a *sparse* representation when covariance matrices have inherent block structures. We consider an ideal situation where variables within the same group are exactly collinear, and variables from different groups are weakly correlated. All calculations are exact and are computed on the population covariance matrix of the variables, i.e. in the limit of the sample size $n \to \infty$. An analysis of convergence rates appears in Sec. 3.2.2. For illustrative purposes, we have ordered the variables.

We begin by analyzing treelets on $p$ random variables that are indistinguishable with respect to their second-order statistics. We show that the treelet algorithm returns scaling functions that are *constant on groups of indistinguishable variables*. In particular, the scaling function on the full set of variables in a block is a constant function. These results, as well as the following Theorem 2 and Corollary 1 are due to the *fully adaptive* nature of the treelet algorithm — a property that sets treelets apart from algorithms that use fixed wavelets on a dendrogram (Murtagh, 2007), or adaptive basis functions on fixed trees (Coifman and Saito, 1996).

**Lemma 1** Assume that $x = (x_1, x_2, \ldots, x_p)^T$ is a random vector with distribution $F$, mean 0 and covariance matrix $\Sigma = \sigma_1^2 1_{p \times p}$, where $1_{p \times p}$ denotes a $p \times p$ matrix with all entries equal to 1. Then, at any level $1 \leq \ell \leq p - 1$ of the tree, the treelet operator $T^\ell$ (defined in Sec. 3.1) returns for the population covariance matrix $\Sigma$, an orthogonal decomposition

$$T^\ell(\Sigma) = \sum_{i=1}^{p-\ell} s_{\ell,i} \phi_{\ell,i} + \sum_{i=1}^{\ell} d_i \psi_i$$

with sum variables $s_{\ell,i} = \frac{1}{\sqrt{|A_{\ell,i}|}} \sum_{j \in A_{\ell,i}} x_j$ and scaling functions $\phi_{\ell,i} = \frac{1}{\sqrt{|A_{\ell,i}|}} I_{s_{\ell,i}}$, defined on disjoint index subsets $A_{\ell,i} \subseteq \{1, \ldots, p\}$ (for $i = 1, \ldots, p-\ell$) with lengths $|A_{\ell,i}|$ and $\sum_{i=1}^{p-\ell} |A_{\ell,i}| = p$ elements total. The expansion coefficients have variances $\mathbb{V}\{s_{\ell,i}\} = |A_{\ell,i}| \sigma_1^2$, and $\mathbb{V}\{d_i\} = 0$ for
all $i$. In particular, for $\ell = p - 1$,
\begin{equation}
T^{p-1}(\Sigma) = s \phi + \sum_{i=1}^{p-1} d_i \psi_i
\end{equation}
where $s = \frac{1}{\sqrt{p}} (x_1 + \ldots + x_p)$ and $\phi = \frac{1}{\sqrt{p}} [1 \ldots 1]^T$.

**Remark 1** Uncorrelated additive noise in $(x_1, x_2, \ldots, x_p)$ adds a diagonal perturbation to the $2 \times 2$ covariance matrices $\Sigma^{(\ell)}$, which are computed at each level in the tree (see Eq. 35). Such noise may affect the order in which variables are grouped, but the asymptotic results of the lemma remain the same.

Next we consider data where the covariance matrix is a $K \times K$ block matrix and white noise has been added to the original variables. The following main result states that, if variables from different blocks are weakly correlated and the noise level is relatively small, then the $K$ maximum variance scaling functions are constant on each block (see Example 2 and Fig. 2 in Sec. 4). We give a sufficient condition for this ideal solution in terms of within-block and between-block correlations of the original data and the noise level (Eq. 13). As before, we are considering the limit $n \to \infty$. A $p \times p$ identity matrix is denoted by $I_p$, and a $p_i \times p_j$ matrix with all entries equal to 1 is denoted by $1_{p_i \times p_j}$.

**Theorem 2** Assume that $x = (x_1, x_2, \ldots, x_p)^T$ 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 and
\begin{equation}
C = \\
\begin{pmatrix}
C_{11} & C_{12} & \ldots & C_{1K} \\
C_{12} & C_{22} & \ldots & C_{2K} \\
\vdots & \vdots & \ddots & \vdots \\
C_{1K} & C_{2K} & \ldots & C_{KK}
\end{pmatrix}
\end{equation}
is a $K \times K$ block matrix with “within-block” covariance matrices $C_{kk} = \sigma^2 1_{p_k \times p_k}$ ($k = 1, \ldots, K$) and “between-block” covariance matrices $C_{ij} = \sigma_{ij} 1_{p_i \times p_j}$ ($i, j = 1, \ldots, K$; $i \neq j$). If
\begin{equation}
\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)}},
\end{equation}

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

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

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

If the conditions of the theorem are satisfied, then all treelets (both scaling and difference functions) associated with levels $\ell > p - K$ are constant on groups of similar variables. In particular, for a full decomposition at the maximum level $\ell = p - 1$ of the tree we have the following result, which follows directly from Theorem 2:

**Corollary 1** Assume that the conditions in Theorem 2 are satisfied. A full treelet decomposition then 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 coefficients $d_1, \ldots, d_{p-K}$ which only reflect noise in the data with variances $\mathbb{V}\{d_i\} = O(\sigma^2)$ for $i = 1, \ldots, p - K$.

**Remark 2** Both Theorem 2 and Corollary 1 can be directly generalized to include $p_0$ uncorrelated noise variables, so that $x = (x_1, \ldots, x_{p-p_0}, x_{p-p_0+1}, \ldots, x_p)^T$, where $\mathbb{E}(x_i) = 0$ and $\mathbb{E}(x_i x_j) = 0$ for $i > p - p_0$ and $j \neq i$. For example, if Eq. 13 is satisfied, then the treelet decomposition at level $\ell = p - p_0$ is

$$T^{p-p_0}(\Sigma) = \sum_{k=1}^{K} s_k \phi_k + \sum_{i=1}^{p-p_0-K} d_i \psi_i + (0, \ldots, 0, x_{p-p_0+1}, \ldots, x_p)^T$$

Furthermore, note that according to Eq. 41 in the appendix, within-block correlations are smallest (“worst-case scenario”) when singletons are merged. Thus, the treelet transform is a stabilizing algorithm; once a few correct coarse-grained variables have been computed, it has the effect of denoising the data.
3.2.2 Convergence Rates

The aim of this section is to provide a rough estimate of the sample size required for treelets to discover the inherent structures of data. For covariance matrices with block structures, we show that treelets find the correct groupings of variables if the sample size $n \gg O(\log p)$, where $p$ is the dimension of the data. This is an interesting result, as standard PCA is consistent if and only if $p/n \to 0$ (Johnstone and Lu, 2004), i.e. when $n \gg O(p)$. The result is also comparable to the result in Bickel and Levina (2007) for regularization of sparse nearly diagonal covariance matrices. One main difference is that their paper assumes an a priori known ordered set of variables in which the covariance matrix is sparse, whereas treelets find such an ordering as part of the algorithm. The argument for treelets is as follows.

