The Annals of Applied Statistics
2008, Vol. 2, No. 2, 472–473
DOI: 10.1214/08-AOAS137A
Main article DOI: 10.1214/07-AOAS137

DISCUSSION OF: TREELETS—AN ADAPTIVE MULTI-SCALE BASIS FOR SPARSE UNORDERED DATA

BY FIONN MURTAGH

University of London

The work of Lee et al. is theoretically well founded and thoroughly motivated by practical data analysis. The algorithm presented has the following important properties:

1. Hierarchical clustering using a novel, adaptive, eigenvector-related, agglomerative criterion.
2. Principal components analysis carried out locally, leading to the required sample size for consistency being logarithmic rather than linear; and computational time being quadratic rather than cubic.
3. Multiresolution transform with interesting characteristics: data-adaptive at each node of the tree, orthonormal, and the tree decomposition itself is data-adaptive.
4. Integration of all of the following: hierarchical clustering, dimensionality reduction, and multiresolution transform.
5. Range of data patterns explored, in particular, block patterns in the covariances, and "model" or pattern contexts.

While I admire the work of the authors, nonetheless I have a different point of view on key aspects of this work:

1. The highest dimensionality analyzed seems to be 760 in the Internet advertisements case study. In fact, the quadratic computational time requirements (Section 2.1 of Lee et al.) preclude scalability. My approach in Murtagh (2007a) to wavelet transforming a dendrogram is of linear computational complexity (for both observations, and attributes) in the multiresolution transform. The hierarchical clustering, to begin with, is typically quadratic for the n observations, and linear in the p attributes. These computational requirements are necessary for the "small n, large p" problem which motivates this work (Section 1). In particular, linearity in p is a *sine qua non* for very high dimensionality data exploration.
   Since $L = O(p)$ in Section 2.1, this cubic time requirement has to be alleviated, in practice, through limiting L to a user-specified value.
2. The local principal components analysis (Section 2.1) inherently helps with data normalization, but it only goes some distance. For qualitative, mixed quantitative and qualitative, or other forms of messy data, I would use a correspondence analysis to furnish a Euclidean data embedding. This, then, can be the

Received October 2007; revised October 2007.

basis for classification or discrimination, benefiting from the Euclidean framework. See Murtagh (2005).

3. My final point is in relation to the following (Section 1): "The key property that allows successful inference and prediction in high-dimensional settings is the notion of sparsity." I disagree, in that sparsity of course can be exploited, but what is far more rewarding is that high dimensions are of particular *topology*, and not just data *morphology*.

   This is shown in the work of Hall et al. (2005), Ahn et al. (2007), Donoho and Tanner (2005) and Breuel (2007), as well as Murtagh (2004). What this leads to, potentially, is the exploitation of the remarkable simplicity that is concomitant with very high dimensionality: Murtagh (2007b). Applications include text analysis, in many varied applications, and high frequency financial and other signal analysis.

In conclusion, I thank the authors for their thought-provoking and motivating work.

REFERENCES

AHN, J., MARRON, J. S., MULLER, K. M. and CHI, Y.-Y. (2007). The high-dimension, low-sample-size geometric representation holds under mild conditions. *Biometrika* **94** 760–766.

BREUEL, T. M. (2007). A note on approximate nearest neighbor methods. Available at http://arxiv.org/pdf/cs/0703101.

DONOHO, D. L. and TANNER, J. (2005). Neighborliness of randomly-projected simplices in high dimensions. *Proc. Natl. Acad. Sci. USA* **102** 9452–9457. MR2168716

HALL, P., MARRON, J. S. and NEEMAN, A. (2005). Geometric representation of high dimension low sample size data. *J. Roy. Statist. Soc. B* **67** 427–444. MR2155347

MURTAGH, F. (2004). On ultrametricity, data coding, and computation. *J. Classification* **21** 167–184. MR2100389

MURTAGH, F. (2005). *Correspondence Analysis and Data Coding with R and Java*. Chapman and Hall/CRC, Boca Raton, FL. With a foreword by J.-P. Benzécri. MR2155971

MURTAGH, F. (2007a). The Haar wavelet transform of a dendrogram. *J. Classification* **24** 3–32. MR2370773

MURTAGH, F. (2007b). The remarkable simplicity of very high dimensional data: Application of model-based clustering. Available at www.cs.rhul.ac.uk/home/fionn/papers.

DEPARTMENT OF COMPUTER SCIENCE
ROYAL HOLLOWAY
UNIVERSITY OF LONDON
EGHAM, SURREY TW20 0EX
UNITED KINGDOM
E-MAIL: fmurtagh@acm.org

The Annals of Applied Statistics
2008, Vol. 2, No. 2, 474–477
DOI: 10.1214/08-AOAS137B
Main article DOI: 10.1214/07-AOAS137

DISCUSSION OF: TREELETS—AN ADAPTIVE MULTI-SCALE BASIS FOR SPARSE UNORDERED DATA

BY PETER J. BICKEL[1] AND YA'ACOV RITOV[2]

University of California and The Hebrew University of Jerusalem

We divide our comments on this very interesting paper into two parts following its own structure:

1. The use of treelets in connection with the correlation matrix of $\mathbf{X} = (X_1, \dots, X_p)^\mathsf{T}$ for which we have n i.i.d. copies, or as the authors refer to it, "unsupervised learning."
2. The use of treelets as a step in best fitting the linear regression of X_1 on $(X_2, \dots, X_p)^\mathsf{T}$.

1. Unsupervised learning. The authors' emphasis is on the method as a useful way of representing data analogous to a wavelet representation where $\mathbf{X} = \mathbf{X}(t)$ with t genuinely identified with a point on the line and observation at p time points, but where the time points have been permuted.

As such, this can be viewed as a clustering method which, from their examples, gives very reasonable answers. However, to make more general theoretical statements and to permit comparison to other methods, they necessarily introduce the model

$$\mathbf{X} = \sum_{j=i}^{K} U_j v_j + \sigma Z_j, \tag{1}$$

where $\mathbf{U} = (U_1, \dots, U_K)^\mathsf{T}$ is an unobservable vector, the v_j are fixed unknown vectors, and $\mathbf{Z} \sim N_p(0, J_p)$, where J_p is the identity, N_p is the p dimensional Gaussian distribution, and $\mathbf{U}, \mathbf{Z}$ are independent.

At this point, we are a bit troubled by the authors' analysis. We believe a key point, that is only stressed implicitly by the authors, is that the population tree structure, as defined, is only a function of the population covariance matrix. This is clear at Step 1, and follows since the Jacobi transformations depend only on the covariance and variances of the coordinates involved. This raises a problematic issue. If $\mathbf{U}$, and hence $\mathbf{X}$, has a Gaussian distribution, then the structure as postulated

Received November 2007; revised November 2007.

[1] Supported in part by NSF Grant DMS-06-05236.

[2] Supported in part by an ISF grant.

in (1) is not identifiable, as in known in factor analysis. Consider, for instance, Example 2. If we redefine $U_j^* = U_j$, $j = 1, 2$, $v_3^* = c_1 v_1 + c_2 v_2$, and $U_3^* = 0$, we are at the same covariance matrix as in (19) with only two nonoverlapping blocks.

The treelets transform evidently gives a decomposition attuned to the authors' beliefs of a block diagonal population structure with high intrablock correlation. But the theoretical burden of exhibiting classes of covariance matrices, other than ones whose eigenvectors are not only orthogonal but have disjoint support, and for which some version of sparse PCA cannot be utilized just as well, remains.

This is an insurmountable problem for any population parameter which is a function only of the covariance matrix.

A second difficulty, special to the treelets parameter $T(\Sigma)$, is that it is not defined uniquely for Σ for which the maximal off diagonal correlation is not uniquely assumed. This is reflected in the authors' discussion in Section 3.1 of the possible instability of the empirical tree. In this context, we don't understand their statement that inferring $T(\Sigma)$ is not the goal. If not, what is?

This issue makes comparison to the other methods difficult. As they state any of the several methods for sparse PCA, for example, d'Aspremont et al. (2007), Johnstone and Lu (2008), would yield the same answer as theirs for their Example 1.

