Refining Genetically Inferred Relationships Using Treelet Covariance Smoothing

Andrew Crossett[3], Ann B. Lee[1*], Lambertus Klei[2], Bernie Devlin[2], and Kathryn Roeder[1]

[1] Department of Statistics
Carnegie Mellon University
Pittsburgh, PA 15213

[2] Department of Psychiatry
University of Pittsburgh School of Medicine
Pittsburgh, PA 15213

[3] Department of Mathematics
West Chester University
West Chester, PA 19383

*Corresponding Author: 5000 Forbes Ave, 229J Baker Hall, Department of Statistics, Carnegie Mellon University, Pittsburgh, PA 15213, phone: 412-268-7831, annlee@cmu.edu

Running title: Treelet Covariance Smoothing

Abstract

Recent technological advances coupled with large sample sets have uncovered many factors underlying the genetic basis of traits and the predisposition to complex disease, but much is left to discover. A common thread to most genetic investigations is familial relationships. Close relatives can be identified from family records, and more distant relatives can be inferred from large panels of genetic markers. Unfortunately these empirical estimates can be noisy, especially regarding distant relatives. We propose a new method for denoising genetically–inferred relationship matrices by exploiting the underlying structure due to hierarchical groupings of correlated individuals. The approach, which we call Treelet Covariance Smoothing, employs a multiscale decomposition of covariance matrices to improve estimates of pairwise relationships. On both simulated and real data, we show that smoothing leads to better estimates of the relatedness amongst distantly related individuals. We illustrate our method with a large genome-wide association study and estimate the "heritability" of body mass index quite accurately. Traditionally heritability, defined as the fraction of the total trait variance attributable to additive genetic effects, is estimated from samples of closely related individuals using random effects models. We show that by using smoothed relationship matrices we can estimate heritability using population-based samples. Finally, while our methods have been developed for refining genetic relationship matrices and improving estimates of heritability, they have much broader potential application in statistics. Most notably, for error-in-variables random effects models and settings that require regularization of matrices with block or hierarchical structure.

Key Words: covariance estimation, cryptic relatedness, genome-wide association, heritability, kinship.

Introduction. In the past decade tremendous progress has been made toward understanding the genetic basis of disease. This challenging endeavor has given rise to numerous study designs with a vast arsenal of statistical machinery. A common theme however is the pivotal role played by familial relationships. Traditionally relationships are encoded in pedigrees of known relatives (Thompson, 1974, 1975; Boehnke and Cox, 1997; Epstein, Duren and Boehnke, 2000; McPeek and Sun, 2000), but for more distantly related individuals, pedigree information can sometimes be erroneous or difficult to obtain. Relatedness can also be calculated from large panels of genetic markers (Milligan, 2003; Albers et al., 2008; Anderson and Weir, 2007; Browning, 2008; Browning and Browning, 2010; Purcell et al., 2007; Day-Williams et al., 2011; Yang et al., 2010a). While this approach has greatly expanded the scope of inference for relationships, empirical estimates are noisy, especially regarding distant relatives.

The search for a disease gene begins with finding unusual sharing of genetic material among individuals who share a trait (phenotype). Linkage analysis involves the study of joint inheritance of genetic material and phenotypes within relatives (Hopper and Mathews, 1982; Almasy and Blangero, 1998). Typically, these studies are restricted to relatives within a pedigree, but more recently the approach has been extended to samples of people who are more distantly related and without known pedigree structure (Day-Williams et al., 2011). Alternatively, genetic associations can be discovered from population samples, which are usually based on case-control studies. In these studies the sample is assumed to be unrelated but the presence of distant relatives (i.e. cryptic relatedness) can reduce power or generate spurious associations (Lander and Schork, 1994; Astle and Balding, 2009). Numerous methods have been proposed to deal with familial structure in genetic association studies (Choi, Wijsman and Weir, 2009; Bravo et al., 2009; Thornton and McPeek, 2010; Kang et al., 2010), all of which require an estimate of family relationships among individuals within the study.

Relationships are also critical for quantitative genetics. A common problem for quantitative genetics is to estimate the fraction of variance of a continuous trait, such as height, due to genetic variation amongst individuals in a population. This feature, known as heritability, delineates the relative contributions of genetic and non-genetic factors to the total phenotypic variance in a population. Heritability is a fundamental concept in genetic epidemiology and disease mapping. Using a variety of close relatives, the heritability of quantitative and qualitative traits can be estimated directly (Fisher, 1918; Devlin, Daniels and Roeder, 1997). With complex pedigrees, applying the same principles, heritability can be estimated using random effects models

(Henderson, 1950). Heritability of height, weight, IQ and many other quantitative traits has been investigated for nearly a century and continues to generate interest (Deary et al., 2012).

Interest in the genetic basis of disease is high because greater understanding of disease etiology will in principle lead to better treatments. Large population-based samples are enhancing our ability to identify DNA variants affecting risk for disease and it has become the standard to search for genetic variants associated with common disease using genome-wide association studies (GWAS). Thousands of associations for common diseases/phenotypes have already been validated (Visscher et al., 2012). Nevertheless, even in the most successful cases, such as Inflammatory Bowel Disease studied in McGovern et al. (2010) and Imielinski et al. (2009), discoveries account for only a fraction of the heritability.

Given the relatively limited discoveries thus far, a reasonable question is whether the heritability of a trait estimated from relatives truly does trace to genetic variation. Yang et al. (2010a) offer a novel approach to genetic analysis that shows that indeed much of it does. They propose to analyze population samples, rather than pedigrees, for the heritability of the trait. To do so they first estimate the correlation between all pairs of individuals in the population sample using a dense set of common genetic variants, such as those typically used for a GWAS. They then take this matrix and relate it to the covariance matrix of phenotypes for these subjects to derive an estimate of heritability. Thus in their application, where essentially all relatives are removed from the sample, heritability refers to the proportion of variance in the trait explained by the measured genetic markers. They provide a fascinating example of how this approach works in the case of human height and they and others applied these techniques to many other traits (reviewed by Visscher et al. (2012)).

Yang et al. (2010a)'s work inspired us to consider applying a related approach to answer a different question. Could estimates of relatedness obtained from a population sample be improved by using smoothing techniques on the variance-covariance matrix? If so, population samples could be used to estimate heritability – in the traditional sense – without requiring close relatives. This approach has application to phenotypes for which extended pedigrees are difficult to obtain. For instance, there is controversy in the literature concerning the heritability of autism, which is typically estimated from twin studies (Hallmayer et al., 2011). Smoothing techniques could also be used to estimate relatedness in samples of distantly related individuals for many other genetic analyses. For example, a version of linkage analysis could be applied to distant relatives.

We propose Treelet Covariance Smoothing – a novel method for smoothing and multiscale decomposition of covariance matrices – as a means to improving estimates of relationships. Treelets were first introduced in Lee and Nadler (2007) and Lee, Nadler and Wasserman (2008) as a multi-scale basis that extends wavelets to unordered data. The method is fully adaptive. It returns orthonormal basis functions supported on nested clusters in a hierarchical tree. Unlike other hierarchical methods, the basis and the tree structure are computed simultaneously, and both reflect the internal structure of the data.

In this work, we extend the original treelet framework for smoothing of one-dimensional signals to smoothing and denoising of variance-covariance matrices with hierarchical block structure and unstructured noise. Smoothing is achieved by a nonlinear approximation scheme in which one discards small elements in a multi-scale matrix decomposition. The basic idea is that if the data have underlying structure in the form of groupings of correlated variables, then we can enforce sparsity by first transforming the data into a treelet representation by a series of rotations of pairs of correlated variables, and then thresholding covariances. We refer to this new regularization approach for covariance matrices with groupings on multiple scales as *Treelet Covariance Smoothing* (TCS).

We apply TCS to genetically inferred relationship matrices, with the goal of improving estimates of pairwise relationships from large pedigrees and population-based samples. On both simulated and real data, we show that TCS leads to better estimates of the relatedness between individuals. Using these estimates allows us to estimate the heritability from population-based samples provided they include some distantly related individuals, a property that is almost inevitable in practice. Finally, we discuss how estimating heritability is simply a case of variance component estimation for an error-in-variables random effects model. Therefore, our method can be applied to a whole family of more general models of similar structure.

Models.

GWAS Panels. The human genome contains many millions of single nucleotide polymorphisms (SNPs) and other genetic variation distributed across the genome. In a GWAS it is now typical to measure a panel of at least 500,000 SNPs from each subject. SNPs typically have only two forms or alleles within a population. Whichever allele is less frequent is called the minor allele. The genotype of an individual at a SNP can then be coded as 0, 1 or 2 depending on the number of minor alleles the individual has at that SNP. Alleles at SNPs in close physical proximity are often highly correlated (i.e., in linkage disequilibrium). When multiple SNPs are in linkage disequilibrium, we say one of these SNPs "tags", or represents, the others. Although estimates vary, well-designed panels of 500,000 SNPs do not tag all of the common SNPs in the genome and they tag very few of the SNPs with rare minor alleles (Yang et al., 2010a). Nevertheless, GWAS provide considerable information about familial relationships.

Estimating Genetic Relationships. The relatedness between a pair of individuals is defined by the frequency by which they share alleles *identical by descent* (IBD). Formally, two alleles are considered IBD if they descended from a common ancestor without an intermediate mutation. Within a pedigree relatives share very recent common ancestors, hence many alleles are IBD. For a more detailed exposition of genetic relationships, see Astle and Balding (2009).

The quantity of interest in this investigation is the *Additive Genetic Relationship* which is defined as the expected proportion of alleles IBD for a pair of individuals. For individuals i and j we use A_{ij} to denote this quantity, which is more familiar when viewed as the *degree of relationship*, where $R_{ij} = -\log_2(A_{ij})$. For example, for siblings, first cousins, and second cousins, who are 1'st, 3'rd and 5'th degree relatives, A is 1/2, 1/8 and 1/32, respectively. Within a non-inbred pedigree A can be computed using a recursive algorithm (Thompson, 1986). For example, if individual i has parents k and l, then $A_{ij} = A_{ji} = 1/2(A_{jk} + A_{jl})$.