Assume that there are $K$ blocks in the population covariance matrix $\Sigma$. Define $A_{L,n}$ as the event that the $K$ maximum variance treelets, constructed at level $L = p - K$ of the tree for a data set with $n$ observations, are supported only on variables from the same block. In other words, let $A_{L,n}$ represent the ideal case where the treelet transform finds the correct groupings of variables. Let $E_\ell$ denote the event that at level $\ell$ of the tree, the largest between-block sample correlation is less than the smallest within-block sample correlation,

$$E_\ell = \{\max|\hat{\rho}_B^{(\ell)}| < \min|\hat{\rho}_W^{(\ell)}|\}.$$

According to Eqs.31-32, the corresponding population correlations

$$\max \rho_B^{(\ell)} < \rho_1 \equiv \max_{1 \leq i,j \leq K} \left(\frac{\sigma_{ij}}{\sigma_i \sigma_j}\right), \quad \min \rho_W^{(\ell)} > \rho_2 \equiv \frac{1}{\sqrt{1 + 3 \max(\delta^2, \delta^4)}},$$

where $\delta = \frac{\sigma_{\min}}{\sigma_k}$, for all $\ell$. Thus, a sufficient condition for $E_\ell$ is that $\{\max|\hat{\rho}_B^{(\ell)} - \rho_B^{(\ell)}| < t\} \cap \{\max|\hat{\rho}_W^{(\ell)} - \rho_W^{(\ell)}| < t\}$, where $t = (\rho_2 - \rho_1)/2 > 0$. We have that

$$\mathbb{P}(A_{L,n}) \geq \mathbb{P}\left(\bigcap_{0 \leq \ell < L} E_\ell\right) \geq \mathbb{P}\left(\bigcap_{0 \leq \ell < L} \{\max|\hat{\rho}_B^{(\ell)} - \rho_B^{(\ell)}| < t\} \cap \{\max|\hat{\rho}_W^{(\ell)} - \rho_W^{(\ell)}| < t\}\right).$$
If (A1) holds, then it follows from Lemma 3 that

$$\mathbb{P}(A_{L,n}^C) \leq \sum_{0 \leq t < L} \left( \mathbb{P}(\max |\hat{\rho}_B^{(l)} - \rho_B^{(l)}| > t) + \mathbb{P}(\max |\hat{\rho}_W^{(l)} - \rho_W^{(l)}| > t) \right) \leq Lc_1 p^2 e^{-nc_2 t^2}$$

for positive constants $c_1, c_2$. Thus, the requirement $\mathbb{P}(A_{L,n}^C) < \alpha$ is satisfied if the sample size

$$n \geq \frac{1}{c_2 t^2} \log \left( \frac{Lc_1 p^2}{\alpha} \right).$$

From the large-sample properties of treelets (Sec. 3.1), it follows that treelets are consistent if $n \gg O(\log p)$.

## 4 Treelets and a Linear Error-In-Variables Mixture Model

In Sec. 3.2 we assumed a general model of the data where the covariance matrix has inherent block structures. In this section, we study a simple error-in-variables linear mixture model (factor model) which, under some conditions, gives rise to such covariance matrices. Under this model, we compare treelets with some common algorithms for supervised and unsupervised tasks. An advantage of introducing a concrete generative model is that we can easily relate our results to the underlying structures or components of the data.

Here we examine a linear mixture model with $K$ components and additive noise, where each multivariate observation $x \in \mathbb{R}^p$ has the form

$$x = \sum_{j=1}^{K} u_j v_j + \sigma z. \tag{15}$$

The components or “factors” $u_j$ are random (but not necessarily independent) variables with variances $\sigma_j^2$. The “loading vectors” $v_j$ are fixed but typically unknown linearly independent vectors. In the last term, $\sigma$ represents the noise level, and $z \sim \mathcal{N}_p(0, I)$ is a $p$-dimensional random vector.

In the unsupervised setting, we are given a training set $\{x_i\}_{i=1}^{n}$ sampled from Eq. 15. Unsupervised learning tasks include, for example, inference on the number of components $K$ and on the underlying vectors $v_j$. In the supervised setting, we consider a data set $\{x_i, y_i\}_{i=1}^{n}$, where the
response value $y$ of an observation $x$ is a linear combination of the variables $u_j$ with a random noise term $\epsilon$,

$$y = \sum_{j=1}^{K} \alpha_j u_j + \epsilon. \quad (16)$$

The standard supervised learning task in regression and classification is prediction of $y$ for new data $x$ given a training set $\{x_i, y_i\}_{i=1}^{n}$.

Linear mixture models are common in many fields, including spectroscopy and gene expression analysis. In spectroscopy Eq. 15 is known as Beer’s law, where $x$ is the logarithmic absorbance spectrum of a chemical substance measured at $p$ wavelengths, $u_j$ are the concentrations of constituents with pure absorbance spectra $v_j$, and the response $y$ is typically one of the components, $y = u_i$. In gene data, $x$ is the measured expression level of $p$ genes, $u_j$ are intrinsic activities of various pathways, and each vector $v_j$ represents the set of genes in a pathway. The quantity $y$ is typically some measure of severity of a disease such as time until recurrence of cancer. A linear relation between $y$ and the values of $u_j$ as in Eq. 16 is commonly assumed.

### 4.1 Treelets and a Linear Mixture Model in the Unsupervised Setting

Consider data $\{x_i\}_{i=1}^{n}$ from the model in Eq. 15. Here we analyze in detail an illustrative example with $K = 3$ components and loading vectors $v_k = I(B_k)$, where $I$ is the indicator function and $B_k \subset \{1, 2, \ldots, p\}$ are sets of variables with sizes $p_k = |B_k|$, for $k = 1, 2, 3$. A more general analysis is possible but may not provide more insight.

The unsupervised task is to uncover the internal structure of the linear mixture model from data, e.g. to infer the unknown structure of the vectors $v_k$, including the sizes $p_k$ of the sets $B_k$. The difficulty of this problem depends, among other things, on possible correlations between the random variables $u_j$, the variances of the components $u_j$, and interferences (overlap) between the loading vectors $v_k$. We present three examples with increasing difficulty. Standard methods, such as principal component analysis, succeed only in the simplest case, whereas more sophisticated methods such as sparse PCA (elastic nets) sometimes require oracle information to correctly fit
tuning parameters in the model. The treelet transform seems to perform well in all three cases and is also, because of its simplicity, very amenable to theoretical analysis.

**Example 1: Uncorrelated Factors and Non-Overlapping Loading Vectors**

Suppose that the random variables $u_j$ are all uncorrelated for $j = 1, 2, 3$, and that the loading vectors $v_j$ are non-overlapping. The population covariance matrix of $x$ is then given by $\Sigma = C + \sigma^2 I_p$ where the noise-free matrix

$$ C = \begin{pmatrix} C_{11} & 0 & 0 & 0 \\ 0 & C_{22} & 0 & 0 \\ 0 & 0 & C_{33} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix} $$

is a $4 \times 4$ block matrix with the first three blocks $C_{kk} = \sigma_k^2 1_{p_k \times p_k}$ ($k = 1, 2, 3$), and the last diagonal block having all entries equal to zero.

Assume that $\sigma_k \gg \sigma$ for $k = 1, 2, 3$. This is a specific example of a “spiked covariance model” (Johnstone, 2001); the three components here correspond to distinct large eigenvalues or “spikes” of a covariance model with added background noise. As $n \to \infty$ (with $p$ fixed), PCA recovers the hidden vectors $v_1, v_2,$ and $v_3$, since these three vectors exactly coincide with the principal eigenvectors of $\Sigma$. A treelet transform with $K = 3$, and a height $\ell$ determined by cross-validation with the maximum energy criterion (see Sec. 2.2) returns the same results asymptotically.