But is there a way of proceeding which teases out explicitly structures such as in (19) without limiting oneself to the covariance matrix? Suppose that we can write $\mathbf{U} = B\mathbf{e}$, where $\mathbf{e} = (e_1, \ldots, e_K)^\mathsf{T}$ is a vector of independent not necessarily identically distributed variables, such that *at most one of them is Gaussian*. That is, we assume the factor loading themselves are obtained structurally. Then we can write for $i = 1, \ldots, n$, $j = 1, \ldots, p$, $X_{ij} = \sum_{l=1}^{K} c_{jl} e_{il} + \sigma Z_{ij}$, where $C = [C_{jl}]$ is a $p \times K$ matrix, the Z_{ij} are i.i.d. $N(0, 1)$, and $\mathbf{e}_i = (e_{i1}, \ldots, e_{iK})^\mathsf{T}$ are independent as above. Here, $C = VB$, where $V = (v_1, \ldots, v_k)$. We conjecture that if $p, n \to \infty$ with K fixed, and the columns of C are sparse, we can recover C up to a scale multiple of each row, and a permutation of the columns. Work on this conjecture is in progress.

2. Supervised learning. Can we select variables based on the X, the predictor variables, themselves? The tempting answer is yes (e.g., using PCA). The theoretical answer is no (Y can be a function of each component). The practical answer is at most a cautious yes; cf. Cook (2007) for a recent discussion. However, one should be careful to justify working with the predictions without the Y, since current regression methods permit one to handle models with almost exponentially many variables.

The LASSO type of estimator can handle sparse models. However, sparsity is an elusive property, since the LASSO can deal with sparsity in a given basis, while a sparse representation may exist only in some other basis. Treelets are proposed as a method which enriches the description of the model, and gives the user an over-rich collection of vectors which span the Euclidean space. Hopefully the tree cluster features are rich enough so the model can be approximated by the linear span of relatively few, say, no more than $o(n/\log n)$ terms.

The suggested algorithm deals with complexity by serial optimization in a fashion similar to standard model selection methods (e.g., forward selection), boosting, etc. It is not clear to us why the authors select the variables from one level and not from their union, since again modern methods can deal with any polynomial number of regressors.

To asses performance of the algorithm, we considered a simple version of the authors' supervised errors-in-variables model, but in an asymptotic setting. Suppose we observe n i.i.d. replicates from the distribution of $(Y, X_1, \dots, X_p)$, where $p = p_n$ and

$$\begin{aligned} Y &= \gamma Z + \varepsilon, \\ X_i &= c_p Z + \eta_i, \qquad i = 1, \dots, p, \end{aligned}$$

where $\varepsilon, Z \sim N(0, 1)$, $\eta_i \sim N(0, \sigma_i^2)$, all independent. This is a classical error in variables model, where the X_i are independent observations on Z_i and the best predictor is given by

$$\hat{y}(X) = \frac{\gamma c_p}{1 + c_p^2 \sum_{i=1}^p \sigma_i^{-2}} \sum_{i=1}^p \sigma_i^{-2} X_i.$$

Consider first $c_p = p^{-1/2}$, with all $\sigma_i = 1$, $\gamma \neq 0$ and, in particular, $c_p^2 \times \sum_{i=1}^p \sigma_i^{-2} = 1$. In this case all variables are interesting, and have the same weight for prediction. However, the covariance matrix of X has all diagonal terms greater than 1, and all off diagonal terms are p^{-1}. This model is not sparse—for instance, in the sense of El Karoui (2008), and is also inaccessible to regularized covariance estimation. The Treelet Algorithm will not be able to find this term. This model is significantly different from the null, and a consistent predictor exists given known parameter values. However, no standard general purpose algorithm will be able to deal with this model. A small set of simulations show that, in fact, there is a range of values of c_p for which PCA works better than treelets. However, for larger values of c_p, treelets work surprisingly well.

The restriction to a basis of a relatively small collection of transform variables is a limitation. In Bickel, Ritov and Tsybakov (2008) a general methodology was suggested for construction of a rich collection of basis functions. Formally, we consider the following hierarchical model selection method. For a set of functions $\mathcal{F}$ with cardinality $|\mathcal{F}| \geq K$, let $\mathcal{MS}_K$ be some procedure to select K functions out of $\mathcal{F}$. We denote by $\mathcal{MS}_K(\mathcal{F})$ the selected subset of $\mathcal{F}$, $|\mathcal{MS}_K(\mathcal{F})| = K$, $K = n^\gamma$ for some $\gamma < \infty$. Define $f \oplus g$ to be the operator combining two base variables, for instance, multiplication. The procedure is defined as follows:

(i) Set $\mathcal{F}_0 = \{X_1, \dots, X_p\}$.

(ii) For $m = 1, 2, \ldots$, let

$$\mathcal{F}_m = \mathcal{F}_{m-1} \cup \{f \oplus g : f, g \in \mathcal{MS}_K(\mathcal{F}_{m-1})\}.$$

(iii) Continue until convergence is declared. The output of the algorithm is the set of functions $\mathcal{MS}_K(\mathcal{F}_m)$ for some m.

Bickel, Ritov and Tsybakov consider $f \oplus g = fg$, since they consider models with interaction. The treelets construction is similar to this one, with each step yielding two new functions, which result from PCA applied to a pair of variables. There is one essential difference between our approach and the treelets algorithm. We also keep at each step the complexity of the over-determined collection in check, but let the complexity increase with the increase with levels.

REFERENCES

D'ASPREMONT, A., EL GHAOUI, L., JORDAN, M. I. and LANCKRIET, G. R. G. (2007). A direct formulation for sparse PCA using semidefinite programming. *SIAM Rev.* **49** 434–448. MR2353806

BICKEL, P. J., RITOV, Y. and TSYBAKOV A. (2008). Hierarchical selection of variables in sparse high-dimensional regression. *J. Roy. Statist. Soc. Ser. B*. To appear.

COOK, R. D. (2007). Fisher Lecture: Dimension reduction in regression (with discussion). *Statist. Sci.* **22** 1–43.

JOHNSTONE, I. AND LU, A. (2008). Sparse principal component analysis. *J. Amer. Statist. Assoc.* To appear.

EL KAROUI, N. (2008). Operator norm consistent estimation of large dimensional sparse covariance matrices. *Ann. Statist.* To appear.

DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA
BERKELEY, CALIFORNIA 94720-3860
USA
E-MAIL: bickel@stat.berkeley.edu

DEPARTMENT OF STATISTICS
THE HEBREW UNIVERSITY OF JERUSALEM
JERUSALEM
ISRAEL
E-MAIL: yaacov.ritov@gmail.com

The Annals of Applied Statistics
2008, Vol. 2, No. 2, 478–481
DOI: 10.1214/08-AOAS137C
Main article DOI: 10.1214/07-AOAS137

DISCUSSION OF: TREELETS—AN ADAPTIVE MULTI-SCALE BASIS FOR SPARSE UNORDERED DATA

BY NICOLAI MEINSHAUSEN AND PETER BÜHLMANN

University of Oxford and ETH Zürich

We congratulate Lee, Nadler and Wasserman (henceforth LNW) on a very interesting paper on new methodology and supporting theory. Treelets seem to tackle two important problems of modern data analysis at once. For datasets with many variables, treelets give powerful predictions even if variables are highly correlated and redundant. Maybe more importantly, interpretation of the results is intuitive. Useful insights about relevant groups of variables can be gained.

Our comments and questions include: (i) Could the success of treelets be replicated by a combination of hierarchical clustering and PCA? (ii) When choosing a suitable basis, treelets seem to be largely an unsupervised method. Could the results be even more interpretable and powerful if treelets would take into account some supervised response variable? (iii) Interpretability of the result hinges on the sparsity of the final basis. Do we expect that the selected groups of variables will always be sufficiently small to be amenable for interpretation?

1. Treelets or hierarchical clustering combined with PCA.

A main part of the treelet algorithm achieves two main objectives:

(1) Variables are ordered in a hierarchical scheme. Highly correlated variables are typically "close" in the hierarchy.
(2) A basis on the tree is chosen. Each node of the tree is associated with a "sum" (and also a "difference" variable).