For distantly related individuals, detailed pedigree information is not often available; however, with GWAS data one can calculate genome-average relatedness directly (Astle and Balding, 2009). Even with complete information regarding IBD status of the chromosomes, the fraction of genetic material shared by relatives will differ slightly from the expectation calculated from the pedigree due to the stochastic nature of the meiotic process (Weir, Anderson and Hepler, 2006). For the purpose of genetic investigations, one could argue that genome-average relatedness is a truer measure of relatedness. For example, while two distantly related individuals are expected to share a small fraction of their genetic material, if they do not inherit anything from their common ancestor it seems appropriate to consider them unrelated.

Under many population genetic models A_{ij} can also be interpreted as a correlation coefficient. Let z_{ik} denote the scaled minor allele count for individual i at SNP k: $z_{ik} = (z^*_{ik} - 2p_k)/(2p_k(1-p_k))^{1/2}$, where z^*_{ik} is the minor allele count and p_k is the minor allele frequency. For individuals i and

j at genetic variant k, it follows from our model that

$$\mathrm{Cov}[z_{ik}, z_{jk}] = A_{ij}. \tag{1}$$

Exploiting this feature leads to a method of moments estimate of A from a panel of m genetic markers. To see this, let $\mathbf{z}_k$ denote a column vector of observed scaled allele counts for all individuals at the k'th SNP, then let

$$\widehat{A} = \frac{1}{m} \sum_{k=1}^{m} \mathbf{z}_k \mathbf{z}_k^t = \frac{ZZ^t}{m}, \tag{2}$$

where $Z = (\mathbf{z}_1, \ldots, \mathbf{z}_m)$. The Genome-wide Complex Trait Analysis (GCTA) software from Yang et al. (2010b) computes this estimator.

The method of moments estimator is unbiased if the population allele frequencies are known (Milligan, 2003). In practice, the p_k's are estimated from the sample data. A criticism of this estimator is that some off-diagonal elements are negative, which doesn't conform to the interpretation of A_{ij} as a probability. Viewed as a correlation coefficient, however, negative quantities suggest the pair of individuals share fewer alleles than expected given the allele frequencies. Alternatively, maximum likelihood estimators of A have been developed (Thompson, 1975; Milligan, 2003), but these estimators are quite computationally intensive for GWAS panels. Hence, while method of moments estimators are typically less precise than maximum likelihood estimators, they are more commonly used when a large SNP panel is available.

Estimating Heritability. By definition, the heritability of a quantitative trait (y) such as height is determined by the additive effect of many genes and genetic variants (g), each of small effect (i.e., the polygenic model). For individuals $i = 1, \ldots, n$, suppose that the genetic effects are explained by J causal SNPs, and we can express the genetic effect as

$$g_i = \sum_{j=1}^{J} z_{ij} u_j, \tag{3}$$

where u_j is the additive random effect of the jth causal variant, weighted by the scaled number z_{ij} of minor alleles at this variant. Let $\mathbf{g} = (g_1, \ldots, g_n)^t$ be the vector of random effects corresponding to the additive genetic effects for individuals $i = 1, \ldots, n$. For $\mathbf{u} = (u_1, \ldots, u_J)^t$ and $Z_c = [z_{ij}]$, we write $\mathbf{g} = Z_c\mathbf{u}$. Define G as the variance-covariance matrix of $\mathbf{g}$. Assuming $\mathrm{Var}[\mathbf{u}] = I\sigma_u^2$, it follows that

$$G = \sigma_g^2 \frac{Z_c Z_c^t}{J} \tag{4}$$

where $\sigma_g^2 = J\sigma_u^2$.

In the traditional model for quantitative traits a continuous phenotype y is modeled as

$$y_i = \mu + g_i + e_i, \tag{5}$$

where $\mathbf{e} = (e_1, \ldots, e_n)^t$ is the vector of residual effects, and $\mathbf{y} = (y_1, \ldots, y_n)^t$ is the vector of phenotypes. In matrix notation, $\mathbf{y} = \mathbf{1}\mu + \mathbf{g} + \mathbf{e}$. The residuals are assumed to be independent with variance-covariance equal to $I\,\sigma_e^2$ and the random effects and residual error are assumed to be normally distributed. Consequently,

$$\text{Var}[\mathbf{y}] = \frac{Z_c Z_c^t}{J}\sigma_g^2 + I\sigma_e^2. \tag{6}$$

The heritability of the phenotype y is defined as

$$h^2 = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}.$$

This quantity is more accurately known as the additive or narrow-sense heritability, in contrast to the broad-sense heritability, which includes non-additive genetic effects such as gene-gene interactions. Our inferences will be confined to narrow-sense heritability.

If the causal SNPs (or good tag SNPs) and the phenotype were directly measured, then one could estimate h^2 based on Eq. 5 and the implied random effects model using maximum likelihood (REML) (Searle et al., 1992). Notationally, Z_c is an $n \times J$ matrix that picks out J columns of the full SNP panel Z. In practice, Z_c is not known. Few of the causal SNPs are known for any phenotype, and many causal SNPs will be missing from Z (i.e. not tagged by any measured SNPs).

How then is h^2 estimated in practice? Assuming various subsets of individuals in the sample are related with relationship matrix A (defined previously), heritability can be estimated without any knowledge of causal genetic variants that constitute $\mathbf{g}$. From Eq. 1 and the polygenic model it follows that $\frac{Z_c Z_c^t}{J} \rightarrow A$ as J gets large. This inspires an alternative random effects model which has long been utilized in population genetics:

$$\text{Var}[\mathbf{y}] = A\sigma_g^2 + I\sigma_e^2. \tag{7}$$

Historically, A has been derived from known pedigree structure. However, provided some subsets of the individuals in the sample are related (even distantly), one can estimate A from genetic markers using either method of moments or maximum likelihood estimation techniques. This approach has been applied frequently in quantitative genetics, especially in breeding studies (Lynch and Ritland, 1999; Eding et al., 2001; Visscher et al., 2006; Hayes and Goddard, 2008). We conjecture that by using TCS, we can improve estimates of A and obtain better estimates of heritability without knowledge of causal variants.

Alternatively, if the sample is completely unrelated then substituting the result of Eq. 2 for 6 does not lead to an estimate of h^2 unless all of the causal SNPs have been recorded. Instead this approach estimates $h_s^2 \leq h^2$, the proportion of the variance in phenotype explained by the SNP panel (Yang et al., 2010a). In this setting, TCS will not improve estimates of h_s^2.

Methods.

Treelet Covariance Smoothing (TCS). The genetic relationship matrix A is a measure of the additive covariance structure that exists between individuals due to a common genetic background. We estimate the relationship matrix using genotyped SNPs, but this estimate is usually noisy. Hence, we propose a method for improving upon this estimate using treelets.

Treelets simultaneously return a hierarchical tree and orthonormal basis functions supported on *nested clusters* in the tree – both reflect the underlying structure of the data. Here we extend the original treelet framework (Lee and Nadler, 2007; Lee, Nadler and Wasserman, 2008) for smoothing one-dimensional signals and functions, to a new means of smoothing and denoising variance-covariance matrices with hierarchical block structure and unstructured noise. The main idea is to first move to a different basis representation through a series of local transformations, and then impose sparsity by thresholding the transformed covariance matrix. We refer to the approach as Treelet Covariance Smoothing (TCS). The general set-up is as follows (see Appendix for reference to code and details on how to compute the treelet transformations):

Let $\mathbf{z}$ be a random vector in $\mathbb{R}^N$ with variance-covariance matrix Σ. In our context, $\mathbf{z}$ represents the scaled minor allele counts for a set of N individuals at any SNP, and the covariance $\Sigma = A$, the additive genetic relationship matrix of the N individuals (Eq. 1). Now at each level of the treelet algorithm, we have an orthonormal multiscale basis. Let $\{\mathbf{v}_1, \ldots, \mathbf{v}_N\}$ denote the basis at the top of the tree (corresponding to level $\ell = N-1$ if using the notation in Appendix). We write

$$\mathbf{z} = \sum_{i=1}^{N} c_i \mathbf{v}_i, \tag{8}$$

where $c_i = \langle \mathbf{z}, \mathbf{v}_i \rangle$ represent the orthogonal projections onto local basis vectors on different scales. It follows that the covariance of $\mathbf{z}$ can be written in terms of a *multi-scale matrix decomposition*

$$\Sigma = \mathrm{Var}(\mathbf{z}) = \sum_{i=1}^{N} \gamma_{i,i}\mathbf{v}_i(\mathbf{v}_i)^t + \sum_{i\neq j}^{N} \gamma_{i,j}\mathbf{v}_i(\mathbf{v}_j)^t, \tag{9}$$

where $\gamma_{i,i} = \mathrm{Var}(c_i)$ and $\gamma_{i,j} = \mathrm{Cov}(c_i, c_j)$. The first term in Eq. 9 describes the diagonally symmetric block structure of the variance-covariance matrix. These blocks are organized in a hierarchical tree. The second term describes more complex structure, including off-diagonal rectangular blocks, which are also hierarchically related to each other in a multi-scale matrix decomposition.

In practice, the covariance Σ is unknown, and both the covariance matrix and the treelet basis need to be estimated from data. For relationship matrices, one can for example derive an estimate $\widehat{\Sigma} = \widehat{A}$ from marker data using method of moments or maximum likelihood methods. Denote the treelet basis derived from $\widehat{\Sigma}$ by $\{\widehat{\mathbf{v}}_1, \ldots, \widehat{\mathbf{v}}_N\}$, and write

$$\widehat{\Sigma} = \sum_{i=1}^{N} \widehat{\gamma}_{i,i}\widehat{\mathbf{v}}_i(\widehat{\mathbf{v}}_i)^t + \sum_{i\neq j}^{N} \widehat{\gamma}_{i,j}\widehat{\mathbf{v}}_i(\widehat{\mathbf{v}}_j)^t,$$

where $\widehat{\gamma}_{i,i} = \widehat{\mathrm{Var}}(c_i)$ and $\widehat{\gamma}_{i,j} = \widehat{\mathrm{Cov}}(c_i, c_j)$.