The difference between PCA and treelets becomes obvious however in the small $n$ — large $p$ regime. As shown in (Johnstone and Lu, 2004), in the joint limit $p, n \to \infty$, standard PCA computes consistent estimators of the vectors $v_j$ if and only if $p(n)/n \to 0$. As described in Sec. 3.2.2 treelets require asymptotically far fewer observations with the condition for consistency being $\log p(n)/n \to 0$. 
Example 2: Correlated Factors and Non-Overlapping Loading Vectors

Suppose now that the random variables $u_j$ are correlated. Specifically, assume that the loading vectors $v_1$, $v_2$ and $v_3$ are non-overlapping, but that the corresponding factors or components are dependent according to

$$u_1 \sim N(0, \sigma_1^2), \quad u_2 \sim N(0, \sigma_2^2), \quad u_3 = c_1 u_1 + c_2 u_2.$$  \hfill (18)

The covariance matrix is now given by $\Sigma = C + \sigma^2 I_p$ where

$$C = \begin{pmatrix} C_{11} & 0 & C_{13} & 0 \\ 0 & C_{22} & C_{23} & 0 \\ C_{13} & C_{23} & C_{33} & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}$$  \hfill (19)

with $C_{kk} = \sigma_k^2 I_{p_k \times p_k}$ (note that $\sigma_3^2 = c_1^2 \sigma_1^2 + c_2^2 \sigma_2^2$), $C_{13} = c_1 \sigma_1^2 I_{p_1 \times p_3}$ and $C_{23} = c_2 \sigma_2^2 I_{p_2 \times p_3}$. Due to the correlations between $u_j$, the loading vectors of the block model no longer coincide with the principal eigenvectors, and it is difficult to extract them with PCA.

We illustrate this problem by the example considered in Zou et al. (2006). Specifically, let

$$v_1 = \begin{bmatrix} B_1 \\ B_2 \\ B_3 \end{bmatrix}^T$$
$$v_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 1 \\ 1 \\ 1 \\ 1 \\ 0 \\ 0 \end{bmatrix}^T$$
$$v_3 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 0 \\ 1 \\ 1 \end{bmatrix}^T$$  \hfill (20)

where each set $B_j$ is disjoint for $j = 1, 2, 3$ with $p_1 = p_2 = 4, p_3 = 2$ variables, respectively, and there are $p = 10$ variables total. Let $\sigma_1^2 = 290$, $\sigma_2^2 = 300$, $c_1 = -0.3$, $c_2 = 0.925$, and $\sigma = 1$. The corresponding variance $\sigma_3^2$ of $u_3$ is 282.8, and the covariances of the off-diagonal blocks are $\sigma_{13} = -87$ and $\sigma_{23} = 277.5$.

The first three PCA vectors for a training set of 1000 samples are shown in Fig. 2 (left). As expected, it is difficult to detect the underlying vectors $v_i$ from these results. Other methods, such as PCA with thresholding (Zou et al., 2006) also fail to achieve this goal, even with an infinite number of observations, i.e. in the limit $n \to \infty$. This example illustrates the limitations of a global approach such as PCA, since ideally, we would detect that for example the variables $(x_5, x_6, x_7, x_8)$
Figure 2: In Example 2, PCA fails to find the important variables in the three-component mixture model, as the computed eigenvectors (left) are perturbed by correlations between different components. On the other hand, the three maximum energy treelets (right) uncover the underlying data structures.

are all related and extract the latent vector \( v_2 \) from only these variables. In Zou et al. (2006), the authors show by simulation that a combined \( L_1 \) and \( L_2 \)-penalized least squares method, which they call “sparse PCA” or “elastic nets”, correctly identifies the sets of important variables if given “oracle information” on the number of variables \( p_1, p_2, p_3 \) in the different blocks. Treelets are similar in spirit to elastic nets as both methods tend to group highly correlated variables together. The treelet algorithm is able to find the hidden loading vectors \( v_i \) knowing only \( K \), the number of components in the linear mixture model, and does not require tuning of any additional sparseness parameters. Fig. 2 (right) shows results from a treelet simulation with a large sample size \( (n = 1000) \) and a height \( L = 7 \) of the tree, determined by cross-validation with the maximum energy criterion. The three maximum energy basis vectors correspond exactly to the hidden loading vectors in Eq. 20.

**Example 3: Uncorrelated Factors and Overlapping Loading Vectors**

Finally, we study a challenging example where the first two loading vectors \( v_1 \) and \( v_2 \) are overlapping and the sample size \( n \) is small. Let \( \{B_1, \ldots, B_4\} \) be disjoint subsets of \( \{1, \ldots, p\} \), and
let
\[ v_1 = I(B_1) + I(B_2), \quad v_2 = I(B_2) + I(B_3), \quad v_3 = I(B_4) \] (21)

where \( I(B_k) \) as before represents the indicator function for subset \( k \) (\( k = 1, \ldots, 4 \)). The population covariance matrix is then given by \( \Sigma = C + \sigma^2 I_p \) where the noiseless matrix has the general form
\[
C = \begin{pmatrix}
C_{11} & C_{12} & 0 & 0 & 0 \\
C_{12} & C_{22} & C_{23} & 0 & 0 \\
0 & C_{23} & C_{33} & 0 & 0 \\
0 & 0 & 0 & C_{44} & 0 \\
0 & 0 & 0 & 0 & 0 
\end{pmatrix}
\] (22)

with diagonal blocks \( C_{11} = \sigma_1^2 \mathbf{1}_{p_1 \times p_1}, C_{22} = (\sigma_1^2 + \sigma_2^2) \mathbf{1}_{p_2 \times p_2}, C_{33} = \sigma_2^2 \mathbf{1}_{p_3 \times p_3}, C_{44} = \sigma_3^2 \mathbf{1}_{p_4 \times p_4} \), and off-diagonal blocks \( C_{12} = \sigma_1^2 \mathbf{1}_{p_1 \times p_2} \) and \( C_{23} = \sigma_2^2 \mathbf{1}_{p_2 \times p_3} \). Consider a numerical example with sample size \( n = 100, p = 500 \) variables and noise level \( \sigma = 0.5 \). We choose the same form for the components \( u_1, u_2, u_3 \) as in [Bair et al., 2006] but in contrast to Bair’s paper, we associate the first two components with overlapping loading vectors \( v_1 \) and \( v_2 \). Specifically, the components are given by \( u_1 = \pm 0.5 \) with equal probability, \( u_2 = I(U_2 < 0.4) \), and \( u_3 = I(U_3 < 0.3) \) where \( I(x) \) is the indicator of \( x \), and \( U_j \) are all independent uniform random variables in \([0,1]\); the corresponding variances are \( \sigma_1^2 = 0.25, \sigma_2^2 = 0.24, \) and \( \sigma_3^2 = 0.21 \). As for the blocks \( B_k \), we consider \( B_1 = \{1, \ldots, 10\}, B_2 = \{11, \ldots, 50\}, B_3 = \{51, \ldots, 100\}, \) and \( B_4 = \{201, \ldots, 400\} \).

Inference in this case is challenging for several different reasons. The sample size \( n < p \), the loading vectors \( v_1 \) and \( v_2 \) are overlapping in the region \( B_2 = \{11, \ldots, 50\} \), the signal-to-noise ratio is low with the variance \( \sigma^2 \) of the noise essentially being of the same size as the variances \( \sigma_j^2 \) of \( u_j \), and the condition in Eq. [13] is not satisfied even for the population covariance matrix. Despite these difficulties, the treelet algorithm is remarkably stable, returning results that by and large correctly identify the internal structures of the data. The details are summarized below and in Fig. [3].