Clearly, treelets are more elegant than any method trying to achieve these two goals separately. As LNW write in Section 1: "The novelty and contribution of our approach is the simultaneous construction of a data-driven multi-scale orthogonal basis *and* a hierarchical cluster tree." We are left wondering, though, how different treelets are to the following scheme. First, variables are ordered in a hierarchical clustering scheme—for concreteness, under complete linkage and using similarities derived from absolute correlations as in (1). Second, a basis on the tree is found. For each node in the hierarchical clustering tree, the "sum" variable of the treelet algorithm would be replaced by the first PCA component of the variables represented by this node. Computationally, this scheme is clearly less efficient than the treelet algorithm, at least if implemented naively. Are there other benefits of taking steps (1) and (2) in one step as in the proposed treelet algorithm? It would

Received November 2007; revised November 2007.

be nice to see whether the tree structure of treelets differs substantially from a hierarchical cluster tree, and whether the treelets bases are very different from local PCA. Unfortunately, we did not obtain the treelet software from LNW, and that is the main reason why we did not pursue our own numerical experiments.

2. Supervised and unsupervised basis selection. In addition to contributions (1) and (2), treelets involve an additional step:

(3) Cut the hierarchical tree at some height, and work with the resulting basis. The chosen height is based on a clever score function; see formula (6).

The choice of the cut-point influences the "resolution" at which one is looking at the data. At one extreme (the leaves of the tree, "high resolution"), all variables are individual basis vectors. At the other extreme (the root of the tree, "low resolution"), basis vectors contain contributions from all variables, just like in global PCA. We understand the motivation behind the approach and the reported results seem to be very favorable. For supervised problems with a response, we are wondering if information in the response variable could be used more extensively to construct the treelet basis.

It is clear that a response variable should influence the choice of the basis. Take an example. If the signal-to-noise ratio (SNR) is very low, then one might be more inclined to work with "low resolution," as there is no hope of recovering the regression coefficients of individual variables. On the other hand, for high SNRs, it might very well be possible to single out individual variables as important. Information in the response variable could be used in various ways. Ranging from weak use of the response to stronger involvement:

(a) *Supervised choice of the cutoff height.* The cutoff of the tree can be influenced by the response. In fact, LNW used some supervised score function in Section 5.1 and also some cross-validation (and hence, supervised) approach in Section 5.3 to choose the best value for K, which in turn determines the cutoff value for the tree through criterion (6). Another possibility for finding the best cutoff in a supervised fashion would be to choose, instead of (6),

$$B_L = \underset{B_\ell : 0 \le \ell \le p-1}{\arg\min} \; CV(B_\ell),$$

where $CV(B_\ell)$ is the cross-validated loss of a favorite prediction method, using the orthogonal basis B_ℓ as predictor variables. Is it better to choose a value of K, and having then an associated best K-basis, or should we rather choose a best basis directly? Note that with the latter, we would also select features from the basis if the prediction method would do variable selection, for example, the Lasso or tree-based methods including boosting or random forests.

(b) *Nonuniform cutoff height.* For a given tree, it is not obvious why cutting at a single height is necessarily optimal. As an example, take 2 predictor variables

x_i and x_j with $i \neq j$ who are quite correlated and both of them are strongly relevant for prediction. They will tend to be merged quite early in the tree, but we would like to keep them separate for interpretation and best predictive performance (while we would like to merge as early as possible less correlated clusters of variables that only have a weak influence on the response).

Instead of cutting the tree at a single height, it might be more advantageous to start toward the root node of the tree. If a given cluster of variables turns out to be important, one could try to add—in a forward selection manner—basis elements from its sub-clusters. If descending deeper into the tree at a particular node improves prediction considerably, one would keep descending and stop otherwise. The selected tree height would not be uniformly the same. The resolution would be high in directions of strong signal and low in directions of weak signal. For related procedures, see also Meinshausen (2008) or Goeman and Mansmann (2008). And also "supervised harvesting" [Hastie, Tibshirani, Botstein and Brown (2001)] has the property that features at different levels of a hierarchical cluster tree are selected.

(c) *Supervised tree growth.* Take again the example in (b) of two rather correlated predictor variables, who are merged quite early in the tree but contribute both strongly to the response. A more principled way of dealing with the issue would be to make the *construction* of treelets, that is, the tree and the bases, supervised. Is it possible? [Besides doing the obvious, viz., to include the response y as another variable, i.e., considering new data $\tilde{x} = (y, x)$.] To our knowledge, there are not many methods for "supervised grouping." It seems to us that among the references in LNW, only the method in Dettling and Bühlmann (2004) remains as "truly supervised," while the elastic net approach in Zou and Hastie (2005), which is supervised, is not extracting a group structure.

We think that it would be worthwhile to extend treelets in the direction of a truly supervised algorithm both for improved prediction performance and better interpretability.

3. Interpretability. One attractive property of treelets is the sparsity of the solution (sparsity is here to be understood as few variables entering a basis vector). Compared with global PCA, which includes contributions from all variables into every basis vector, treelet basis vectors contain in general only a few variables in each basis vector. This increases the interpretability of results dramatically.

There is clearly a tradeoff, though: increasing the sparsity increases interpretability by performing variable selection among the treelet features. Increasing sparsity increases at the same time, however, the variance of the solution. Making the results very sparse carries, in general, the risk that the results are unstable. We might see a completely different result on repeated measurements (or on repeated bootstrap samples). We would thus like to make the results "as sparse as possible, but not any sparser." A very sparse yet unstable result is not suitable for interpretation either.

What should we do if the selected groups of variables will be too large for interpretation? For example, groups of genes of size more than 20 are often an idea attractive to statisticians or computer scientists, but it is very likely that such large groups will never be validated by biological experiments. Is the solution as simple as cutting the tree at a level such that the group size is bounded by a value which is desired for a specific application?

Bounding the maximal group size can potentially render the algorithm unstable. As a possible solution to the sparsity–stability tradeoff, we can cut the tree at a height that gives maximal sparsity of results under the condition that the obtained groups of variables are—in some sense—stable under permutations of the data. LNW show in Figure 3 some bootstrap confidence bands which are supported by some asymptotic theory in Section 3.1. It would be interesting to have a more complete way of visualizing the stability of the treelet procedure.

4. Conclusions. We think that treelets are a very interesting and promising proposal for high-dimensional modern data analysis. Open-source high quality software would be desirable: it would help promoting the method to a large community of users and researchers and it would allow reproducibility of results.

REFERENCES

DETTLING, M. and BÜHLMANN, P. (2004). Finding predictive gene groups from microarray data. *J. Multivariate Anal.* **90** 106–131. MR2064938

GOEMAN, J. and MANSMANN, U. (2008). Multiple testing on the directed acyclic graph of Gene Ontology. *Bioinformatics* **24** 537–544.

HASTIE, T., TIBSHIRANI, R., BOTSTEIN, D. and BROWN, P. (2001). Supervised harvesting of expression trees. *Genome Biology* **2** 1–12.

MEINSHAUSEN, N. (2008). Hierarchical testing of variable importance. *Biometrika* **95** 265–278.

ZOU, H. and HASTIE, T. (2005). Regularization and variable selection via the elastic net. *J. Roy. Statist. Soc. Ser. B* **67** 301–320. MR2137327

DEPARTMENT OF STATISTICS
UNIVERSITY OF OXFORD
1 SOUTH PARKS ROAD
OXFORD OX1 3TG
UNITED KINGDOM
E-MAIL: meinshausen@stats.ox.ac.uk

SEMINAR FÜR STATISTIK
ETH ZÜRICH
LEONHARDSTRASSE 27
CH-8092 ZÜRICH
SWITZERLAND
E-MAIL: buhlmann@stat.math.ethz.ch

The Annals of Applied Statistics
2008, Vol. 2, No. 2, 482–483
DOI: 10.1214/07-AOAS137D
Main article DOI: 10.1214/07-AOAS137
© Institute of Mathematical Statistics, 2008

DISCUSSION OF: TREELETS—AN ADAPTIVE MULTI-SCALE BASIS FOR SPARSE UNORDERED DATA

BY ROBERT TIBSHIRANI

Stanford University

This is a very interesting paper on an important topic—the problem of extracting features in an unsupervised way from a dataset. There is growing evidence that unsupervised feature extraction can provide an effective set of features for supervised learning: see, for example, the interesting recent work on learning algorithms for Boltzmann machines [Hinton, Osindero and Teh (2006)].

The ideas in this paper are exciting—treelets are a neat construction that combine clustering and wavelets, and are simple enough to be theoretically tractible. The connection to the latent variable model is also interesting: this kind of model is also the basis of supervised principal components, a method that I co-developed recently [Bair et al. (2006)] for regression and survival analysis in the $p > N$ setting.