Let $T(\widehat{\Sigma})$ be the covariance estimate after a treelet transformation, i.e. after applying a full set of $N-1$ Jacobi rotations of pairs of correlated variables. A calculation shows that

$$\widehat{\gamma}_{i,i} = \widehat{\mathrm{Var}}(c_i) = [T(\widehat{\Sigma})]_{ii} \quad \text{and} \quad \widehat{\gamma}_{\mathrm{i,j}} = \widehat{\mathrm{Cov}}(\mathrm{c_i}, \mathrm{c_j}) = [\mathrm{T}(\widehat{\Sigma})]_{\mathrm{ij}}, \tag{10}$$

where $c_i \equiv \langle \mathbf{z}, \mathbf{v}_i \rangle$ and $c_j \equiv \langle \mathbf{z}, \mathbf{v}_j \rangle$. This suggests[1] a smoothed estimate of the covariance by thresholding:

$$\widetilde{\Sigma}(\lambda) = \sum_{i=1}^{N} f_\lambda[\widehat{\gamma}_{i,i}]\widehat{\mathbf{v}}_i(\widehat{\mathbf{v}}_i)^t + \sum_{i\neq j}^{N} f_\lambda[\widehat{\gamma}_{i,j}]\widehat{\mathbf{v}}_i(\widehat{\mathbf{v}}_j)^t, \tag{11}$$

with the thresholding function

$$f_\lambda[a] = \begin{cases} a & \text{when } |a| \geq \lambda \\ 0 & \text{when } |a| < \lambda, \end{cases} \tag{12}$$

[1] The special case $c_i \equiv \langle \mathbf{z}, \delta_i \rangle$ and $c_j \equiv \langle \mathbf{z}, \delta_j \rangle$, where δ_i denotes the Kronecker delta function, corresponds to simple thresholding of the original covariance estimate. Here we consider more general groupings of correlated variables on different scales.

where λ is a smoothing parameter.

To summarize and in matrix short-hand notation: The smoothed genetic relationship matrix is given by

$$\widetilde{A}(\lambda) = B\ f_\lambda[T(\widehat{A})]\ B^t, \tag{13}$$

where $B = (\widehat{\mathbf{v}}_1, \ldots, \widehat{\mathbf{v}}_N)$ and $T(\widehat{A})$, respectively, denote the treelet basis and the covariance matrix at the top of the tree, and f_λ corresponds to element-wise thresholding (Eq. 12). Note that to compute B we only need to know the Jacobi rotations at each level of the tree. This is because of Eq. 17 (in Appendix). More precisely, the treelet basis, $B = J^{(1)} \cdot J^{(2)} \cdot \ldots \cdot J^{(N-1)}$, where the Jacobi rotation matrix $J^{(\ell)}$ is the rotation matrix at level ℓ. The covariance estimate after a treelet transformation and before smoothing, $\widehat{\Sigma}^\ell \equiv T(\widehat{A}) = B^t \widehat{A} B$.

Choosing a smoothing parameter. The goal is to choose a threshold (λ) that reduces noise in the estimated relationships. Traditional cross-validation is not an option because we cannot predict A_{ij} without including persons i and j. Alternatively, we have an abundance of genetic information from which to estimate $\widehat{A}$. We propose a SNP subsampling procedure to estimate the tuning parameter.

We begin by breaking the genome into independent *training* and *test* sets by randomly placing half the chromosomes into each set. To improve the efficiency of our estimate of A, we utilize a "blackout window" of length b to avoid sampling SNPs that are highly correlated. This b can be considered either in terms of physical location along the chromosome or the number of SNPs between any two SNPs in question. From the set of training chromosomes, select a relatively large sample of M independent SNPs to get a reliable estimate of $\widehat{A}$. We train our algorithm by smoothing $\widehat{A}$ using TCS to get $\widetilde{A}(\lambda)$, for all $\lambda \in \Lambda$, where Λ is a grid of reasonable threshold values.

Once we have $\widetilde{A}(\lambda)$, for a given λ, we subsample L SNP sets of size k from the test set of chromosomes. Here, $k \ll M$ and the SNPs within each of the L subsampled sets follow our defined blackout window, b. Then, for all $l = 1, \ldots, L$, estimate the relationship matrix, $\widehat{A}_l$, based on the subset of SNPs. We then compare our smoothed relationship matrix, $\widetilde{A}(\lambda)$, from the training chromosomes to each of the L non-smoothed relationship matrices, $\widehat{A}_l$, via a weighted risk function:

$$H(\lambda) = \frac{1}{(N-1)NL} \sum_{l=1}^{L} \sum_{i<j}^{N} w_{ij} \left(\widehat{A}_{ij,l} - \widetilde{A}_{ij}(\lambda) \right)^2, \tag{14}$$

where w_{ij} is a weight associated with each element in A. Clearly, the optimal tuning parameter is $\widehat{\lambda} = \text{argmin}_{\lambda \in \Lambda} H(\lambda)$.

The reason for introducing the weighting scheme is because many subjects are nearly unrelated. Thus, we aim to upweight the loss function so that the preponderance of near-zero elements in the off-diagonal do not overwhelm the loss function. We suggest using the learned hierarchical tree to get the weights. More specifically, $w_{ij} = |[T(\widehat{A})]_{ij}|$, corresponding to the absolute value of the correlations between the final groupings of individuals after $N-1$ rotations (Eq. 10). Also, we set $w_{ii} = 0$ because we are not interested in estimating inbreeding coefficients. It should be noted that this is a rather general weighting method. Other schemes may be more appropriate if there is *a priori* information suggesting the importance of particular relationships.

Results.

Simulations. To produce realistic simulations we started with the phased genomes (haplotypes) of individuals from the HapMap 3 database[2], selecting two populations with European ancestry (CEU and TSI). Utilizing the small sample of available haplotypes, our first objective was to generate a large sample of haplotypes, representative of those that might be sampled from unrelated founders of a population. The challenge was to keep intact the realistic haplotype structure for a human population, including linkage disequilibrium (LD), without generating unusual sharing between the founders. To accomplish this goal we took the HapMap data on CEU and Tuscan samples, which were phased quite accurately into haplotypes, as the initial sample of chromosomes from which to generate founders. Now each founder haplotype was created by sampling pieces of chromosomes (or haplotypes) from the initial sample. To do so, the number of recombination spots per chromosome was determined using an overall recombination of $\theta = 10^{-6}$ per Mb, which is 100 times the normal rate of recombination for humans. The actual location of the recombination spots were then determined using the recombination map provided by HapMap, a procedure that successfully keeps intact the LD structure of the chromosome. From this pool of generated haplotype pairs, chromosomes were randomly assigned to each of the 39 founders in each of 100 families. These founder chromosomes were then dropped through a seven generation pedigree. At each generation the chromosomes underwent recombination with an overall rate of $\theta = 10^{-8}$ at locations determined by HapMap's recombination map. Within each pedigree, the genotype information of twenty individuals was collected (colored

[2] http://hapmap.ncbi.nlm.nih.gov/downloads/phasing/2009-02_phaseIII/HapMap3_r2/

yellow). We then sampled ten individuals of varying relatedness from this group with a random sampling strategy that favored individuals of distant relatedness within the pedigree. The highest pairwise relatedness within a family is .125, corresponding to $R = 3$, and the lowest is $< .001$. Individuals from different families are unrelated. Each simulation produced a total of 1,000 individuals made up of 100 ten-member families of varying levels of relatedness. Finally, the entire process was repeated fifty times.

Because we know the pedigree structure, we can compare the unsmoothed estimate $\widehat{A}$ to $\widetilde{A}$ found via TCS. Here, we use the GCTA software (Yang et al., 2010b) to estimate $\widehat{A}$ using 100,000 randomly chosen SNPs with MAF $> .05$. The optimal level of smoothing $(\widehat{\lambda})$ is chosen via the subsampling scheme described previously using $M = 5,000$, $b = 10$, $k = 50$, $L = 50$ and repeating everything ten times. Here, b is in terms of number of SNPs. We choose $\widehat{\lambda}$ by examining a plot of $H(\lambda)$ across a grid of λ values. The optimal smoothing parameter is the one that minimizes the risk function, H. For one such simulation sample we can see from Figure 2 that $\widehat{\lambda} \approx .051$.

The question then becomes, does TCS improve estimates of relatedness? Figures 3 and 4 display boxplots comparing the root mean square error (RMSE) of $\widehat{A}$ to $\widetilde{A}$ at varying levels of known pairwise relationship values. For a full comparison, we have included two smoothing methods: TCS, as previously described, as well as “simple thresholding”, wherein the elements of $\widehat{A}$ are directly thresholded. (The latter approach is a degenerate case of TCS models, at level $\ell = 0$, for which the basis is the Dirac basis, i.e. $\mathbf{v}_i = \widehat{\mathbf{v}}_i = \delta_i$ for $i = 1, \ldots, N$ in Eqs. 8-13.) Moving from left to right in the figures, the true degree of relatedness increases from $R = 4, \ldots, 11$, to no relatedness. Over the entire matrix of estimates, the RMSE is .0055, .0015 and .0019 for the unsmoothed $(\widehat{A})$, TCS $(\widetilde{A}_{\mathrm{t}})$ and simple thresholding $(\widetilde{A}_{\mathrm{S}})$ methods respectively, demonstrating an overall advantage of TCS. As with many shrinkage methods, TCS introduces a slight bias that is reflected in a higher RMSE for closely related individuals. Consequently, TCS has a larger RMSE than the unsmoothed estimate for smaller values of R. Where TCS gains a notable advantage over the unsmoothed estimate is in differentiating between more distantly related individuals and noise. From Figures 3 and 5 we can see that simple thresholding incurs a substantially larger RMSE for closer relationships because it thresholds too aggressively. For $R = 4$, 70% of the pairs are zeroed out, and for $R > 4$ virtually all pairs are zeros out. Naturally, this method has the smallest RMSE for the sample of unrelated pairs because thresholding zeros out all of these entries. Notably, TCS does almost as well in this setting. A direct comparison of RMSE does not fully reflect the true loss incurred in practice. In most genetic studies close relatives are often recorded in pedigrees and hence estimates are not required. Alternatively, considering distant relatives to be unrelated leads to a substantial loss for estimating heritability and most other genetic applications.