Fig. [3] (top left), for example, shows the energy score of the best \( K \)-basis at different levels of the tree. We used 5-fold cross-validation; i.e. we generated a single data set of \( n = 100 \)
observations, but in each of the 5 computations the treelets were constructed on a subset of 80 observations with 20 observations left out for the energy score computation. The five curves as well as their average clearly indicate a “knee” at the level $L = 300$. This is consistent with our expectations that the treelet algorithm mainly merges noise variables at levels $L \geq |\bigcup_k B_k|$. For a tree with “optimum” height $L = 300$, we then constructed a treelet basis on all $p = 500$ variables. Fig. 3 (top right) shows the energy of these treelets sorted according to descending energy score. The results indicate that we have two dominant treelets, while the remaining treelets have an energy that is either slightly higher or of the same order as the variance of the noise. In Fig. 3 (bottom left), we plot the loadings of the four highest energy treelets: “Treelet 1” (red) is approximately constant on the set $B_4$, “Treelet 2” (blue) is approximately piecewise constant on blocks $B_1$, $B_2$ and $B_3$, while the low-energy degenerate treelets 2 (green) and 3 (magenta) seem to take differences between variables in the sets $B_1$, $B_2$ and $B_3$. Finally, we computed 95% confidence bands of the treelets using 1000 bootstrap samples (see Sec. 3.1). Fig. 3 (bottom right) indicate that the treelet results for the two maximum energy treelets are rather stable despite the small sample size and the low signal-to-noise ratio. Most of the time the first treelet selects variables from $B_4$, and most of the time the second treelet selects variables from $B_2$ and either $B_1$ or $B_3$ or both sets. The low-energy treelets seem to pick up differences between blocks $B_1$, $B_2$ and $B_3$ but the exact order in which they select the variables vary from simulation to simulation. As described in the next section, for the purpose of regression, the key point is that the linear span of the first few highest energy treelets is a good approximation to the span of the unknown loading vectors, $\text{Span}\{v_1, \ldots, v_K\}$.

4.2 The Treelet Transform as a Feature Selection Scheme Prior to Regression

Next we consider a typical regression or classification problem with a training set $\{x_i, y_i\}_{i=1}^n$ from Eqs. (15) and (16). Since the data $x$ is noisy, this is an error-in-variables type problem. Given the training data, the goal is to construct a linear function $f : \mathbb{R}^p \rightarrow \mathbb{R}$ to predict $\hat{y} = f(x)$ for a new
The predicted response has the form $\hat{y} = r \cdot x$, where $r$ is the regression vector that minimizes the mean squared error (MSE), $\mathbb{E}\{(\hat{y} - y)^2\}$.

For simplicity, we assume that $y = u_1$ in Eq. 16. We denote by $P_1 : \mathbb{R}^p \to \mathbb{R}^p$ the projection operator onto the space spanned by the vectors $\{v_2, \ldots, v_K\}$. When $n \to \infty$, the regression vector of the unbiased estimator with smallest mean squared error is proportional to $v_y = v_1 - P_1 v_1$. For example, in the case of only two components this vector is given by

$$v_y = v_1 - \frac{v_1 \cdot v_2}{\|v_2\|^2} v_2$$

(23)

The vector $v_y$ is the part of the loading vector $v_1$ that is unique to the response variable $y = u_1$. It plays a central role in chemometrics, where it is known as the net analyte signal (Lorber et al., 2010).
The corresponding mean squared error is given by
\[ \mathbb{E}\{(\hat{y} - y)^2\} = \frac{\sigma^2}{\|v_y\|^2} \]  
(24)

As common in point estimation, one can lower the MSE further via *shrinkage* (Gruber, 1998).

In the case of a linear mixture model, the optimal-MSE regression vector shrinks the net analyte signal vector towards zero, while adding a suitable linear combination of the different loading vectors,
\[ v_{\text{opt}} = v_y(1 - c\sigma^2) + \sigma^2 \sum_{j=1}^{K} \beta_j v_j. \]  
(25)

The shrinkage coefficients \(c, \beta_1, \ldots, \beta_K\) are complicated expressions which depend both on the covariance matrix of the different components as well as on the inner products of their loading vectors (Nadler and Coifman, 2005b).

In typical applications, the number of variables is much larger than the number of observations \(p \gg n\). Two common approaches to overcome this problem include principal component regression (PCR) and partial least squares (PLS). Both methods first perform a global dimensionality reduction from \(p\) to \(k\) variables, and then apply linear regression on these \(k\) features. While in the limit \(n \to \infty\) both methods are MSE-optimal, e.g., they compute the regression vector of Eq. 25, their main limitation is that in the presence of noisy high dimensional data, the computed projections are noisy themselves. In fact, by performing a Taylor expansion of the regression coefficient as a function of noise level \(\sigma\), for fixed \(p, n\), the averaged prediction error of these methods has the form (Nadler and Coifman, 2005b)
\[ \mathbb{E}\{(\hat{y} - y)^2\} \simeq \frac{\sigma^2}{\|v_y\|^2} \left[ 1 + \frac{c_1}{n} + \frac{c_2 \sigma^2}{\mu \|v_y\|^2} \frac{p^2}{n^2} (1 + o(1)) \right]. \]  
(26)

where \(\mu\) is a measure of the variances and covariances of the different components \(u_j\), and \(c_1, c_2\) are both \(O(1)\) constants, independent of \(\sigma, p, n\). This formula shows that when \(p \gg n\) the last term in (26) can dominate and lead to large prediction errors. This emphasizes the need for robust feature
Figure 4: **Left**: The vector $v_y$ (only first 150 coordinates are shown, the rest are zero). **Right**: Averaged prediction errors of 20 simulation results for the methods from top to bottom: PLS on all variables (blue), supervised PLS with variable selection (purple), PLS on treelet features (green), PLS on projections onto the true vectors $v_i$ (red).

...selection and dimensionality reduction of the underlying noise-free data prior to application of learning algorithms such as PCR and PLS.

Variable selection schemes, and specifically those that choose a small subset of variables based on their individual correlation with the response $y$ are also common approaches to dimensionality reduction in this setting. To analyze their performance we consider a more general dimensionality reduction transformation $T: \mathbb{R}^p \rightarrow \mathbb{R}^k$ defined by $k$ orthonormal projections $w_i$,

$$T x = (x \cdot w_1, x \cdot w_2, \ldots, x \cdot w_k)$$ (27)

This family of transformations includes variable subset selection methods, where each projection $w_j$ selects a single variable, as well as wavelet-type methods and our proposed treelet transform. Since an orthonormal projection of a Gaussian noise vector in $\mathbb{R}^p$ is a Gaussian vector in $\mathbb{R}^k$, the prediction error in the new variables admits the form

$$\mathbb{E}\{(\hat{y} - y)^2\} \approx \frac{\sigma^2}{\|T v_y\|^2} \left[ 1 + \frac{c_1}{n} + \frac{c_2 \sigma^2}{\mu \|T v_y\|^2 n^2 (1 + o(1))} \right]$$ (28)

Eq. (28) indicates that a dimensionality reduction scheme should ideally preserve the signal vector...
unique to $y$ ($\|Tv_y\| \simeq \|v_y\|$), while at the same time represent the signals by as few features as possible ($k \ll p$). The main problem of PCA is that it optimally fits the noisy data, yielding for the noise-free response $\|Tv_y\|/\|v_y\| \approx (1 - \frac{c\sigma^2 p^2}{n^2})$. The main limitation of variable subset selection schemes is that in complex settings with overlapping vectors $v_j$, such schemes may at best yield $\|Tv_y\|/\|v_y\| = \alpha < 1$. However, due to high dimensionality, variable subset selection schemes may still achieve better prediction errors than methods that use all the original variables. If the data $x$ is a priori known to be smooth continuous signals, then this feature selection can be performed by wavelet compression, which is known to be asymptotically optimal. In the case of unstructured data, we propose to use treelets.