I have no practical experience with treelets, so my remaining comments will be brief and mostly in the form of questions for the authors. A much simpler approach to this problem would be to hierarchically cluster the predictors, and then take the average at every internal node of the dendrogram. Let's call this the "simple averaging" method. As noted by the authors, this has already been proposed in the literature, for example, in the "Tree-harvesting" procedure. In this approach we keep all of the original predictors and all of the internal node averages and so end up with an over-complete basis of $2p$ basis functions.

How are treelets different from simple averaging? Treelets do an orthogonalization after each node merge, but does this change the clustering in a material way? What advantage is there to the orthogonal basis delivered by treelets? After all, it looks like the resulting linear combinations of variables are not uncorrelated. Does the simple averaging method perform as well as treelets in the kind of examples of the paper? Do the authors' theorems apply to the simple averaging method as well, or are treelets uniquely good in their estimation of the components of a latent variable model?

The contrast between treelets and simple averaging is analogous to the contrast between wavelets and basis pursuit [Chen, Donoho and Saunders (1998)]. The former is an orthogonal basis while the latter is over-complete; when fitting is done with an L_1 (lasso) penalty, the over complete basis, can provide a very good predictive model.

Received December 2007; revised December 2007.

One small point—hierarchical clustering is usually done with average linkage between pairs of predictors. A variation, commonly used in genomics and sometimes called Eisen clustering (since it's implemented in Eisen's Cluster program), uses instead the distance (or correlation) between centroids. The Treelet construction looks more like Eisen clustering. The point is that one could apply Eisen clustering, and then simply average the predictors in every internal node.

REFERENCES

BAIR, E., HASTIE, T., PAUL, D. and TIBSHIRANI, R. (2006). Prediction by supervised principal components. *J. Amer. Statist. Assoc.* **101** 119–137. MR2252436

CHEN, S. S., DONOHO, D. L. and SAUNDERS, M. A. (1998). Atomic decomposition by basis pursuit. *SIAM J. Sci. Comput.* **20** 33–61. MR1639094

HINTON, G., OSINDERO, S. and TEH, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Comput.* **18** 1527–1554. MR2224485

DEPARTMENTS OF HEALTH RESEARCH & POLICY,
AND STATISTICS
STANFORD UNIVERSITY
STANFORD, CALIFORNIA 94305
USA
E-MAIL: tibs@stanford.edu

The Annals of Applied Statistics
2008, Vol. 2, No. 2, 484–488
DOI: 10.1214/08-AOAS137E
Main article DOI: 10.1214/08-AOAS137

DISCUSSION OF: TREELETS—AN ADAPTIVE MULTI-SCALE BASIS FOR SPARSE UNORDERED DATA

BY XING QIU

University of Rochester

This is a discussion of paper "Treelets—An adaptive multi-scale basis for sparse unordered data" by Ann B. Lee, Boaz Nadler and Larry Wasserman. In this paper the authors defined a new type of dimension reduction algorithm, namely, the treelet algorithm. The treelet method has the merit of being completely data driven, and its decomposition is easier to interpret as compared to PCR. It is suitable in some certain situations, but it also has its own limitations. I will discuss both the strength and the weakness of this method when applied to microarray data analysis.

1. The design of the treelet algorithm. A lot of modern technologies require analyzing noisy, high-dimensional and unordered data. As an example, in the field of microarray analysis, researchers are often interested in analyzing gene expessions sampled from n different subjects. These expression data can be seen as n independent realizatons of a p-dimensional random vector $\vec{x} = (x_1, \dots, x_p)^T$, each x_i represents (usually log tranformed) an expression level of a given gene. In practice, p (number of genes) is measured in thousands or tens of thousands, and n (sample size) is more than often less than a dozen. Due to this "large p, small n" nature, dimension reduction such as hierarchical clustering (denoted as HC henthforth) is often conducted prior to regression or classification analysis.

The treelet algorithm can be best described as a data driven local PCA (Principal Component Analysis). It can be summarized in the following steps:

1. Find the two most similar variables (genes) by a well-defined metric of similarity such as covariance. Denote this pair of genes as x_α, x_β.
2. Perform a local PCA on this pair to decorrelate them. More specifically, find a Jacobi rotation matrix J such that $x^{(2)} = J^T(x)$ has this property: $\operatorname{cov}(x_\alpha^{(2)}, x_\beta^{(2)}) = 0$. Then drop the less important one of them (the one with smaller variance) and update the similar matrix.

   In other words, after this step, a summary variable will be chosen to replace the two most similar variables from the original data.
3. Update the similarity matrix with this new summary variable and then find the next most similar pair of variables.

Received December 2007; revised December 2007.
Key words and phrases. Treelet, microarray, dependence.

4. Build up a multi-resolution analysis accordingly. At each step, we have a representation of x as the sum of the coarse-grained representation of the signal and the sum of the residuals.

2. Comparisons to other methods. Dimension reduction is not a new technique in data analysis. Principal component analysis [Jolliffe (2002)] and hierarchical clustering methods [Eisen et al. (1998), Tibshirani et al. (1999)] are among the most used methods in this arena.

PCA As pointed out by the authors, PCA computes a *global* representation of data. The principal components are linear combinations of all variables. This poses an obstacle for interpreting the results. On the other hand, the treelet method is a *local* method by design. For example, when the underlying dependence structure of data can be modeled as disjoint groups of variables which are uncorrelated to each other groupwise, in principle, local dimension reduction methods should perform better than their global counterpart.

HC In a sense, the treelet can be viewed as yet another way of constructing the dendrogram from the bottom up. So the treelet method is a legitimate member of the family of agglomerative hierarchical clustering algorithms. However, there is a novelty in the treelet method approach. By construction, at each step only the sum variable (the variable which contributes more variance) remains as the representative of the pair of closely related variables. At the end of the day, the dendrogram will reflect the "skeleton" of the given data rather than the dependence structure of the data themselves. If in a specific application we have evidence that the unused residual terms reflect nothing but noise, then the treelet method provides us invaluable information about hierarchical dependence of the data which is noise resistant.

3. Applicability in the field of microarray data analysis. As mentioned in Section 1, microarray data analysis is a good example where the treelet method may shine. It is a well-known biological fact that genes work together instead of independently. As a consequence, their expressions are highly correlated.

Storey and Tibshirani (2003) hypothesized that most likely the form of intergene dependence is *weak dependence*, which can be "...loosely described as any form of dependence whose effect becomes negligible as the number of features increases to infinity." And their argument is that genes can be grouped into essentially independent *pathways*.

If this hypothesis is true, then the treelet method would work beautifully, as illustrated in Chapter 3.2 of Lee, Nadler and Wasserman (2008).

However, a series of study conducted by Qiu et al. (2005a, 2005b, 2006) on St. Jude Children's Research Hospital Database (see sjcrh database on childhood leukemia) showed that, on average, the intergene correlation level is too high to be explained by the within pathway dependence (weak dependence) alone. There is strong long ranged *global* dependence between pathways. Whether this global dependence is due to technical noise or not is up to debate [Klebanov and Yakovlev (2007)]. If the observed high intergene correlation is due to biological reasons rather than noise, then the treelet method may be harmful since it will reduce and distort useful information contained in the dependence structure.

It is also interesting to compare the treelet method with various normalization methods, such as the global normalization [Yang et al. (2002)]. Apparently, global normalization (or any other normalization method) is not a dimension reduction procedure, nor does it give us a dendrogram. However, one similarity can be found between the global normalization and the treelet method: they both replace data variables with surrogate variables which are linear combinations of the original variables. In the case of global normalization (assuming expression levels are log transformed), the ith variable (gene) x_i is replaced by $x_i - \bar{x}$, where $\bar{x}$ is the sample average of x over all genes for a given slide. From this point of view, global normalization is a *global* basis transformation. A hidden assumption in doing global normalization is that $\bar{x}$ represents slide-specific noise thus needs to be removed from the observed signal. While I personally think that technical noise cannot be removed in such an overly simplistic way, it provides an example where a global method may better capture the most useful information at a much faster rate.