Heritability in Health ABC Study. Body Mass Index (BMI) is one of several traits measured as part of the study entitled "Whole Genome Association Study of Visceral Adiposity" as part of the Health Aging and Body Composition (Health ABC) Study. These data are archived in the Database for Genotypes and Phenotypes (dbGaP)[3]. We restrict our attention to those 1644 individuals with self-reported European ancestry. To control for confounding, prior to analysis, we adjust BMI scores by regressing out age, gender and collection site. Our objective is to estimate heritability of BMI from this population sample. Published heritability estimates range from as low as 0.05 to as high as 0.90 (Allison et al., 1996); however, based on estimates derived from known pedigrees, the heritability of BMI is estimated to be approximately 50-75% (Kangas-Kontio et al., 2010; Zabaneh et al., 2009).

From the full sample of SNPs (Illumina 1M platform) we remove those with missingness greater than .1% and MAF $< .01$. From these we select a subpanel of $90,000$ SNPs, chosen to be nearly evenly spaced. Based on these SNPs, we calculated the relationship matrix $\widehat{A}$, and find that the individuals are predominately unrelated. The most highly related pair appear to be third degree relatives, such as first cousins. And more than half of the pairs appear to be more distantly related than 10'th degree relatives.

To estimate the heritability in this setting, we input the smoothed relationship matrix in Eq. 11 into the REML algorithm. The required smoothing parameter λ is selected in two ways: (i) minimizing the loss function in Eq. 14 via the subsampling approach; and (ii) using a profile likelihood approach. With both techniques, we get estimates of the heritability that are very close to what is found in the literature.

For a range of smoothing parameters, $0 \leq \lambda \leq .40$, we calculate the smoothed relationship matrix, $\widetilde{A}_\lambda$, and plug this value into the REML model to obtain a profile likelihood (Figure 6). Also plotted in this figure is $\widehat{h}^2_\lambda$, the heritability that maximizes REML as a function of λ (or minimizes -2 times the log-likelihood). Without smoothing ($\lambda = 0$), which is not shown in the plot, $\widehat{h}^2 = 0.23$. Smoothing the relationship matrix results in an increasing estimate of the heritability which stabilizes at about 70%. Further smoothing beyond the range displayed leads to a numerically unstable optimization problem and diminished likelihood. Using the profile likelihood approach, λ

[3] http://www.ncbi.nlm.nih.gov/gap

is chosen to be the point at which REML is maximized. This method results in an estimate of $\widehat{\lambda} = .20$ corresponding to $\widehat{h}^2 = 0.71$. Smoothing using our SNP subsampling scheme results in $\widehat{\lambda} = .18$ and $\widehat{h}^2 = 0.72$.

For comparison, we have repeated the above experiments with an orthogonal basis computed by principal component analysis (PCA) in lieu of a treelet basis. Such an approach does not improve the estimates of family relationships or heritability. When noise is present, PCA is unable to uncover the underlying sparse structure of the relationship matrix. In fact, the results with PCA are identical to those without smoothing (with the profile likelihood peaking when the tuning parameter is set to 0).

Another trait that was measured in this study is Abdomen Visceral Fat Density (AVFD). As was the case with BMI, we restricted our attention to individuals of European descent and regressed out age, sex and collection site. According to the literature, the heritability of AVFD should be between 45-70% (Katzmarzyk, Perusse and Bouchard, 1999). According to Figure 6, one can see that using the smoothing parameter based on our subsampling scheme ($\widehat{\lambda} = .18$) we get $\widehat{h}^2 = .29$. On the other hand, exploiting the profile likelihood plot results in $\widehat{\lambda} = .09$ and $\widehat{h}^2 = .36$. When no smoothing was used (not shown in figure), $\widehat{\lambda} = .11$. Thus, both methods for choosing the smoothing parameter used in TCS resulted in estimates of the heritability that are closer to what's established in the literature than without smoothing.

It is notable that $\widehat{h}^2$ for both traits increased towards the established estimate of heritability regardless of how we estimate the optimal smoothing parameter, because only a small fraction of the genome was sampled by the SNP panel. Thus, our results underscore the fact that the quantitative trait model given in Eq. 5 does not require measurement of the causal SNPs that constitute Eq. 3. What is required is a good estimate of A based on relatives.

Our analysis of BMI and AVFD illustrates the difference between estimates of heritability in the traditional sense and estimates of h_s^2, the heritability attributable to the SNPs in the panel. From Eqs. 6 and 7 it is clear that heritability derived from the classic quantitative traits model can distinguish between variance explained by relatives and variance explained by causal SNPs only if either (i) all causal SNPs are excluded, or (ii) all relatives are excluded. Because a large number of undiscovered SNPs scattered across the genome are likely to be causal, and large samples invariably contain distantly related individuals, some ambiguity will always be present.

Clearly the $90,000$ SNPs in our panel do not explain a substantial fraction of the variation in BMI and yet we obtain an accurate estimate of heritability using TCS. The increase in estimated heritability of BMI from 23% to 72% suggests that smoothing improves the estimate of A and that a substantial fraction of the correlation in BMI in our sample is due to genetic relatedness. In a similar study with a larger population sample Yang et al. (2011) estimated h_S^2 of BMI at 17% when using the full SNP panel, but excluding all detectable relatives. Assuming relatives were successfully removed, they conclude that approximately 17% of the variability in BMI is explained by common variants included or tagged by the SNP panel.

Discussion. Recently, there has been an upsurge of papers on sparse covariance matrix estimation; see Bickel and Levina (2008), Cai and Liu (2011) and references within. Most of this research concerns the problem of estimating population covariance matrices from samples of multivariate data in the "large p–small n" regime using banding or thresholding techniques in the original coordinate system. Our setting is slightly different with a more complex data structure: We want to improve estimates of a large covariance matrix (A) in which we expect a hierarchical block structure due to clustering of distantly related individuals. A noisy estimate of covariance is obtained from a large sample of SNPs, each of which contains very little information. This matrix is interpreted as the additive genetic relationship matrix and it can be used to infer degree of relationship between pairs of individuals.

We propose a new method, which we call treelet covariance smoothing (TCS), for regularizing real symmetric matrices with hierarchical block structure and unstructured noise. We show how a subsampling strategy applied to SNPs can be used to chose the tuning parameter for the smoothing procedure. For simulated data, we show that TCS does indeed improve estimates of family relationships. As an application we show how TCS can be used to estimate heritability of quantitative traits from a genome-wide sample of SNPs by smoothing relationships estimated from those SNPs. We then apply TCS to the problem of estimating the heritability of body mass index (BMI) and abdomen visceral fat density (AVFD) in the Health ABC data set. In particular, BMI heritability is usually quoted to be at least 0.50, but an estimate based on a noisy estimate of A yields a much lower value of 0.23. By denoising the estimated relationship matrix with treelets, we increase the estimated heritability of BMI from 0.23 to 0.72. AVFD heritability analysis produces similar results. Thus, a careful examination of heritability estimates using more distant relatives demonstrates that one may substantially improve relationship estimates using TCS.

Other covariance regularization schemes exist in the literature, but systematic comparison is beyond the scope of this work. Direct application of regularization methods for a sample covariance matrix ($Z_c Z_c^t$) is sometimes further complicated if we do not have direct access to the multivariate data matrix Z_c. Cai and Liu (2011), for example, describe a state-of-the-art adaptive thresholding method for heteroscedastic problems that requires an estimate of the variability of the entries of a sample covariance matrix. To our knowledge, TCS is the only principled approach to regularization of general similarity matrices with block structure on multiple scales. In addition, the computed basis vectors themselves contain information of the internal structure of the data – a topic that we will explore in a separate paper with applications to complex extended pedigrees. One can also easily modify the TCS algorithm so that positive semi-definiteness is always guaranteed.

Our results are relevant to a recent area of burgeoning interest in genetics, namely, the estimation of heritability from population samples (Yang et al., 2010a). However our purpose is to estimate heritability, as traditionally defined, rather than to determine the fraction of variation explained by measured SNPs. We expect that the TCS-refined genetic relationships will find wide application to other problems in genetics, such as population-based linkage analysis (Day-Williams et al., 2011), along with linear mixed models for testing association (Kang et al., 2010).