To illustrate these points, we revisit Example 3 in Sec. 4.1 and compare treelets to the variable subset selection scheme of [Bair et al., 2006] for PLS, as well as standard (global) PLS. As before, we consider a relatively small training set of size $n = 100$ but here we include 1500 additional noise variables, so that $p = 2000 \gg n$. We assume that the response is given by $y = 2u_1$. The vectors $v_j$ are shown in Fig. 3 (top left). The two vectors $v_1$ and $v_2$ overlap, but $v_1$ and $v_3$ are orthogonal. Therefore, the response vector unique to $y$ is given by Eq. 23; see Fig. 4 (left).

To compute $v_y$, all the 100 first coordinates (the set $B_1 \cup B_2 \cup B_3$) are needed. However, a feature selection scheme that chooses variables based on their correlation to the response will pick the first 10 coordinates and then the next 40, i.e. only variables in the set $B_1 \cup B_2$ or the support of the loading vector $v_1$. Variables numbered 51 to 100 (set $B_3$), although critical for prediction of the response $y = 2u_1$, are uncorrelated with it (as $u_1$ and $u_2$ are uncorrelated) and are thus not chosen, even in the limit $n \to \infty$. In contrast, even in the presence of moderate noise and a relatively small sample size of $n = 100$, the treelet algorithm correctly joins together the subsets of variables 1-10, 11-50, 51-100 and 201-400 (i.e. variables in the sets $B_1, B_2, B_3, B_4$). The rest of the variables, which contain only noise are combined only at much higher levels in the treelet algorithm, as they are asymptotically uncorrelated. Therefore, using only coarse-grained sum variables in the treelet transform yields near optimal prediction errors. In Fig. 4 (right) we plot the mean squared
error of prediction (MSEP) for 20 different simulations tested on an independent test set of 500 observations. The different methods are PLS on all variables (MSEP=0.17), supervised PLS with variable selection as in Bair et al. (2006) (MSEP=0.09), PLS on the 50 treelet features with highest variance, with the level of the treelet determined by leave-one-out cross validation (MSEP=0.035), and finally PLS on the projection of the noisy data onto the true vectors $v_i$ (MSEP = 0.030). In all cases, the optimal number of PLS projections (latent variables) is also determined by leave-one-out cross validation. Due to the high dimensionality of the data, choosing a subset of the original variables performs better than full-variable methods. However, choosing a subset of treelet features performs even better yielding an almost optimal prediction error ($\sigma^2/\|v_y\|^2 \approx 0.03$); compare the green and red curves in the figure.

5 Examples with Real Data Sets

5.1 A Classification Example with an Internet Advertisement Data Set

To illustrate the applicability of treelets on real data sets, we here study classification of an internet advertisement data set from the UCI ML repository (Kushmerick, 1999). After removal of the first three continuous variables, this data set contains 1555 binary variables and 3279 observations, labeled as belonging to one of two classes. The goal is to predict whether a new observation (an image in an internet page) is an internet advertisement or not, given values of its 1555 variables (various features of the image).

With standard classification algorithms, one can easily obtain a generalization error of about 5%. The first column in Table 1 labeled “full data set” shows the misclassification error rate for linear discriminant analysis (LDA), with the additional assumption of a diagonal covariance matrix, and for 1-nearest neighbor (1-NN) classification. The average is taken over 25 randomly selected training and test sets, with 3100 and 179 observations each.

The internet-ad data set, however, has several distinctive properties that are clearly revealed if
Table 1: Classification test errors for an internet advertisement data set

| classifier | full data set (1555 variables) | reduced data set (760 variables) | compressed representation with treelet features |
|------------|---------------------------------|----------------------------------|-----------------------------------------------|
|            | 5.5%                            | 5.1%                            | 4.5%                                          |
| LDA        | 4.0%                            | 4.0%                            | 3.7%                                          |

one applies the treelet algorithm as a feature extraction step prior to learning: First of all, several of the original variables are exactly linearly related. As the data is binary (-1 or 1), these variables are either identical or with opposite values. In fact, one can reduce the dimensionality of the data from 1555 to 760 without loss of information. The second column in the table labeled “reduced data set” shows the decrease in error rate after such a lossless compression. Furthermore, of these remaining 760 variables, many are highly related, with groups or clusters of similar variables. The treelet algorithm automatically identifies these clusters, as the algorithm effectively orders these variables bottom-up in the tree and defines treelets on nested clusters of similar variables. Fig. 5, left, shows the correlation matrix of the first 200 out of 760 variables in the order they are given. To the right, we see the correlations of the first 200 variables that are combined and sorted by the treelet transform.

In fact, an analysis of the reduced data set with 760 variables shows that there are more than 200 distinct pairs of variables with a correlation coefficient larger than 0.95. Not surprisingly, as shown in the right column of Table 1, treelets not only reduce the dimensionality but also increase the predictive performance on this dataset, yielding results competitive with other feature selection methods in the literature (Zhao and Liu, 2007). All results in Table 1 are averaged over 25 different simulations. The treelet results were achieved by projecting the data onto all coarse-grained treelets, at a level $L$ found by 10-fold cross-validation and a minimum prediction error criterion.
5.2 Classification and Analysis of DNA Microarray Data

We conclude with an application to DNA microarray data. In the analysis of gene expression, many methods first identify groups of highly correlated variables and then choose a few representative genes for each group (a so called gene signature). The treelet method also identifies sets of genes that exhibit similar expression patterns, but in contrast, replaces each such group by a coarse-grained variable that encodes the information from all variables in that group. Apart from the benefits of dimensionality reduction, such a representation also typically regularizes the data; this regularization could improve the performance of regression and classification algorithms (as illustrated below).

Another advantage to our method is that it is multi-scale. The treelet transform constructs orthogonal basis functions that are supported on nested clusters in a hierarchical tree. The idea of a hierarchical basis for microarray data is novel, but the benefits of hierarchical clustering in exploring and visualizing microarray data are well recognized in the field (Eisen et al., 1998; Tibshirani et al., 1999). It is, for example, known that a hierarchical clustering (or dendrogram) of
genes can sometimes reveal interesting clusters of genes worth further investigation. Similarly, a dendrogram of samples may identify cases with similar medical conditions. The treelet algorithm automatically provides such a re-arrangement and interpretation of the data.

We illustrate our method on the leukemia data set of Golub et al. (1999). This data monitor expression levels for 7129 genes and 72 patients, suffering from acute lymphoblastic leukemia (ALL, 47 cases) or acute myeloid leukemia (AML, 25 cases). The data are known to have a low intrinsic dimensionality, with groups of genes having similar expression patterns across samples (cell lines). The data are available at http://www.genome.wi.mit.edu/MPR, and include a training set of 38 samples and a test set of 34 samples.