Another dangerous behavior of the treelet method is that it uses variance as a means to evaluate which variable should be retained (sum variable), and which one should be disregarded (difference variable). This approach may look very plausible mathematically, yet it ignores the possibility that genes with lesser variability may actually be the important ones. It may very well be the case that in evolution genes that are responsible for essential functionalities are more likely to have smaller variation than those less important ones.

One of the major advantage of the treelet method is that the sum variables it produces use only a subset of variables. This makes it easier to interpret than PCR, which gives linear combinations of *all* variables as outcome. However, the sum variables of the treelet method can also be linear combinations of *many* variables. It is a huge leap forward in the right direction, yet it is still hard to find its way into another important field of microarray analysis: testing differential expressions. Being hard to interpret is just an apparent disadvantage. A more subtle disadvantage is that there is no guarantee that the multiple testing procedures designed to work with original expressions still control the same false positive level when we replace them with some "noise-free" surrogate variables. Much future work can be done in this direction.

4. Discussion. Overall, I think the treelet method has the merit of being completely data driven and being local. I am very impressed by its performance when data variables are divided into uncorrelated groups.

However, when talking about its applicability to gene expression data, I think a lot of careful investigation still needs to be done. This is due to the complexity of the dependence structure exhibits in this type of data. This complexity is probably the reason why the treelet method (in its original form) did not outperform other classification methods on the leukemia data set of Golub et al.

In the future more attention should be paid to the nature of inter-pathway dependence. Should we model pathways as disjoint, uncorrelated "super variables"? Or should we also model some long range, inter-pathway correlation? I think this question can be answered only through joint efforts from both statisticians and biologists.

I also want to point out that I disagree with the authors in that PCA cannot reveal the underlying noiseless structure of the data while the treelet method can. As pointed out by numerous researchers [Storey et al. (2007), Barbujani et al. (1997), Akey et al. (2002), Rosenberg et al. (2002)], most human genetic variation is due to variation among individuals within a population rather than among populations. This implies that the majority of "noise" in the data is actually true biological information. So being too good at removing "noise" may not always be a merit.

In short, I believe there is no one-size-fits-all solution for noisy, high-dimensional data. The treelet method provides us a very good solution in some situations, and it opens many research possibilities in the future.

Possible future improvements:

- The leukemia data set of Golub et al. used for classification of DNA microarray data is not the largest data available. The authors may want to try St. Jude Children's Research Hospital Database on childhood leukemia too.
- In the same chapter, the authors claim that they use a novel "two-way treelet decomposition scheme." They first compute treelets on the genes, then compute treelets on the *samples*. It looks very suspicious. I have a feeling that the gained performance is due to some subtle violation of the principle of external cross-validation. The authors should definitely provide more details about this approach.
- A recent paper by Klebanov, Jordan and Yakovlev (2006) proposed a new model of the long range intergene correlation structure. In a loose way, they hypothesize that there exist "gene drivers" and "gene modulators," such that the expression of a "gene-modulator" is stochastically proportional to that of a "gene-driver" (without log transformation). It would be nice to see if the treelet method works in this situation.

REFERENCES

ST. JUDE CHILDREN'S RESEARCH HOSPITAL (sjcrh) DATABASE ON CHILDHOOD LEUKEMIA.

AKEY, J. M., ZHANG, G., ZHANG, K., JIN, L. and SHRIVER, M. D. (2002). Interrogating a high-density snp map for signatures of natural selection. *Genome Res.* **12** 1805–1814.

BARBUJANI, G., MAGAGNI, A., MINCH, E. and CAVALLI-SFORZA, L. L. (1997). An apportionment of human dna diversity. *Proc. Natl. Acad. Sci. USA* **94** 4516–4519.

EISEN, M., SPELLMAN, P., BROWN, P. and BOTSTEIN, D. (1998). Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA* **95** 14863–14868.

JOLLIFFE, I. T. (2002). *Principal Component Analysis*. Springer, New York. MR2036084

KLEBANOV, L., JORDAN, C. and YAKOVLEV, A. (2006). A new type of stochastic dependence revealed in gene expression data. *Stat. Appl. Genet. Mol. Biol.* **5** Article 7. MR2221298

KLEBANOV, L. and YAKOVLEV, A. (2007). How high is the level of technical noise in microarray data? *Biol. Direct.* **2** 9.

LEE, A. B., NADLER, B. and WASSERMAN, L. (2008). Treelets—An adaptive multi-scale basis for sparse unordered data. *Ann. Appl. Statist.* To appear.

QIU, X., BROOKS, A. I., KLEBANOV, L. and YAKOVLEV, A. (2005a). The effects of normalization on the correlation structure of microarray data. *BMC Bioinformatics* **6** 120.

QIU, X., KLEBANOV, L. and YAKOVLEV, A. (2005b). Correlation between gene expression levels and limitations of the empirical bayes methodology in microarray data. *Statist. Appl. Genet. Mol. Biol.* **4** Article 3. MR2183944

QIU, X. and YAKOVLEV, A. (2006). Some comments on instability of false discovery rate estimation. *J. Bioinformatics Computational Biology* **4** 1057–1068.

ROSENBERG, N. A., PRITCHARD, J. K., WEBER, J. L., CANN, H. M., KIDD, K. K., ZHIVOTOVSKY, L. A. and FELDMAN, M. W. (2002). Genetic structure of human populations. *Science* **298** 2381–2385.

STOREY, J. D. and TIBSHIRANI, R. (2003). Statistical significance for genomewide studies. *Proc. Nat. Acad. Sci. USA* **100** 9440–9445. MR1994856

STOREY, J. D., MADEOY, J., STROUT, J. L., WURFEL, M., RONALD, J. and AKEY, J. M. (2007). Gene-expression variation within and among human populations. *Am. J. Hum. Genet.* **80** 502–509.

TIBSHIRANI, R., HASTIE, T., EISEN, M., ROSS, D., BOTSTEIN, D. and BROWN, P. (1999). Clustering methods for the analysis of dna microarray data. Technical report, Dept. Statistics, Stanford Univ.

YANG, Y. H., DUDOIT, S., LUU, P., LIN, D. M., PENG, V., NGAI, J. and SPEED, T. P. (2002). Normalization for cdna microarray data: A robust composite method addressing single and multiple slide systematic variation. *Nucleic Acids Res.* **30** e15.

DEPARTMENT OF BIOSTATICS
AND COMPUTATIONAL BIOLOGY
UNIVERSITY OF ROCHESTER
601 ELMWOOD AVE
BOX 630
ROCHESTER, NEW YORK 14642
USA
E-MAIL: xqiu@bst.rochester.edu

The Annals of Applied Statistics
2008, Vol. 2, No. 2, 489–493
DOI: 10.1214/08-AOAS137F
Main article DOI: 10.1214/07-AOAS137

DISCUSSION OF: TREELETS—AN ADAPTIVE MULTI-SCALE BASIS FOR SPARSE UNORDERED DATA

BY CATHERINE TUGLUS AND MARK J. VAN DER LAAN

University of California, Berkeley

We would like to congratulate Lee, Nadler and Wasserman on their contribution to clustering and data reduction methods for high p and low n situations. A composite of clustering and traditional principal components analysis, treelets is an innovative method for multi-resolution analysis of unordered data. It is an improvement over traditional PCA and an important contribution to clustering methodology. Their paper presents theory and supporting applications addressing the two main goals of the treelet method: (1) Uncover the underlying structure of the data and (2) Data reduction prior to statistical learning methods. We will organize our discussion into two main parts to address their methodology in terms of each of these two goals. We will present and discuss treelets in terms of a clustering algorithm and an improvement over traditional PCA. We will also discuss the applicability of treelets to more general data, in particular, the application of treelets to microarray data.

1. Uncover the underlying structure of the data.

In order to determine the underlying structure of a given data set, the statistician will often employ various clustering algorithms, or projection-based methods such as principal components analysis in an effort to tease apart the data which is often highly correlated and very noisy. The authors, Lee, Nadler and Wasserman, propose a new method targeted at detecting the multi-resolution internal structure of the data. In wavelet-fashion, the results are presented on multiple scales, providing detail only when necessary. However, unlike wavelet-analysis, their technique is applicable to unordered data. Though presented initially as an extension of wavelets, treelets are built upon a hierarchical clustering framework and can be illustrated as such.