Furthermore, TCS can be applied to a whole family of mixed effects "error-in-variables" models of the form

$$\mathbf{y} = W\beta + Z\mathbf{u} + \mathbf{e}, \tag{15}$$

where $\mathbf{y} \in \mathbb{R}^n$ is a vector of response variables; $\beta \in \mathbb{R}^p$ is a vector of fixed effects; $\mathbf{u} \in \mathbb{R}^q$ represents random effects; and $\mathbf{e} \in \mathbb{R}^n$ is a vector of residual errors. In the general case, we assume that there are c random effects, where each random effect originates from a specific distribution with zero mean and unknown variance. In vector-matrix notation:

$$\mathbf{u} = \begin{pmatrix} \mathbf{u}_1 \\ \vdots \\ \mathbf{u}_c \end{pmatrix} \quad \text{and} \quad Z = (Z_1, \ldots, Z_c)$$

where $\mathbf{u}_i$ is a $q_i \times 1$ vector whose elements are the levels of the ith random factor, $q = q_1 + \ldots + q_c$, and Z_i is an $n \times q_i$ matrix of regressors for the ith random factor. Assuming $\mathbb{E}(\mathbf{u}) = \mathbb{E}(\mathbf{e}) = \mathbf{0}$ and

$$\text{Var}\begin{bmatrix} \mathbf{u} \\ \mathbf{e} \end{bmatrix} = \begin{bmatrix} D & 0 \\ 0 & E \end{bmatrix},$$

where $D = \text{diag}(\sigma_1^2 I_{q_1}, \ldots, \sigma_c^2 I_{q_c})$, yields $\mathbb{E}[\mathbf{y}] = W\beta$ and

$$\text{Var}[\mathbf{y}] = ZDZ^t + E = \sum_{i=1}^{c} \sigma_i^2 Z_i Z_i^t + E$$

where the variance components $\sigma_1^2, \ldots, \sigma_c^2$ and E are unknown and to be estimated. Now consider an *error-in-variables* scenario in which the matrix W of regressors of fixed effects is known, but we only have *noisy* estimates of some or all of the positive semi-definite (psd) matrices $Z_i Z_i^t$ associated with the random effects. If these matrices have block structure and the noise is unstructured, then one could potentially improve estimates of variance components by first applying TCS. In our application, for example, we looked at a special case where we first estimate the psd matrix $Z_c Z_c^t$ in an additive polygenic model using marker-based data, and then use a denoised estimate of $Z_c Z_c^t$ to estimate the variance components, σ_g^2 and σ_e^2 in a random effects model where $D = \sigma_g^2 I$ and $E = \sigma_e^2 I$.

In summary, we have introduced a new method, called Treelet Covariance Smoothing (TCS), that regularizes a relationship matrix estimated from a large panel of genetic markers. In the context of a GWAS study a huge number of SNPs are measured, each of which provides information about the relationship between individuals in the sample. We proposed a SNP subsampling procedure that exploits this rich source of information to choose a tuning parameter for the algorithm. We illustrated one instance of the utility of such estimates by substituting the resulting smoothed relationship matrix into a random effects model to estimate the heritability of body mass index. While others have used genetically inferred estimates of relatedness from samples of close relatives to estimate heritability, we believe this is the first time such estimates have been applied to population-based samples when the goal is to estimate heritability in the traditional sense.

Acknowledgments. This work was supported by National Institute of Mental Health grant MH057881, ONR grant N0014-08-1-0673 and NSF grant DMS0943577. Funding support for the CIDR Visceral Adiposity Study (Study accession number: phs000169.v1.p1) was provided through the Division of Aging Biology and the Division of Geriatrics and Clinical Gerontology, NIA. The CIDR Visceral Adiposity Study includes a genome-wide association study funded as part of the Division of Aging Biology and the Division of Geriatrics and Clinical Gerontology, NIA. Assistance with phenotype harmonization and genotype cleaning, as well as with general study coordination, was provided by Health ABC Study Investigators. We would like to thank Daniel Weeks, Nadine Melhem and Cosma Shalizi for comments on the manuscript, Elizabeth Thompson for guidance in designing the simulations, and Evan Klei for assistance with the simulations.

References.

Albers, C. A., Stankovich, J., Thomson, R., Bahlo, M. and Kappen, H. J. (2008). Multipoint approximations of identity-by-descent probabilities for accurate linkage analysis of distantly related individuals. *The American Journal of Human Genetics* **82** 607–622.

Allison, D., Kaprio, J., Korkeila, M., Koskenvuo, M., Neale, M., Hayakawa, K. et al. (1996). The heritability of body mass index among an international sample of monozygotic twins reared apart. *International journal of obesity* **20** 501–506.

Almasy, L. and Blangero, J. (1998). Multipoint quantitative-trait linkage analysis in general pedigrees. *The American Journal of Human Genetics* **62** 1198–1211.

Anderson, A. D. and Weir, B. S. (2007). A maximum likelihood method for estimation of pairwise relatedness in structured populations. *Genetics* **176** 421–420.

Astle, W. and Balding, D. J. (2009). Population structure and cryptic relatedness in genetic association studies. *Statistical Science* **24** 451–471.

Bickel, P. J. and Levina, E. (2008). Regularized estimation of large covariance matrices. *The Annals of Statistics* **36** 199–227.

Boehnke, M. and Cox, N. J. (1997). Accurate inference of relationships in sib-pair linkage studies. *The American Journal of Human Genetics* **61** 423–429.

Bravo, H. C., Lee, K. E., Klein, B. E. K., Klein, R., Iyengar, S. K. and Wahba, G. (2009). Examining the relative influence of familial, genetic, and environmental covariate information in flexible risk models. *Proceedings of the National Academy of Sciences* **106** 8128.

Browning, S. R. (2008). Estimation of pairwise identity by descent from dense genetic marker data in a population sample of haplotypes. *Genetics* **178** 2123.

Browning, S. R. and Browning, B. L. (2010). High-resolution detection of identity by descent in unrelated individuals. *The American Journal of Human Genetics* **86** 526–539.

Cai, T. and Liu, W. (2011). Adaptive thresholding for sparse covariance matrix estimation. *Journal of the American Statistical Association* **106** 672–684.

Choi, Y., Wijsman, E. M. and Weir, B. S. (2009). Case-control association testing in the presence of unknown relationships. *Genetic Epidemiology* **33** 668–678.

Day-Williams, A. G., Blangero, J., Dyer, T. D., Lange, K. and Sobel, E. M. (2011). Linkage analysis without defined pedigrees. *Genetic Epidemiology* **35** 360–370.

Deary, I. J., Yang, J., Davies, G., Harris, S. E., Tenesa, A., Liewald, D., Luciano, M., Lopez, L. M., Gow, A. J., Corley, J. et al. (2012). Genetic contributions to stability and change in intelligence from childhood to old age. *Nature* **482** 212–215.

Devlin, B., Daniels, M. and Roeder, K. (1997). The heritability of IQ. *Nature* **388** 468–471.

Eding, H. et al. (2001). Marker-based estimates of between and within population kinships for the conservation of genetic diversity. *Journal of Animal Breeding and Genetics* **118** 141–159.

Epstein, M. P., Duren, W. L. and Boehnke, M. (2000). Improved inference of relationship for pairs of individuals. *The American Journal of Human Genetics* **67** 1219–1231.

Fisher, R. A. (1918). The correlation between relatives on the supposition of Mendelian inheritance. *Transactions of the Royal Society of Edinburgh* **52** 399–433.

Golub, G. H. and Van Loan, C. F. (1996). *Matrix Computations*. Johns Hopkins Univ Pr.

Hallmayer, J., Cleveland, S., Torres, A., Phillips, J., Cohen, B., Torigoe, T., Miller, J., Fedele, A., Collins, J., Smith, K., Lotspeich, L., Croen, L. A., Ozonoff, S., Lajonchere, C., Grether, J. K. and Risch, N. (2011). Genetic Heritability and Shared Environmental Factors Among Twin Pairs With Autism. *Archives of General Psychiatry* **Epub ahead of print**.

Hayes, B. J. and Goddard, M. (2008). Technical note: Prediction of breeding values using marker-derived relationship matrices. *Journal of Animal Science* **86** 2089.

Henderson, C. R. (1950). Estimation of genetic parameters. *Biometrics* **6** 186–187.

Hopper, J. and Mathews, J. (1982). Extensions to multivariate normal models for pedigree analysis. *Annals of Human Genetics* **46** 373–383.

Imielinski, M., Baldassano, R. N., Griffiths, A., Russell, R. K., Annese, V., Dubinsky, M., Kugathasan, S., Bradfield, J. P., Walters, T. D., Sleiman, P. et al. (2009). Common variants at five new loci associated with early-onset inflammatory bowel disease. *Nature Genetics* **41** 1335–1340.

Kang, H. M., Sul, J. H., Zaitlen, N. A., Kong, S., Freimer, N. B., Sabatti, C., Eskin, E. et al. (2010). Variance component model to account for sample structure in genome-wide association studies. *Nature genetics* **42** 348–354.

Kangas-Kontio, T., Huotari, A., Ruotsalainen, H., Herzig, K. H., Tamminen, M., Ala-Korpela, M., Savolainen, M. J. and Kakko, S. (2010). Genetic and environmental determinants of total and high-molecular weight adiponectin in families with low HDL-cholesterol and early onset coronary heart disease. *Atherosclerosis* **210** 479–485.

Katzmarzyk, P., Perusse, L. and Bouchard, C. (1999). Genetics of abdominal vesceral fat levels. *American Journal of Human Biology* **11** 225–235.

Lander, E. S. and Schork, N. J. (1994). Genetic dissection of complex traits. *Science* **265** 2037.

Lee, A. and Nadler, B. (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*.

Lee, A. B., Nadler, B. and Wasserman, L. (2008). Treelets–An adaptive multi-scale basis for sparse unordered data. *The Annals of Applied Statistics* **2** 435–471.

Lynch, M. and Ritland, K. (1999). Estimation of pairwise relatedness with molecular markers. *Genetics* **152** 1753.

Mallat, S. G. (1999). *A Wavelet Tour of Signal Processing*. Academic Pr.

McGovern, D. P. B., Gardet, A., Törkvist, L., Goyette, P., Essers, J., Taylor, K. D., Neale, B. M., Ong, R. T. H., Lagacé, C., Li, C. et al. (2010). Genome-wide association identifies multiple ulcerative colitis susceptibility loci. *Nature Genetics* **42** 332–337.

McPeek, M. S. and Sun, L. (2000). Statistical tests for detection of misspecified relationships by use of genome-screen data. *The American Journal of Human Genetics* **66** 1076–1094.

Milligan, B. G. (2003). Maximum-likelihood estimation of relatedness. *Genetics* **163** 1153.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., De Bakker, P. I. W., Daly, M. J. et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. *The American Journal of Human Genetics* **81** 559–575.

Searle, S. R., Casella, G., McCulloch, C. E. et al. (1992). *Variance components*. Wiley Online Library.

Thompson, E. (1974). Gene identities and multiple relationships. *Biometrics* **30** 667–680.