Prior to analysis, we use a standard two-sample t-test to select the most “significant” genes, i.e. we choose a subset of genes that are differentially expressed in the two leukemia types. Such a pre-screening is common practice in microarray data analysis; here we use it so we can compare our results to benchmark results in Zou and Hastie (2005) for some common classification algorithms. Thus, using the training data, we perform a full (i.e. maximum height) treelet decomposition of the $p = 1000$ most significant genes. We sort the treelets according to their energy content (Eq. 5) on the training samples, and project the test data onto the $K$ treelets with the highest energy score. The reduced data representation of each sample (from $p$ genes to $K$ features) is then used to classify the samples into the two leukemia types, ALL or AML. We examine two different classification schemes:

In the first case, we apply a linear Gaussian classifier (LDA). As in Sec. 5.1, the treelet transform serves as a feature extraction and dimensionality reduction tool prior to classification. The appropriate value of the dimension $K$ is chosen by 10-fold cross-validation (CV). We divide the training set at random into 10 approximately equal-size parts, perform a separate $t$-test in each fold, and choose the $K$-value that leads to the smallest CV classification error (Fig. 6, left).

In the second case, we classify the data using a treelet “two-way clustering” scheme — we first compute treelets on the genes, and then we compute treelets on the samples. As before, each
Table 2: Leukemia misclassification rates; courtesy of Zou and Hastie (2005).

| Method                          | Ten-fold CV error | Test error |
|---------------------------------|-------------------|------------|
| Golub et al. (1999)             | 3/38              | 4/34       |
| Support vector machines         | 2/38              | 1/34       |
| Nearest shrunken centroids      | 2/38              | 2/34       |
| Penalized logistic regression   | 2/38              | 1/34       |
| Elastic nets                    | 3/38              | 0/34       |
| LDA on treelet features         | 2/38              | 3/34       |
| Treelets and two-way clustering | 0/38              | 1/34       |

sample (patient) is represented by $K$ treelet features instead of the $p$ original genes. The dimension $K$ is chosen by cross-validation on the training set. However, instead of applying a standard classifier, we construct new treelets on the samples using the reduced $N \times K$ data matrix. The two main branches of the associated dendrogram divide the $N$ samples into two classes; the classes are labeled using the training data and a majority vote. Such a two-way clustering procedure – of both genes and samples – seems to lead to classification results competitive with other algorithms; see Fig. 6 right, and Table 2. Moreover, the proposed method returns orthogonal basis functions that contain potentially useful information on groupings of genes or samples.

Fig. 7 (left) displays the original microarray data, with rows (genes) and columns (samples) ordered according to a hierarchical two-way clustering with treelets. The graph to the right shows the three maximum energy treelets on ordered samples. Note that the loadings are small for the two cases that are misclassified. The results for the treelets on genes are similar. The key point is that whenever there is a group of highly correlated variables (genes or samples), the algorithm tends to choose a coarse-grained variable for that whole group (see, for example, “Treelet 3” in the figure).

6 Conclusions

In this paper we presented treelets – a novel construction of a multi-scale orthonormal basis for unstructured or unordered data, which has many potential applications in dimensionality reduc-
Figure 6: Number of misclassified cases as a function of the number of treelet features. **Left:** LDA on treelet features; ten-fold cross-validation gives the lowest misclassification rate (2/38) for \( K = 3 \) treelets; the test error rate is then 3/34. **Right:** Two-way clustering scheme with treelets on both genes and samples; the lowest CV misclassification rate (0/38) is for \( K = 4 \); the test error rate is then 1/34.

The authors would like to thank P. Bickel, R.R. Coifman and S. Lafon for.

Acknowledgments: The authors would like to thank P. Bickel, R.R. Coifman and S. Lafon for
helpful discussions, and K. Roeder for comments that improved the manuscript. The research of ABL and LW was funded in part by NSF award CCF-0625879, and the research of BN was supported by the Hana and Julius Rosen fund and by the William Z. and Eda Bess Novick Young Scientist fund.

A Proof of Theorem 1

Let $x = (x_1, \ldots, x_p)^T$ be a random vector with distribution $F$ and covariance matrix $\Sigma = \Sigma_F$. Let $\rho_{ij}$ denote the correlation between $x_i$ and $x_j$. Let $x^1, \ldots, x^n$ be a sample from $F$, and denote the sample covariance matrix and sample correlations by $\hat{\Sigma}$ and $\hat{\rho}_{ij}$. Let $S_p$ denote all $p \times p$ covariance matrices. Let

$$\mathcal{F}_n(b) = \{ F : \Sigma_F \text{ is positive definite}, \min_{1 \leq j \leq p_n} \sigma_j \geq b \}.$$  

Any of the assumptions (A1a), (A1b) or (A1c) are sufficient to guarantee certain exponential inequalities.
Lemma 2  There exist positive constants $c_1, c_2$ such that, for every $\epsilon > 0$,

$$
\mathbb{P}(||\hat{\Sigma}_{jk} - \Sigma_{jk}||_\infty > \epsilon) \leq c_1p^2n^2e^{-nc_2\epsilon^2}.
$$

(29)

Hence,

$$
||\hat{\Sigma}_{jk} - \Sigma_{jk}||_\infty = O_P\left(\sqrt{\log \frac{n}{n}}\right).
$$

Proof. Under (A1), (29) is an immediate consequence of standard exponential inequalities and the union bound. The last statement follows by setting $\epsilon_n = K\sqrt{\log \frac{n}{n}}$ for sufficiently large $K$ and applying (A2). □

Lemma 3  Assume either that (i) $x$ is multivariate normal or that (ii) $\max_{1 \leq j \leq p}|x_j| \leq B$ for some finite $B$ and $\min_j \sigma_j \geq b > 0$. Then, there exist positive constants $c_3, c_4$ such that, for every $\epsilon > 0$,

$$
\mathbb{P}(\max_{jk} |\hat{\rho}_{jk} - \rho_{jk}| > \epsilon) \leq c_3p^2e^{-nc_4\epsilon^2}.
$$

(30)

Proof. Under normality this follows from Kalisch and Bühlmann (2007). Under (ii) note that

$$
h(\sigma_1, \sigma_2, \sigma_{12}) = \frac{\sigma_{12}}{\sigma_1 \sigma_2}
$$

satisfies

$$
|h(\sigma_1, \sigma_2, \sigma_{12}) - h(\sigma_1', \sigma_2', \sigma_{12}')| \leq \frac{3 \max\{|\sigma_1 - \sigma_1'|, |\sigma_2 - \sigma_2'|, |\sigma_{12} - \sigma_{12}'|\}}{b^2}.
$$

The result then follows from the previous lemma. □

Let $J_{\theta}$ denote the $2 \times 2$ rotation matrix of angle $\theta$. Let

$$
J_{\Sigma} = \begin{pmatrix}
\cos(\theta(\Sigma)) & -\sin(\theta(\Sigma)) \\
\sin(\theta(\Sigma)) & \cos(\theta(\Sigma))
\end{pmatrix}
$$

(31)

denote the Jacobi rotation where

$$
\theta(\Sigma) = \frac{1}{2}\tan^{-1}\left(\frac{2\Sigma_{12}}{\Sigma_{11} - \Sigma_{22}}\right).
$$

(32)
Lemma 4 Let $F$ be a bivariate distribution with $2 \times 2$ covariance matrix $\Sigma$. Let $J = J_\Sigma$ and $\hat{J} = J_{\hat{\Sigma}}$. Then,

$$\mathbb{P}(|\hat{J}^T \hat{\Sigma} \hat{J} - J^T \Sigma J|_\infty > \epsilon) \leq c_5 p^2 e^{-nc_6 \epsilon^2}. \quad (33)$$

**Proof.** Note that $\theta(\Sigma)$ a bounded, uniformly continuous function of $\Sigma$. Similarly, the entries of $J_\theta$ are also bounded, uniformly continuous functions of $\Sigma$. The result then follows from (29). ■

For any pair $(\alpha, \beta)$, let $\theta(\alpha, \beta)$ denote the angle of the principal component rotation and let $J(\alpha, \beta, \theta)$ denote the Jacobi rotation on $(\alpha, \beta)$. Define the selection operator

$$\Delta : \mathcal{S}_p \to \{(j, k) : 1 \leq j < k \leq p\}$$

by $\Delta(\Sigma) = (\alpha, \beta)$ where $\rho_{\alpha,\beta} = \arg\max_{ij} \rho_{ij}$. In case of ties, define $\Delta(\Sigma)$ to be the set of pairs $(\alpha, \beta)$ at which the maximum occurs. Hence, $\Delta$ is multivalued on a subset $\mathcal{S}_p^* \subset \mathcal{S}_p$ of measure 0.