As outlined in the overview van der Laan, Pollard and Bryan (2003), clustering methods are described by three major components: the distance measure, the grouping criteria, and the algorithm. The authors in this paper present treelets in terms of a correlation distance matrix, while we have argued for algorithms which allow arbitrary distance metrics since different applications can require different uses of the notion of proximity. Though they elude that other distance measures can be applied, all theory and simulation is presented and proven using a covariance or correlation measure of similarity. When alternate distance measures are used the benefit of using this method over other clustering methods seems questionable, and the final interpretation of the multi-resolution basis is unclear.

Received February 2008; revised February 2008.

When the underlying structure of the data does not reflect a sparse diagonal correlation matrix, using more adaptable clustering methods such as Hierarchical Partitioning and Collapsing Hybrid (HOPACH) [Pollard and van der Laan (2005), van der Laan and Pollard (2003)] would be more appropriate and seem to provide more flexibility and more interpretable results. HOPACH takes as input an arbitrary distance or dissimilarity matrix, combines top-down and agglomerative clustering into a hybrid algorithm, allows for data adaptively deciding on the number of children cluster in each node, orders the clusters in each layer of the hierarchical tree based on the distance so that neighboring clusters are close to each other w.r.t. the specified dissimilarity, and it allows the use of a data adaptive as well as visual criteria (including output of bootstrap) to decide on the depth and number of clusters in the tree.

The treelet algorithm is a binary agglomerative hierarchical clustering algorithm. In terms of a hierarchical graph only, the two most correlated nodes are combined at a given step. For an n by p data matrix, there are total $p - 1$ layers for a graph combined to completion. The binary combination allows for the multiresolution interpretability of the resulting basis. At each node a principal components analysis is applied to the pair of variables. The node is then represented by the two components, the first component becoming a “sum” variable, and the second the “difference” variable. Since only the sum variable is allowed to combine in higher levels of the graph, the difference variable remains behind as a residual measure of the combination. Each treelet, comprised of one node (sum variable) and its associated difference variables can be represented by a orthonormal basis.

The treelet method is applicable given any agglomerative hierarchical algorithm. However, the graph is solely built on the similarity between two variables. This does not take advantage of all information present in the data. Clustering algorithms have advanced beyond simple similarity measures and use informative measures such as the Mean Silhouette [Kaufman and Rousseeuw (1990)], the Median Silhouette, or the Split Mean/Median Silhouette [van der Laan, Pollard and Bryan (2003)]. Each of these grouping criteria reflects how similar variables are in relation to how dissimilar they are from others.

The authors do present a measure to determine the optional height of the tree, a normalized energy score reflecting the percent variance explained on a given basis conditional on the number of variables chosen to represent the treelet—the best K-dimensional basis. According to the authors, the best height and dimension K can be chosen using cross-validation - though the exact method of cross validation is not presented clearly in terms of choosing K. If the goal is to use treelets for the purpose of prediction, then this is easily defined, but it becomes unclear what is meant otherwise.

In terms of a clustering algorithm, we applaud the authors for having a well defined goal: estimation of the true correlation matrix. Generally cluster analysis, though built from localized structure, does not identify that as its far-reaching goal leaving consistency theory nonexistent. We would like to point out that in terms of clustering, a particular consistency theory for the estimation of the mean and covariance matrix based on Bernstein's inequality, as well as the sensitivity and reproducibility of the estimate based on bootstrap resampling, was presented in van der Laan and Bryan (2001) and subsequent articles.

Beyond a clustering interpretation, treelets can also be viewed as an improved robust version of PCA. Traditional PCA is a global method, highly sensitive to noise in the data. Treelets focus on detecting localized structure and by performing binary data-driven rotations, are much more robust to noise. The authors show the improved finite sample properties of treelets over traditional PCA, and we believe this is a fundamental contribution to the field. Treelets will be able to perform well in many practical settings, while PCA will often rely on too large sample sizes. Treelets also incorporate hierarchical clustering giving the method a wavelet-like property, preserving detailed structure in only the necessary region, unlike PCA which splits the data into orthogonal projections, each with a linear basis relating to the entire data set.

In terms of detecting the underlying structure of data given a sparse correlation matrix, treelets are a great contribution providing a new summary metric for binary clustering algorithms, and providing a localized PCA. In application, however, the method is potentially limited to only data where the underlying correlation structure is assumed to be sparse, such as many image and spatial analyses. Given a more complex correlation structure, which is often seen in biological data such as microarray data, treelets do not necessarily perform better than clustering or standard PCA. The improvement in convergence rate over PCA is contingent on the sparsity of the correlation matrix.

2. Data reduction. In terms of data reduction, treelets are a data-driven method which provides a more concise representation of a data matrix with sparse correlation. Reducing the dimension of the initial data set before applying a learning algorithm can improve the accuracy of the predictor. In the spirit of the superlearning approach [van der Laan, Polley and Hubbard (2007)], involving an aggressive approach for data adaptively selecting among a continuum of different strategies for construction of a prediction, for the purposes of dimension reduction in prediction, we recommend in practice that the height of the tree (L) and the dimension of the basis (K) should be chosen with respect to the cross-validated risk of the prediction in all applications. The authors elude to this.

The practical application of treelets as a dimension reduction technique for high-dimensional microarray data is unclear. Microarray data is generally not sparsely correlated with a nice diagonal block structure. In fact, the correlation structure is often very complex and noisy. Though the treelets may provide a set of summary measures for the data set, the benefit of using these summary measures over those obtained using a traditional PCA for this type of data is not demonstrated. Also, we note that though they present the benefits of using their method as data reduction prior to prediction in Sections 5.1 and 5.3, in the case of the Glob

DNA microarray data in Section 5.3 the authors chose to reduce the data prior to the application of treelets using univariate regression. They restrict their data to the 1000 most "significant" genes. The reasons for this initial reduction are not stated, nor are the reasons for the arbitrary cut-off of 1000.

Often the truncation of a data set using a p-value cut-off is used to improve computational speed or improve accuracy. Regardless of the reasoning, the use of simple linear regression may not achieve an accurate ranking of "significant" genes. Univariate regression is notorious for detecting false positive genes. Constraining the data to the more "significant" genes may decrease the noise of the data, but it will not decrease the complexity of the correlation structure. We argue the use of targeted variable importance using targeted Maximum Likelihood or comparable double robust locally efficient estimation method would provide a more accurate ranking of the potentially causal genes [Bembom et al. (2007), Tuglus and van der Laan (2008)] than univariate regression. We also argue that if the initial reduction was completed to improve accuracy for the sake of prediction, the cut-off should be chosen with respect to the overall prediction performance. The Golub data, though commonly used to demonstrate prediction methods, is also commonly easy to obtain accurate results. The improvement accuracy of the treelet method over others is difficult to see when in general methods seem to perform so well.

3. Final comments. In general we believe treelets to be a great contribution to the field. With respect to clustering methodology, it provides a framework which actively searches for the correct underlying correlation structure. Its improvement over PCA when the correlation matrix is believed to be sparse is also impressive. Given the appropriate data and application, treelets will be a very useful and practical tool for statistical analysis.

REFERENCES

BEMBOM, O., PETERSEN, M. L., RHEE, S.-Y., FESSEL, W. J., SINISI, S. E., SHAFER, R. W. and VAN DER LAAN, M. J. (2007). Biomarker discovery using targeted maximum likelihood estimation: Application to the treatment of antiretroviral resistant hiv infection. U.C. Berkeley Division of Biostatistics Working Paper Series, Working Paper 221.

KAUFMAN, L. and ROUSSEEUW, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. Wiley, New York. MR1044997

POLLARD, K. and VAN DER LAAN, M. (2005). Cluster analysis of genomic data with applications in r. U.C. Berkeley Division of Biostatistics Working Paper Series, Working Paper 167.

TUGLUS, C. and VAN DER LAAN, M. (2008). Targeted methods for biomarker discovery: The search for a standard. Univ. California, Berkeley Division of Biostatistics Working Paper Series, Working Paper 233.

VAN DER LAAN, M. and BRYAN, J. (2001). Gene expression analysis with the parametric bootstrap. *Biostatistics* **2** 1–17.