Thompson, E. (1975). The estimation of pairwise relationships. *Annals of Human Genetics* **39** 173–188.

Thompson, E. A. (1986). *Pedigree analysis in human genetics*. Johns Hopkins University

Press Baltimore, MD.

Thornton, T. and McPeek, M. S. (2010). ROADTRIPS: case-control association testing with partially or completely unknown population and pedigree structure. *The American Journal of Human Genetics* **86** 172–184.

Visscher, P. M., Medland, S. E., Ferreira, M. A. R., Morley, K. I., Zhu, G., Cornes, B. K., Montgomery, G. W. and Martin, N. G. (2006). Assumption-free estimation of heritability from genome-wide identity-by-descent sharing between full siblings. *PLoS Genetics* **2** e41.

Visscher, P. M., Brown, M. A., McCarthy, M. I. and Yang, J. (2012). Five Years of GWAS Discovery. *The American Journal of Human Genetics* **90** 7–24.

Weir, B. S., Anderson, A. D. and Hepler, A. B. (2006). Genetic relatedness analysis: modern data and new challenges. *Nature Reviews Genetics* **7** 771–780.

Yang, J., Benyamin, B., McEvoy, B. P., Gordon, S., Henders, A. K., Nyholt, D. R., Madden, P. A., Heath, A. C., Martin, N. G., Montgomery, G. W. et al. (2010a). Common SNPs explain a large proportion of the heritability for human height. *Nature Genetics* **42** 565–569.

Yang, J., Lee, S. H., Goddard, M. E. and Visscher, P. M. (2010b). GCTA: a tool for genome-wide complex trait analysis. *The American Journal of Human Genetics* **88** 76–82.

Yang, J., Manolio, T. A., Pasquale, L. R., Boerwinkle, E., Caporaso, N., Cunningham, J. M., De Andrade, M., Feenstra, B., Feingold, E., Hayes, M. G. et al. (2011). Genome partitioning of genetic variation for complex traits using common SNPs. *Nature genetics* **43** 519–525.

Zabaneh, D., Chambers, J., Elliott, P., Scott, J., Balding, D. and Kooner, J. (2009). Heritability and genetic correlations of insulin resistance and component phenotypes in Asian Indian families using a multivariate analysis. *Diabetologia* **52** 2585–2589.

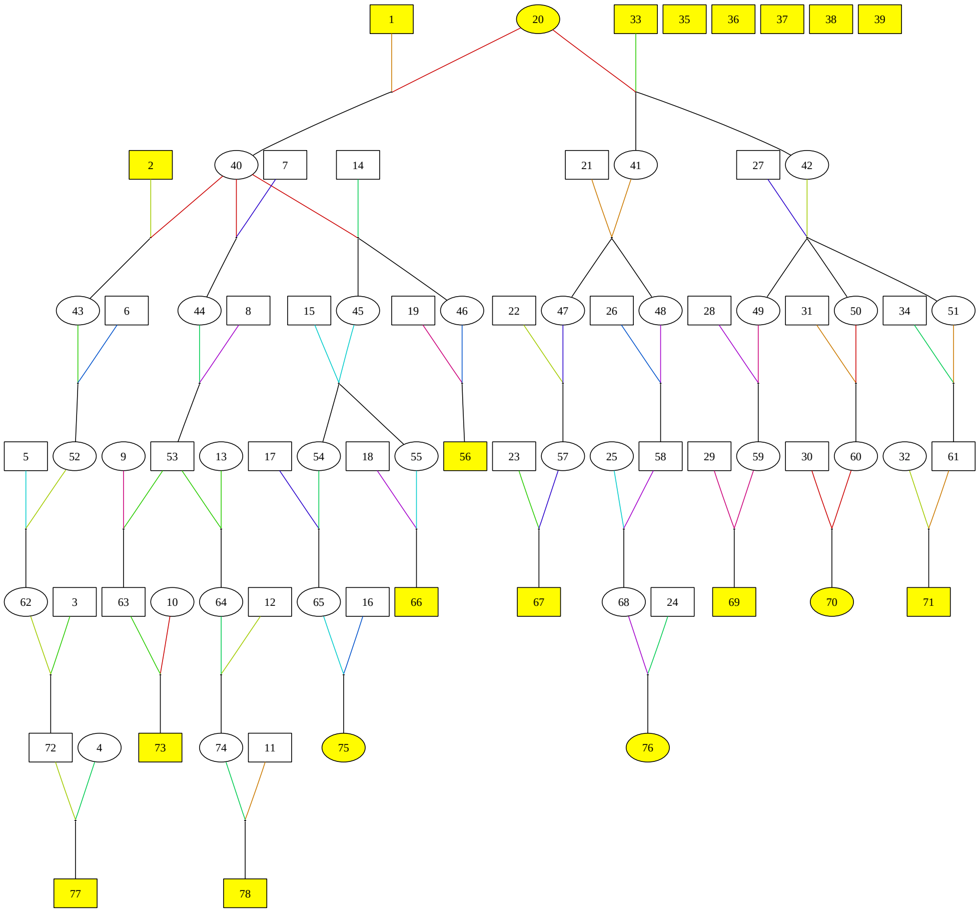


FIG 1. *Pedigree of a single family used for simulations. Genomes were dropped through the entire pedigree and ten individuals were sampled from the twenty possible highlighted individuals. Individuals 35-39 are unrelated to everyone else in the pedigree.*

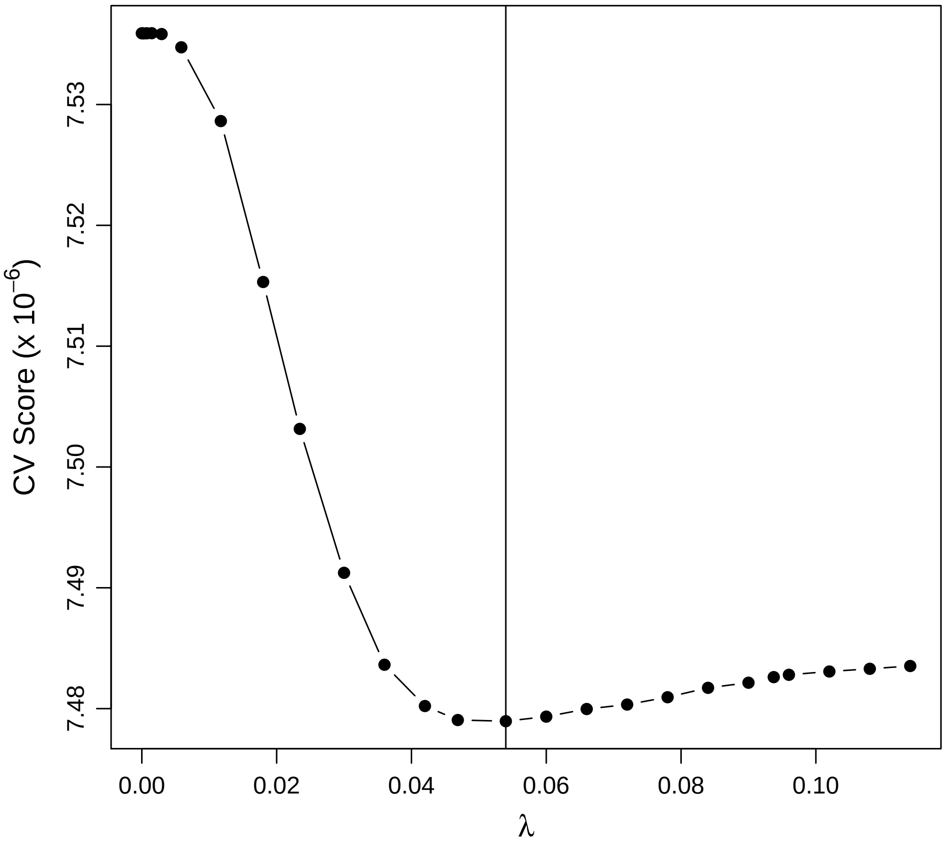


FIG 2. *Cross-validation plot showing the weighted risk function at varying levels of the thresholding parameter, λ. The optimal λ is the point where the $H(\lambda)$ (CV Score) is minimized.*