The one-step treelet operator $T : \mathcal{S}_p \to \mathcal{S}_p$ is defined by

$$T(\Sigma) = \left\{ J^T \Sigma J : J = J(\alpha, \beta, \theta(\alpha, \beta)), (\alpha, \beta) \in \Delta(\Sigma) \right\}. \quad (34)$$

Formally, $T$ is a multivalued map because of potential ties.

**Proof of Theorem.** The proof is immediate from the lemmas. For the matrices $\hat{\Sigma}_n$, we have that $||\hat{\Sigma}_n - \Sigma||_\infty < \delta_n$ except on a set $A^*_n$ of probability tending to 0 at rate $O(n^{-(K-2c)})$. Hence on the set $A_n = \{ \hat{\Sigma}_n : ||\hat{\Sigma}_{n,b} - \hat{\Sigma}_n||_\infty < \delta_n \}$, we have that $T(\hat{\Sigma}_n) \in \mathcal{T}_n(\Sigma)$. The same holds at each step. ■

**B Proof of Lemma.**

Consider first the case, where at each level in the tree, the treelet operator combines a coarse-grained variable with a singleton according to $\{ \{x_1, x_2\}, x_3\}, \ldots$. Let $s_0 = x_1$. For $\ell = 1$, the $2 \times 2$ covariance submatrix $\Sigma^{(0)} \equiv \mathbb{V}\{(s_0, x_2)\} = \sigma_1^2 \begin{pmatrix} 1 & 1 \\ 1 & 1 \end{pmatrix}$. A principal component
analysis of $\Sigma^{(0)}$ gives $\theta_1 = \pi/4$ and $s_1 = \frac{1}{\sqrt{2}}(x_1 + x_2)$. By induction, for $1 \leq \ell \leq p - 1$, $\Sigma^{(\ell-1)} \equiv \mathcal{V}\{(s_{\ell-1}, x_{\ell+1})\} = \sigma_1^2 \left( \frac{\ell}{\sqrt{\ell}} \begin{array}{c} \sqrt{\ell} \\ 1 \end{array} \right)$. PCA on $\Sigma^{(\ell-1)}$ gives the (unconstrained) rotation angle $\theta_\ell = \arctan \sqrt{\ell}$, and the new sum variable $s_\ell = \frac{1}{\sqrt{\ell+1}} \sum_{i=1}^{\ell+1} x_i$.

More generally, at level $\ell$ of the tree, the treelet operator combines two sum variables $u = \frac{1}{\sqrt{m}} \sum_{i \in A_u} x_i$ and $v = \frac{1}{\sqrt{n}} \sum_{j \in A_v} x_j$, where $A_u, A_v \subseteq \{1, \ldots, p\}$ denote two disjoint index subsets with $m = |A_u|$ and $n = |A_v|$ number of terms, respectively. The $2 \times 2$ covariance submatrix

$$
\Sigma^{(\ell-1)} \equiv \mathcal{V}\{(u, v)\} = \sigma_1^2 \left( \frac{m}{\sqrt{mn}} \begin{array}{c} \sqrt{mn} \\ n \end{array} \right). \tag{35}
$$

The correlation coefficient $\rho_{uv} = 1$ for any pair $(u, v)$; thus, the treelet operator $T_\ell$ is a multivariate function of $\Sigma$. A principal component analysis of $\Sigma^{(\ell-1)}$ gives the eigenvalues $\lambda_1 = m + n$, $\lambda_2 = 0$, and eigenvectors $e_1 = \frac{1}{\sqrt{m+n}}(\sqrt{m}, \sqrt{n})^T$, $e_2 = \frac{1}{\sqrt{m+n}}(-\sqrt{n}, \sqrt{m})^T$. The rotation angle

$$
\theta_\ell = \arctan \sqrt{\frac{m}{n}}. \tag{36}
$$

The new sum and difference variables at level $\ell$ are given by

$$
\begin{align*}
    s_\ell &= \frac{1}{\sqrt{m+n}}(\sqrt{m}u + \sqrt{n}v) = \frac{1}{\sqrt{m+n}} \sum_{i \in \{A_u, A_v\}} x_i \\
    d_\ell &= \frac{1}{\sqrt{m+n}}(-\sqrt{m}u + \sqrt{mn}) = \frac{1}{\sqrt{m+n}} \left(-\sqrt{\frac{m}{n}} \sum_{i \in A_u} x_i + \sqrt{\frac{n}{m}} \sum_{j \in A_v} x_j\right) \tag{37}
\end{align*}
$$

The results of the lemma follow.

### C Proof of Theorem 2

Assume that variables from different blocks have not been merged for levels $\ell' < \ell$, where $1 \leq \ell \leq p$. From Lemma 1, we then know that any two sum variables at the preceding level $\ell - 1$ have the general form $u = \frac{1}{\sqrt{m}} \sum_{i \in A_u} x_i$ and $v = \frac{1}{\sqrt{n}} \sum_{j \in A_v} x_j$, where $A_u$ and $A_v$ are two disjoint index subsets with $m = |A_u|$ and $n = |A_v|$ number of terms, respectively. Let $\delta_k = \sigma_k/\sigma$. If $A_u \subseteq B_i$ and $A_v \subseteq B_j$, where $i \neq j$, i.e. the subsets belong to different blocks, then

$$
\Sigma^{(\ell-1)} = \mathcal{V}\{(u, v)\} = \left( \frac{m\sigma_i^2}{\sqrt{mn\sigma}} \begin{array}{c} \sqrt{mn\sigma} \\ \sigma_j^2 \end{array} \right) + \sigma^2 I. \tag{38}
$$
The corresponding “between-block” correlation coefficient

\[ \rho_B^{(\ell-1)} = \frac{\sigma_{ij}}{\sqrt{m + \delta_i^2 n + \delta_j^2}} \leq \frac{\sigma_{ij}}{\sqrt{m_n n + \delta_i^2 n + \delta_j^2}} \]  

with equality (“worst-case scenario”) if and only if \( \sigma = 0 \).

If \( \mathcal{A}_u, \mathcal{A}_v \subset B_k \), i.e. the subsets belong to the same block, then

\[ \Sigma^{(\ell-1)} = \mathbb{V}\{(u, v)\} = \sigma_k^2 \left( \frac{m}{\sqrt{mn}} \sqrt{\frac{mn}{n}} \right) + \sigma^2 I. \]  

The corresponding “within-block” correlation coefficient

\[ \rho_W^{(\ell-1)} = \frac{1}{\sqrt{1 + \frac{m+n}{mn} \delta_k^2 + \frac{1}{mn} \delta_4^4}} \geq \frac{1}{\sqrt{1 + 3 \max(\delta_k^2, \delta_4^4)}} \]  

with the “worst-case scenario” occurring when \( m = n = 1 \), i.e. when singletons are combined. Finally, the main result of the theorem follows from the bounds in Eq. 39 and Eq. 41 and the fact that

\[ \max \rho_B^{(\ell-1)} < \min \rho_W^{(\ell-1)} \]  

for \( \ell = 1, 2, \ldots, p - K \) is a sufficient condition for not combining variables from different blocks.