VAN DER LAAN, M. and POLLARD, K. (2003). A new algorithm for hierarchical hybrid clustering with visualization and the bootstrap. *J. Statist. Plann. Inference* **117** 275–303. MR2004660

VAN DER LAAN, M., POLLARD, K. and BRYAN, J. (2003). A new partitioning around medoids algorithm. *J. Statist. Comput. Simul.* **73** 575–584. MR1998670

VAN DER LAAN, M., POLLEY, E. and HUBBARD, A. (2007). Super learner. U.C. Berkeley Division of Biostatistics Working Paper Series, Working Paper 222.

DIVISION OF BIOSTATISTICS
UNIVERSITY OF CALIFORNIA—BERKELEY
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: ctuglus@berkeley.edu

DIVISION OF BIOSTATISTICS
DEPARTMENT OF STATISTICS
UNIVERSITY OF CALIFORNIA—BERKELEY
BERKELEY, CALIFORNIA 94720
USA
E-MAIL: laan@berkeley.edu

The Annals of Applied Statistics
2008, Vol. 2, No. 2, 494–500
DOI: 10.1214/08-AOAS137REJ
Main article DOI: 10.1214/07-AOAS137

REJOINDER OF: TREELETS—AN ADAPTIVE MULTI-SCALE BASIS FOR SPARE UNORDERED DATA

BY ANN B. LEE, BOAZ NADLER AND LARRY WASSERMAN

Carnegie Mellon University, Weizmann Institute of Science and Carnegie Mellon University

We are grateful to all of the discussants for their thoughtful comments. Their remarks have added significant insight and perspective on the work. As a variety of issues have been raised, we have organized our rejoinder according to main topics that have been brought up by the discussants.

1. A multiresolution transform guided by the second-order statistics of the data. The treelet transform is a multiresolution transform that allows one to represent the original data in an alternative form. Rather than describe the data in terms of the original set of covariates, we perform a series of rotations which gradually reveal the hierarchical grouping structure of the covariates. The idea is very similar to the Grand Tour by Asimov (1985). The treelet transform is a tour "guided" by the covariance structure of the data.

Once the treelet transform has been completed, there are multiple ways of choosing an orthogonal basis (see Section 2.2). We never directly discard residual terms as noise. These terms are in fact an integral part of the final representation. In the simulated example of Section 4.2, most of the detail variables represent noise with small expansion coefficients; consequently, only certain coarse-grained variables are chosen for regression. In general, however, detail variables may convey crucial information. The latter point is illustrated in Sections 5.1 and 5.3, where we use the standard choice of one scaling term and $p-1$ difference terms; that is, an observation $\mathbf{x}$ is decomposed according to

$$\mathbf{x} = s\phi + \sum_{i=1}^{p-1} d_i \psi_i,$$

where the first term is a coarse-grained representation of the signal and the d-terms represent "differences" between node representations at two consecutive levels in the tree.

2. Orthogonal versus overcomplete bases. Tibshirani and Bickel/Ritov correctly point out that one need not restrict attention to one treelet level. An overcomplete dictionary of treelets can certainly be used for prediction. The "tree

Received March 2008; revised March 2008.

harvesting" scheme by Hastie et al. (2001), for example, takes node averages of all $2p - 1$ nodes in a hierarchical tree and uses these averages as new predictors for regression. The same scheme could be applied to treelets, but one would then also lose some of the strengths of treelets: Regression/classification is just one application of the treelet transform. More generally, the method yields a multiresolution analysis and a *coordinate system* of the data: we have a multi-scale basis function expansion of the data $\mathbb{X} = (\mathbf{x}_1, \dots, \mathbf{x}_n)$ and the covariance matrix $S = \frac{1}{n}\mathbb{X}^T\mathbb{X}$. An *orthonormal* basis also has many advantages compared to an overcomplete basis: (i) The representation is easy to interpret and computationally simple, (ii) the solution is *stable* in the sense that adding or omitting a covariate does not change the fit of the other covariates, (iii) the theoretical analysis is much simpler and (iv) the expansion coefficients sometimes carry information on the effective dimension of the data set and the relative importance of the coordinates; removing terms with small coefficients then has the effect of regularizing and denoising the data.

3. Treelets versus averaging predictors on preclustered trees. Meinshausen and Bühlmann ask how treelets are different from the following scheme: First order variables in a hierarchical cluster tree (under, e.g., complete linkage) and then find a basis on the tree by Principal Component Analysis (PCA). Tibshirani suggests a related scheme where one first builds a hierarchical cluster tree and then simply averages predictors in each cluster. Tuglus and van der Laan suggest other more sophisticated clustering techniques. We have not completed a full comparison of treelets and the schemes proposed by the discussants but would like to mention a few theoretical and practical advantages of treelets.

First, there are relatively few theoretical results on hierarchical clustering algorithms. Many popular procedures are not stable to noise, or even consistent. In Hartigan (1981), Hartigan writes that "standard hierarchical techniques, such as average and complete linkage, are hopelessly inconsistent [for density estimation]"; he then shows that single-linkage clustering is only weakly consistent or "fractionally consistent." Unfortunately, even less is known about the statistical properties of more complex methods that combine hierarchical clustering, averaging of predictors and regression. The treelet method has the advantage of being simple. The construction of an orthogonal basis and a tree in one step, using the covariance structure of the data, makes the algorithm computationally efficient and the method amenable to theoretical analysis. In our paper, we examine the large sample properties of treelets. We also show, for a block covariance model, that the required sample size for detecting the underlying population tree structure is logarithmic rather than linear in p. It is not clear if the same results apply to PCA on pre-clustered trees. It would be interesting to see more theoretical results on the many promising hierarchical clustering algorithms that have been suggested in the literature.

Compared to "simple averaging" of predictors in clusters, treelets also have other advantages: (i) The method yields an orthonormal basis; see item 2 above. (ii)

There is information *in the basis functions themselves*. Simple averaging does not provide such information and can also not adapt to the data. Treelets are constant on groups of indistinguishable variables (see Section 3.2.1, Corollary 1); this is not the case for simple averaging where the loadings are sensitive to the *exact order* in which one merges the variables. Moreover, if the groupings are less well defined and more "fuzzy," the loadings in treelets will also adapt accordingly. The latter point is illustrated by the waveforms in Figures 6, 7 and 10.

4. Identifiability and uniqueness. Sparse PCA. Bickel and Ritov (BR) raise two theoretical issues: identifiability and uniqueness of treelets. As BR point out, the treelet transform $T(\Sigma)$ viewed as a population parameter is a function of the population covariance matrix Σ only. The underlying structure in linear mixture models is indeed nonidentifiable, as there exist more than one solution for the loading vectors. The treelet transform chooses a representation that reflects groupings of highly correlated variables. These groups of variables, however, do not have to be disjoint for an approximate block covariance structure (as Example 3 in the paper also shows).

Why do we need treelets and what is the advantage of a treelet transform compared to other sparse methods? A notorious difficulty of least squares and variable selection methods lies in the collinearity between covariates; see Fan and Lv (2008), Section 4.1.2, on the need of a transform that takes advantage of the joint information among the predictors. Sparse PCA [Zou, Hastie and Tibshirani (2006)] with a combined l_1- and l_2-penalization scheme does find groupings of correlated variables but the results depend on the *particular* choice of tuning parameters. The latter choice defines the scale of the analysis. Real data sets, however, are often rather complex and groupings can occur on multiple scales. One of the strengths of the treelet method is that it captures hierarchical groupings by construction. The series of transformations in the method helps weaken correlation among the covariates. We do not think that treelet transform is a replacement of other sparse methods. On the other hand, it can be a useful tool if combined with other sparse methods as suggested by Fan and Lv.

Bickel and Ritov also raise the issue of uniqueness. We would like to point out that if we use covariances as a similarity measure, the treelet transform is unique up to a permutation of second-order statistically exchangeable variables. In most applications, correlations seem to be a better measure of similarity. The treelet transform T with a correlation measure is, however, multivalued: formally, $T(P)$ is a set of transforms rather than a single transform. If treelets are viewed as an exploratory tool, then we do not find this fact troubling. An analogy with mixture models might be helpful. Mixture models are famous for suffering numerous irregularities: local nonidentifiability, intractable limiting distributions of test statistics, nonunique maxima of the likelihood function, infinite likelihood values and slow convergence rates, to name a few. For theoretical analysis, they can be a nightmare. Nonetheless, they are used in many applications with great practical success. Like

BR, we find the nonidenitifability and multivalued properties of treelets disquieting but, like mixture models, they nonetheless do seem to be useful. Ultimately, the effectiveness of treelets in real problems will determine their utility. On the other hand, any theoretical ideas that provide insight are welcome. Thus, we are intrigued by BR's conjecture at the end of their Section 1. We look forward to hearing about future progress on this idea.