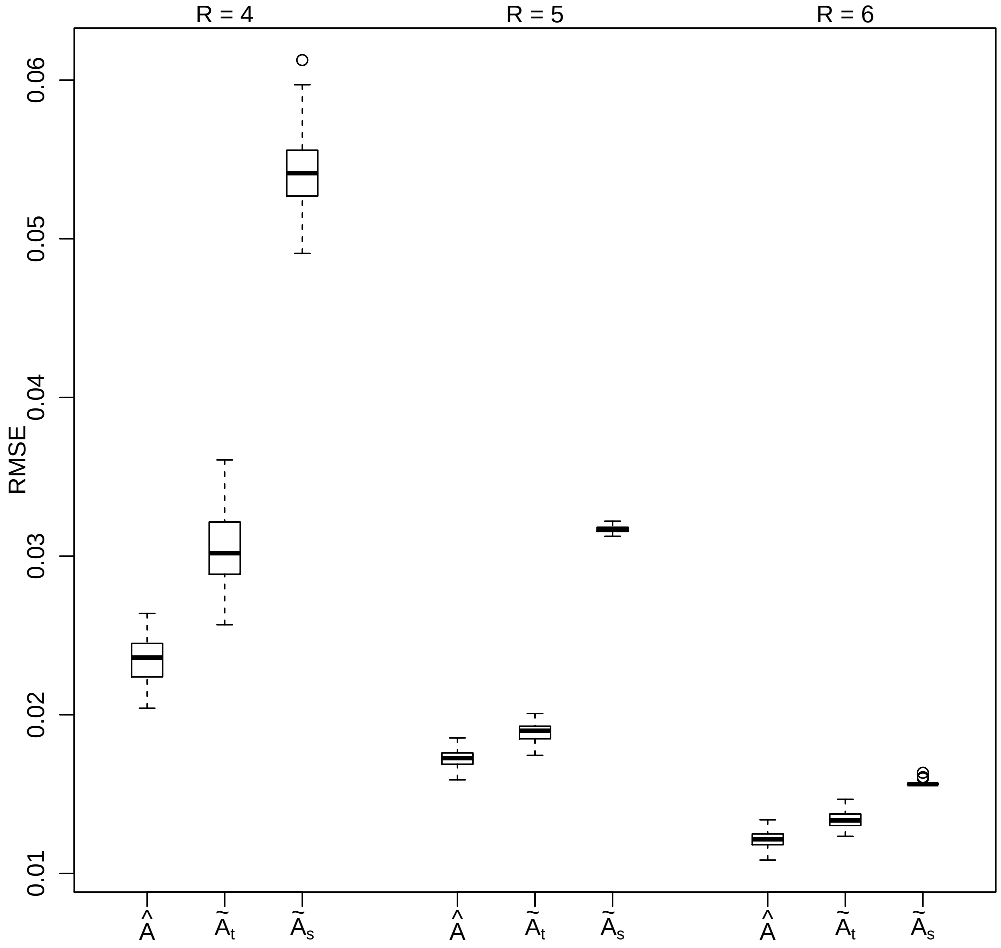


FIG 3. *Boxplots of RMSE for unsmoothed ($\widehat{A}$) along with smoothed using TCS ($\widetilde{A}_t$) and simple thresholding ($\widetilde{A}_S$) at increasing degrees of relatedness ($R = 4, 5, 6$; see header). Here, TCS is better than simple thresholding as the latter method thresholds too aggressively.*

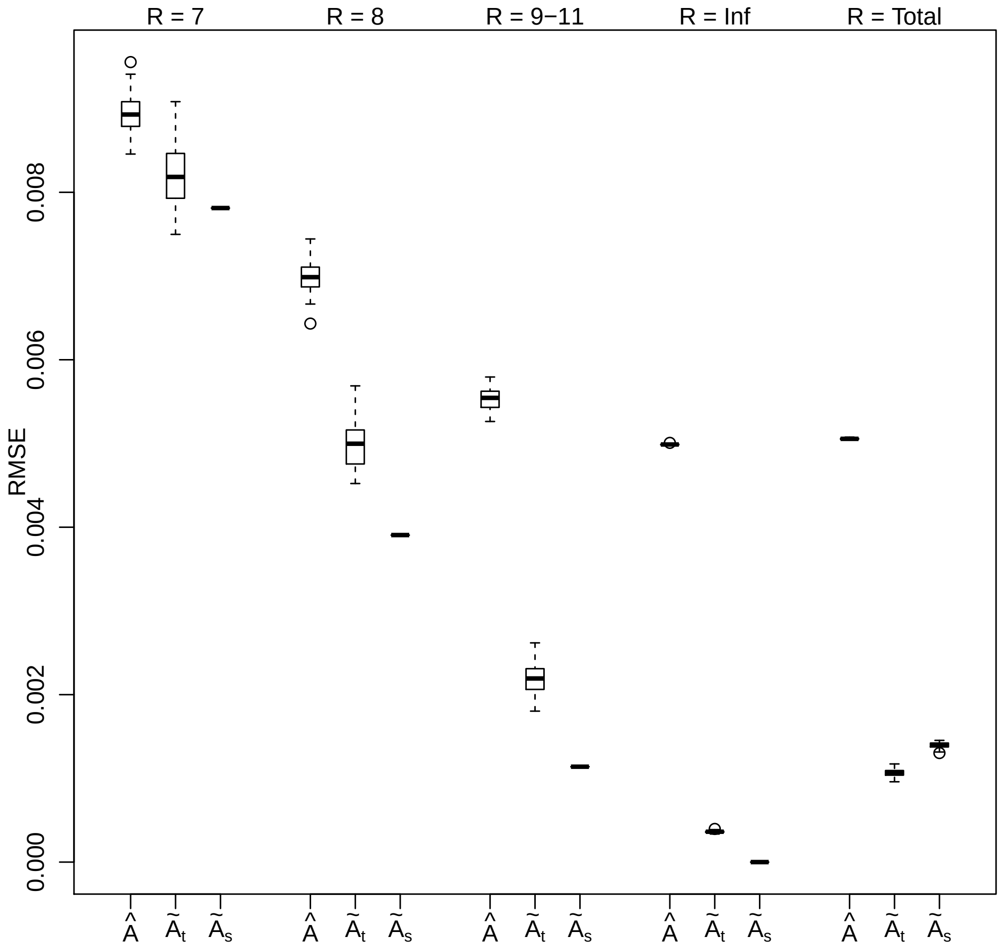


FIG 4. *Boxplots of RMSE for unsmoothed ($\widehat{A}$) along with smoothed using TCS ($\widetilde{A}_t$) and simple thresholding ($\widetilde{A}_S$) at increasing degrees of relatedness ($R = 7, 8, 9 - 11$). Also included is the comparison of RMSE values for unrelated pairs ($R =$ Inf) and average RMSE for the entire relationship matrix ($R =$ Total). We see that both thresholding methods remove noise, but TCS works better than simple thresholding overall.*

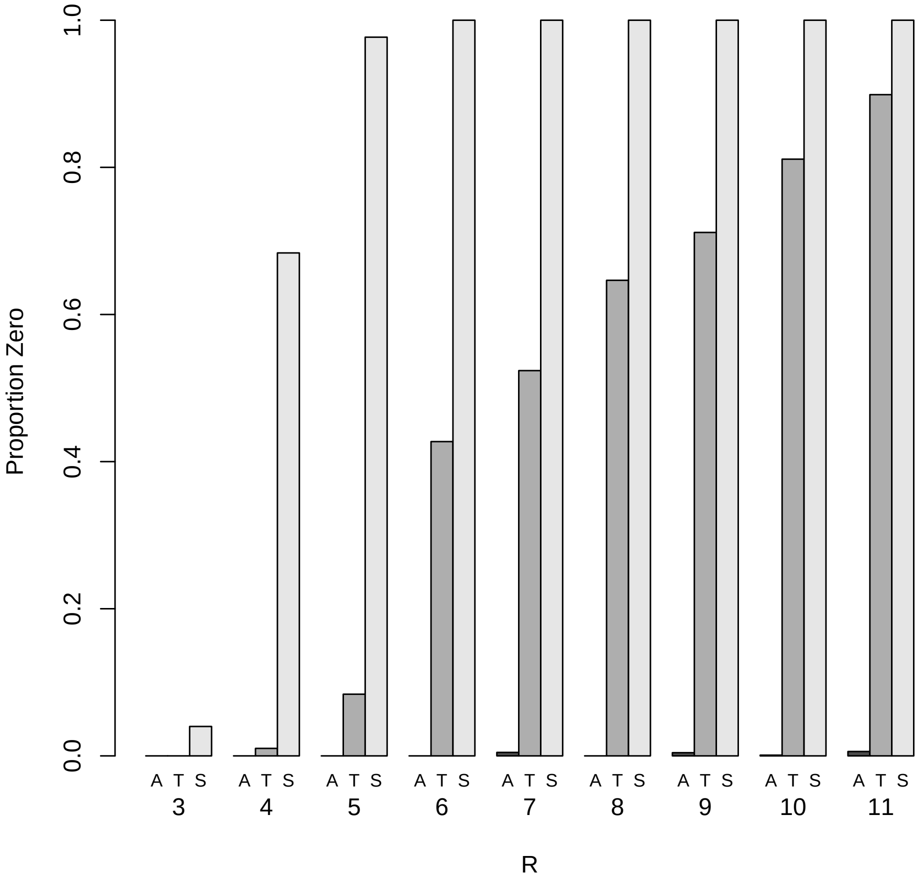


Fig 5. *Barplots of the percentage of relationships that are equal to 0 for no smoothing (A), smoothing using TCS (T) and simple thresholding (S). The three cases are compared at increasing degrees of relatedness ($R = 3, \ldots, 11$). Any value below $\epsilon = 10^{-5}$ is considered to be 0.*