If the inequality Eq. 13 is satisfied, then the coefficients in the treelet expansion have the general form in Eq. 37 at any level \( \ell \) of the tree. With white noise added, the expansion coefficients have variances \( \mathbb{V}\{s_\ell\} = (m + n)\sigma_k^2 + \sigma^2 \) and \( \mathbb{V}\{d_\ell\} = \sigma^2 \frac{m^2 + n^2}{mn(m + n)} \). Furthermore, \( \mathbb{E}\{s_\ell\} = \mathbb{E}\{d_\ell\} = 0 \).

References

Asimov, D. (1985). The Grand Tour: A tool for viewing multidimensional data. SIAM Journal of Scientific Computing 6(1), 128–143.

Bair, E., T. Hastie, D. Paul, and R. Tibshirani (2006). Prediction by supervised principal components. Journal of the American Statistical Association 101(473), 119–137.

Belkin, M. and P. Niyogi (2005). Semi-supervised learning on riemannian manifolds. Machine Learning 56, 209–239.

Beran, R. and M. Srivastava (1985). Bootstrap tests and confidence regions for functions of a covariance matrix. The Annals of Statistics 13(1), 95—115.

39
Bickel, P. J. and E. Levina (2007). Regularized estimation of large covariance matrices. *The Annals of Statistics*. To appear.

Buckheit, J. and D. Donoho (1995). Improved linear discrimination using time frequency dictionaries. In *Proc. SPIE*, Volume 2569, pp. 540–551.

Candes, E. and T. Tao (2005). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Stat.*

Coifman, R., S. Lafon, A. Lee, M. Maggioni, B. Nadler, F. Warner, and S. Zucker (2005). Geometric diffusions as a tool for harmonics analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences* 102(21), 7426–7431.

Coifman, R. and N. Saito (1996). The local Karhunen-Loeve basis. In *Proc. IEEE International Symposium on Time-Frequency and Time-Scale Analysis*, pp. 129–132. IEEE Signal Processing Society.

Coifman, R. and M. Wickerhauser (1992). Entropy-based algorithms for best basis selection. In *Proc. IEEE Tran. Inf. Th.*, Volume 32, pp. 712–718.

Dettling, M. and P. Bühlmann (2004). Finding predictive gene groups from microarray data. *Journal of Multivariate Analysis* 90(1), 106–131.

Donoho, D. and M. Elad (2003). Maximal sparsity representation via $l_1$ minimization. *Proc. Nat. Acad. Sci.* 100, 2197–2202.

Donoho, D. and I. Johnstone (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association* 90, 1200–1224.

Eisen, M., P. Spellman, P. Brown, and D. Botstein (1998). Cluster analysis and display of genome-wide expression patterns. *Proceedings of the National Academy of Sciences* 95(25), 14863–14868.

Golub, G. and C. F. van Loan (1996). *Matrix Computations*. Johns Hopkins University Press.

Golub, T. R., D. K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J. P. Mesirov, H. Coller, M. L. Lob, J. R. Downing, M. Caliguri, C. Bloomfield, and E. Lander (1999). Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science* 286, 531–537.

Gruber, M. (1998). *Improving Efficiency by Shrinkage: The James-Stein and Ridge Regression Estimators*. Marcel Dekker, NY.

Guyon, I., J. Weston, S. Barnhill, and V. Vapnik (2002). Gene selection for cancer classification using support vector machines. *Machine Learning* 46, 389–422.

Hastie, T., R. Tibshirani, D. Botstein, and P. Brown (2001). Supervised harvesting of expression trees. *Genome Biology* 2(1), research0003.1–0003.12.

Hastie, T., R. Tibshirani, and J. Friedman (2001). *The Elements of Statistical Learning*. Springer.

Jain, A. K., M. N. Murty, and P. J. Flynn (1999). Data clustering: A review. *ACM Computing Surveys* 31(3), 264–323.

Johnstone, I. and A. Lu (2004). Sparse principal component analysis. *Journal of the American Statistical Association*. To appear.

Johnstone, I. M. (2001). On the distribution of the largest eigenvalue in principal component analysis. *The Annals of Statistics* 29(2), 295–327.
Jolliffe, I. T. (2002). *Principal Component Analysis*. Springer.

Kalisch, M. and P. Bühlmann (2007). Estimating high-dimensional directed acyclic graphs with the pc-algorithm. *Journal of Machine Learning Research* 8, 613–636.

Kushmerick, N. (1999). Learning to remove internet advertisements. In *Proceedings of the Third Annual Conference on Autonomous Agents*, pp. 175–181.

Lee, A. and B. Nadler (2007). Treelets — a tool for dimensionality reduction and multi-scale analysis of unstructured data. In *Proc. of the Eleventh International Conf. on Artificial Intelligence and Statistics*.

Levina, E. and J. Zhu (2007). Sparse estimation of large covariance matrices via a hierarchical lasso penalty. Submitted.

Lorber, A., K. Faber, and B. R. Kowalski (1997). Net analyte signal calculation in multivariate calibration. *Anal. Chemometrics* 69, 1620–1626.

Mallat, S. (1998). *A Wavelet Tour of Signal Processing*. Academic Press.

Meinshausen, N. and P. Bühlmann (2006). High-dimensional graphs and variable selection with the lasso. *The Annals of Statistics* 34(3), 1436–1462.

Murtagh, F. (2007). The Haar wavelet transform of a dendrogram. *Journal of Classification*. In Press.

Nadler, B. and R. Coifman (2005a). Partial least squares, Beer’s law and the net analyte signal: statistical modeling and analysis. *Journal of Chemometrics* 19, 45–54.

Nadler, B. and R. Coifman (2005b). The prediction error in CLS and PLS: the importance of feature selection prior to multivariate calibration. *Journal of Chemometrics* 19, 107–118.

Ogden, R. T. (1997). *Essential Wavelets for Statistical Applications and Data Analysis*. Birkhäuser.

Saito, N. and R. Coifman (1995). On local orthonormal bases for classification and regression. In *On local orthonormal bases for classification and regression*, pp. 1529–1532. IEEE Signal Processing Society.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B, Methodological* 58, 267–288.

Tibshirani, R., T. Hastie, M. Eisen, D. Ross, D. Botstein, and P. Brown (1999). Clustering methods for the analysis of DNA microarray data. Technical report, Department of Statistics, Stanford University.

Tibshirani, R., T. Hastie, B. Narasimhan, and G. Chu (2002). Diagnosis of multiple cancer types by shrunken centroids of gene expression. *Proceedings of the National Academy of Sciences* 99(10), 6567–6572.

Whittaker, J. (2001). *Graphical Models in Multivariate Statistics*. Wiley.

Xu, R. and D. Wunsch (2005). Survey of clustering algorithms. *IEEE Tran. Neural Networks* 16(3), 645–678.

Zhao, Z. and H. Liu (2007). Searching for interacting features. In *Proceedings of the 20th International Joint Conference on AI (IJCAI-07)*.

Zhu, J. and T. Hastie (2004). Classification of gene microarrays by penalized logistic regression. *Biostatistics* 5, 427–444.
Zou, H. and T. Hastie (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B* 67(2), 301–320.

Zou, H., T. Hastie, and R. Tibshirani (2006). Sparse principal component analysis. *Journal of Computational and Graphical Statistics* 15(2), 265–286.