5. Supervised learning. We agree with Meinshausen and Bühlmann (MB) that constructing predictors without using the response Y does fail in some cases. The advantage of treelets is the intepretability of the derived features. Sometimes constructing predictors without reference to Y is a necessity. An example is the problem of semi-supervised inference. In this case we observe labeled data $(X_1, Y_1), \dots, (X_n, Y_n)$ but we also have access to unlabeled data $X_{n+1}, \dots, X_N$, where N is much larger than n. Evidence that the unlabeled data alone can be used to construct effective predictors abounds in the machine learning literature. As Tibshirani writes in his discussion, there is also growing empirical evidence that unsupervised feature extraction can provide an effective set of features for supervised learning. Tibshirani cites the recent work by Hinton, Osindero and Teh (2006) on learning algorithms for Boltzmann machines as an example.

MB point out that information in the response variable can be used in various ways "ranging from weak use of the response to stronger involvement." They give some innovative suggestions on how the response could potentially be incorporated into a treelet framework. As MB writes, the current supervised choice of basis functions by cross-validation represents one use of the response, but perhaps a weaker one. In their discussion, they mention "fully supervised" schemes where Y is used to construct the groupings themselves. We plan to look into various such extensions of treelets in the future.

Regarding supervised learning of predictors, we are intrigued by Bickel and Ritov's suggested method for iteratively growing a class of basis functions. Independently, we have been experimenting with a similar algorithm in the context of modeling phenotypes on interactions of SNPs. Like BR, we start with main effects and gradually add interaction terms in an adaptive fashion. We have recently begun a theoretical analysis of this idea and we look forward to comparing our results with those of BR.

6. Scalability and other computational issues. Murtagh raises questions about the scalability of the treelet algorithm. Our current implementation of the treelets uses an exhaustive search at each level of the tree. This is typical of bottom-up hierarchical algorithms and corresponds to a computational cost of $O(Lp^2) + m$, where L is the level of the tree, p is the number of variables (or leaves in the tree) and m is the initial cost of computing the data covariance matrix. However, by keeping track of local changes in the covariance matrix (see

Section 2.1), the complexity of the treelet algorithm can further be reduced to $O(Lp)+m$.

We do not believe our method has any computational disadvantage compared to Murtagh's method with fixed Haar wavelets on precomputed dendrograms [Murtagh (2007)]. The cost in computing an adaptive basis is neglible compared to the cost of computing the dendrogram itself. The experimental evaluations in the paper are on $p = 1000$ variables because of the nature of the problems and, in the case of the analysis of the Golub microarray data, because of the availability of benchmark results for this choice of p. One can run the computations efficiently in higher dimensions, such as $p \gtrsim 10000$. While we disagree with Murtagh regarding scalability, we agree that treelets may not be appropriate for "ultra-high" dimensional settings (e.g., $p \gtrsim 100000$), where certain topological phenomena may dominate the data.

We plan to post open-source code in both $C++$ and R at http://www.stat.cmu.edu/~annlee/software.htm by the end of the summer of 2008. Until then, we have made available some Matlab test code at the same URL. This code, however, has not been optimized for speed or efficiency in memory use.

7. Applicability to microarray data. Finally, Qiu and Tuglus/van der Laan (TV) comment on the applicability of treelets to microarray data. We are not experts on the analysis of such data, but would like to bring up a few potentially important points.

TV correctly state that treelets are built upon a hierarchical scheme of grouping variables and that the graph structure is solely based on correlations. They suggest that other similarity or distance measures may be more appropriate for clustering. We agree on this point but would like to emphasize that the goal of treelets is not clustering per-se. It is the construction of a multi-resolution representation of data. Should other distance measures be used, one needs to define how to aggregate the resulting sets of variables. In principle, one can also think of graph-theoretic measures of similarity between variables, and nonlinear treelet-inspired local transformations between them (for example, for data lying on nonlinear manifolds). The theoretical analysis becomes increasingly difficult once one goes beyond second-order statistics.

Qiu remarks that a possible pitfall of the treelet methods is its preference for sum variables with higher variance than the corresponding detail variables. He argues that genes with smaller variability may be the ones responsible for essential biological functionalities. In our framework, detail variables are not discarded. They are only removed from further merging in the tree. These detail variables can certainly be included in a regression or classification model, as is also shown in the paper. Furthermore, correlation-based treelets can actually be useful in unraveling groups of genes with low variance. Consider, for example, data with sets of genes with very different variances and different intrinsic noise levels. Global variance-based methods such as PCA or sparse PCA would not pick up groups of genes with individual low variances. However, if these variables are highly correlated, they will be among the first ones to be identified and merged with the treelet algorithm.

In our paper (Section 5.3) we describe a "two-way" classification scheme for the Golub leukemia data set. Qiu asks for a clarification of this method. Our main goal here was to show that treelets can be built on both variables (genes) and samples (patients). We are not claiming that the method is superior—only that a general method such as treelets can be competitive with state-of-the-art algorithms that are especially tuned for the analysis of microarray data. The proposed scheme is as follows: First compute treelets on the genes using the training data. This part is the same as for "LDA on treelet features." The second step, however, is different. Here we express *all* 72 samples (patients) in terms of their new profiles over the K maximum variance treelets. We build treelets on the new patient profiles and find the two main branches of the tree. The two groups represent the two cancer classes (ALL or AML); these groups are labeled using the training data and a majority vote. The error is evaluated on the test set (see Figure 9, right). Note that the second step, the labeling of samples, is an example of *semi-supervised learning* (see item 5). It is not a violation of cross-validation. On the contrary, semi-supervised learning (SSL) is a powerful method that is often used to improve classification; see, for example, Belkin and Niyogi (2005). The key idea behind SSL is that *unlabeled* data can be used to uncover the underlying structure of the data (e.g., low-dimensional manifolds, groupings etc.) and that this knowledge can lead to better prediction than if only labeled data had been used.

To summarize, we do not claim that the treelets are the optimal method to model microarray data. They might miss important effects in certain settings. However, treelets or some of their possible generalizations may turn out to be useful in the analysis of such data. Further research is required in this direction.

Acknowledgments. It has been a real pleasure to participate in this discussion. We would like to thank the AOAS Biology Sciences Editor Michael Newton, the AOAS Editor-In-Chief Bradley Efron and the discussants for making this exchange of ideas possible.

REFERENCES

ASIMOV, D. (1985). The Grand Tour: A tool for viewing multidimensional data. *SIAM J. Sci. Comput.* **6** 128–143.

BELKIN, M. and NIYOGI, P. (2005). Semi-supervised learning on Riemannian manifolds. *Machine Learning* **56** 209–239.

FAN, J. and LV, J. (2008). Sure independence screening for ultra-high dimensional feature space. *J. Roy. Statist. Soc. B*. To appear.

HARTIGAN, J. A. (1981). Consistency of single linkage for high-density clusters. *J. Amer. Statist. Assoc.* **76** 388–394.

HASTIE, T., TIBSHIRANI, R., BOTSTEIN, D. and BROWN, P. (2001). Supervised harvesting of expression trees. *Genome Biology* **2** research0003.1–0003.12.

HINTON, G. E., OSINDERO, S. and TEH, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Comput.* **18** 1527–1554.

MURTAGH, F. (2007). The Haar wavelet transform of a dendrogram. *J. Classification* **24** 3–32.

ZOU, H., HASTIE, T. and TIBSHIRANI, R. (2006). Sparse principal component analysis. *J. Comput. Graph. Statist.* **15** 265–286.

A. B. LEE
L. WASSERMAN
DEPARTMENT OF STATISTICS
CARNEGIE MELLON UNIVERSITY
PITTSBURGH, PENNSYLVANIA 15206
USA
E-MAIL: annlee@stat.cmu.edu
larry@stat.cmu.edu

B. NADLER
DEPARTMENT OF COMPUTER SCIENCE
AND APPLIED MATHEMATICS
WEIZMANN INSTITUTE OF SCIENCE
REHOVOT
ISRAEL
E-MAIL: boaz.nadler@weizmann.ac.il