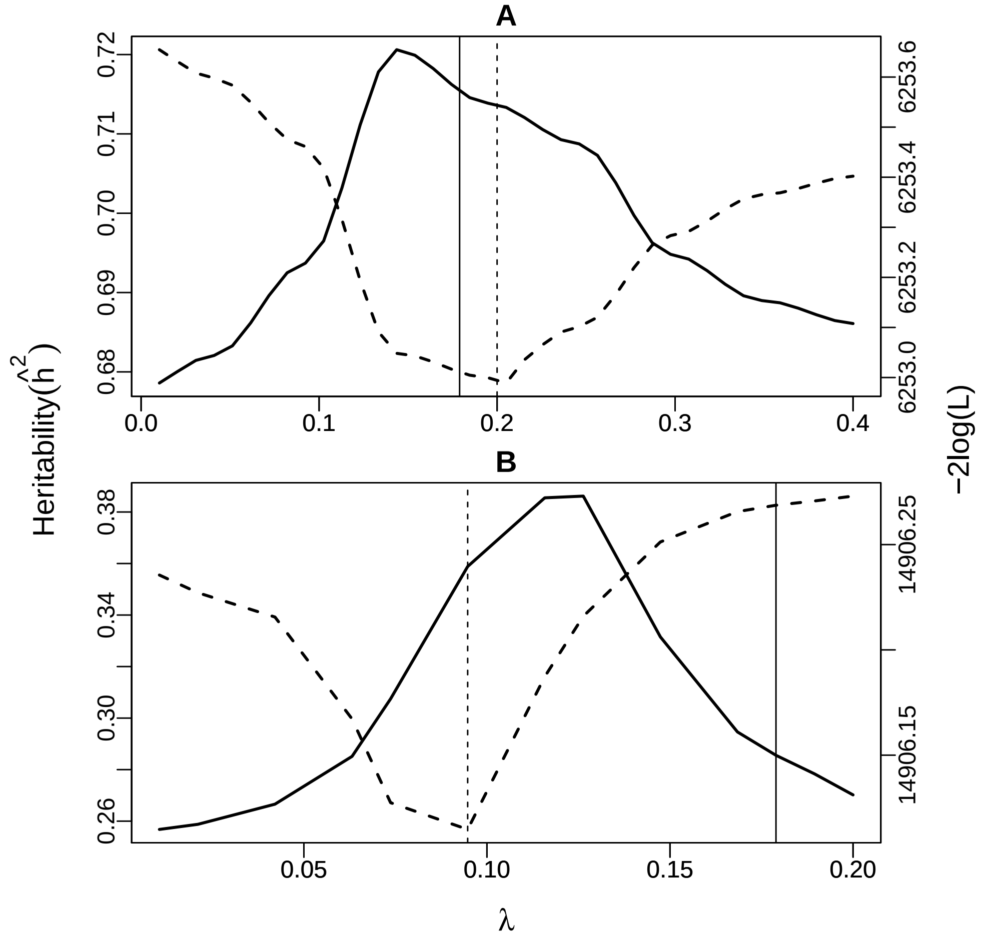


FIG 6. *Estimating heritability in the Health ABC data set. Solid curve is the estimated heritability at increasing values of the smoothing parameter λ. The dashed curve is $\propto -2log(\mathcal{L})$, where, $\mathcal{L}$ is the maximum profile likelihood obtained from the REML algorithm. The solid vertical line is the optimally chosen threshold value using our subsampling scheme. The dashed vertical line represents the optimally chosen threshold value when minimizing the likelihood profile.* **A:** *For BMI, $h^2 = .72$ when using subsampling to choose an optimal smoothing parameter ($\widehat{\lambda} = .18$). Similarly, $h^2 = .71$ when using the profile likelihood plot ($\widehat{\lambda} = .20$). With no smoothing ($\lambda = 0$), $h^2 = .23$. This is not shown on the plot.* **B:** *For AVFD, $h^2 = .29$ when using our subsampling approach to choose an optimal smoothing parameter. However, $h^2 = .36$ when using the profile likelihood plot ($\widehat{\lambda} = .09$). These are compared to $h^2 = .11$, the heritability when there is no smoothing (not shown).*

APPENDIX: THE TREELET ALGORITHM

Hierarchical clustering algorithms offer an easily interpretable description of data in terms of a dendrogram and some measure of similarity between the observations. So called agglomerative hierarchical methods start at the bottom of the tree and merge, at each level, the two groups with the highest inter-group similarity into one larger cluster. The novelty of the treelet algorithm proposed in Lee, Nadler and Wasserman (2008) is the construction of not only clusters or groupings of variables, but also functions on the data. More specifically, treelets construct a multi-scale orthonormal basis on a hierarchical tree. As in standard multi-resolution analysis (MRA) (Mallat, 1999), the algorithm returns 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 $V_\ell \oplus W_\ell = V_{\ell-1}$. These functions are well-localized in space; in fact, they are supported on nested clusters in a hierarchical tree.

The treelet algorithm, as well as its implementation is available in R on CRAN as the *treelet* library. Here we summarize the original published algorithm; see Lee, Nadler and Wasserman (2008) for a theoretical analysis. 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". In this paper, we measure similarity by Pearson's correlation coefficient but other choices are also possible. The new variables are then computed by a local PCA or Jacobi rotation in two dimensions. Unlike Jacobi's classical method for eigendecompositions (Golub and Van Loan, 1996), difference variables (i.e. 2nd local principal components) are stored, and *only* sum variables (i.e. 1st local principal components) are processed at higher levels of the tree. Hence, the multi-resolution analysis. The details of the complete treelet algorithm are as follows:

- At level $\ell = 0$ (the "bottom" of the tree), each observation or "signal" $\mathbf{z}$ is represented by the original variables $\mathbf{z}^{(0)} = (s_{0,1}, \ldots, s_{0,N})^t$, where $s_{0,k} = z_k$. Associate to these coordinates, the Dirac basis $B_0 = (\phi_{0,1}, \phi_{0,2}, \ldots, \phi_{0,N})$ where B_0 is the $N \times N$ identity matrix and $\phi_{0,i}$ are unit vectors. Compute estimates of the covariance and similarity matrices, $\widehat{\Sigma}^{(0)}$ and $\widehat{M}^{(0)}$, where $\widehat{M}^{(0)}_{ij} = \frac{\widehat{\Sigma}^{(0)}_{ij}}{\sqrt{\widehat{\Sigma}^{(0)}_{ii}\widehat{\Sigma}^{(0)}_{jj}}}$. Initialize the index set of "sum variables", $\mathcal{S}_0 = \{1, 2, \ldots, N\}$.
- Repeat for $\ell = 1, \ldots, L$, where $L \leq N - 1$.

  1. **Find the two most similar sum variables according to the**

similarity matrix $\widehat{M}^{(\ell-1)}$. Let

$$\{\alpha_\ell, \beta_\ell\} = \arg \max_{i,j \in \mathcal{S}_{\ell-1}} \widehat{M}_{ij}^{(\ell-1)} \ . \tag{16}$$

where $i < j$, and maximization is only over pairs of sum variables that belong to the set $\mathcal{S}_{\ell-1}$. 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

$$J(\alpha_\ell, \beta_\ell, \theta_\ell) = \begin{bmatrix} 1 & \cdots & 0 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & \vdots & & \vdots & & \vdots \\ 0 & \cdots & c & \cdots & -s & \cdots & 0 \\ \vdots & & \vdots & \ddots & \vdots & & \vdots \\ 0 & \cdots & s & \cdots & c & \cdots & 0 \\ \vdots & & \vdots & & \vdots & \ddots & \vdots \\ 0 & \cdots & 0 & \cdots & 0 & \cdots & 1 \end{bmatrix} \tag{17}$$

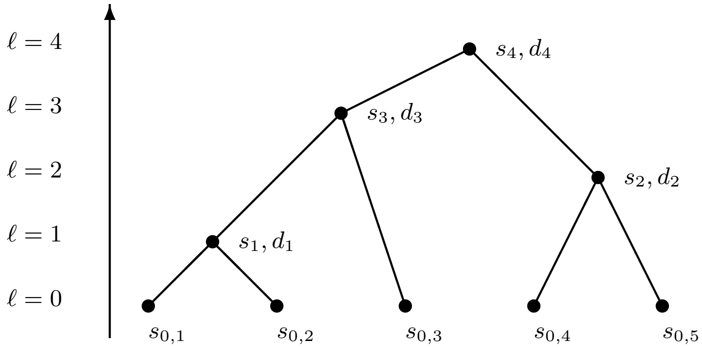
where $c = \cos(\theta_\ell)$ and $s = \sin(\theta_\ell)$, that decorrelates z_α and z_β; more specifically, find a rotation angle θ_ℓ such that $|\theta_\ell| \leq \pi/4$ and $\widehat{\Sigma}_{\alpha\beta}^{(\ell)} = \widehat{\Sigma}_{\beta\alpha}^{(\ell)} = 0$, where $\widehat{\Sigma}^{(\ell)} = J^t \widehat{\Sigma}^{(\ell-1)} J$. This transformation corresponds to a change of basis $B_\ell = B_{\ell-1} J$, and new coordinates $\mathbf{z}^{(\ell)} = J^t \mathbf{z}^{(\ell-1)}$.

Update the similarity matrix $\widehat{M}^{(\ell)}$ accordingly.

3. **Multi-resolution analysis.** For ease of notation, assume that $\widehat{\Sigma}_{\alpha\alpha}^{(\ell)} \geq \widehat{\Sigma}_{\beta\beta}^{(\ell)}$ after the Jacobi rotation, where the indices α and β denote the first and second principal components, respectively. Define the sum and difference variables at level ℓ as $s_\ell = z_\alpha^{(\ell)}$ and $d_\ell = z_\beta^{(\ell)}$. Similarly, define the scaling and detail functions ϕ_ℓ and ψ_ℓ as columns α_ℓ and β_ℓ of the basis matrix B_ℓ. Remove the difference variable from the set of sum variables, $\mathcal{S}_\ell = \mathcal{S}_{\ell-1} \setminus \{\beta_\ell\}$. At level ℓ, we have the orthonormal *treelet decomposition*

$$\mathbf{z} = \sum_{i=1}^{N-\ell} s_{\ell,i} \phi_{\ell,i} + \sum_{i=1}^{\ell} d_i \psi_i. \tag{18}$$

where the new set of scaling vectors $\{\phi_{\ell,i}\}_{i=1}^{N-\ell}$ is the union of the vector ϕ_ℓ and the scaling vectors $\{\phi_{\ell-1,j}\}_{j \neq \alpha,\beta}$ from the previous level, and the new coarse-grained sum variables $\{s_{\ell,i}\}_{i=1}^{N-\ell}$ 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.



$$\begin{array}{l}
\mathbf{z}^{(4)} = (\; s_4 \quad d_1 \quad d_3 \quad d_4 \quad d_2 \;)^t \\
\mathbf{z}^{(3)} = (\; s_3 \quad d_1 \quad d_3 \quad s_2 \quad d_2 \;)^t \\
\mathbf{z}^{(2)} = (\; s_1 \quad d_1 \quad s_{0,3} \quad s_2 \quad d_2 \;)^t \\
\mathbf{z}^{(1)} = (\; s_1 \quad d_1 \quad s_{0,3} \quad s_{0,4} \quad s_{0,5} \;)^t \\
\mathbf{z}^{(0)} = (\; s_{0,1} \quad s_{0,2} \quad s_{0,3} \quad s_{0,4} \quad s_{0,5} \;)^t
\end{array}$$

FIG 7. *(***Left***) A toy example of a hierarchical tree for data of dimension $N = 5$. At $\ell = 0$, the signal is represented by the original N variables. At each successive level $\ell = 1, 2, \ldots, N-1$ the two most similar sum variables are combined and replaced by the sum and difference variables (s_ℓ, d_ℓ) corresponding to the first and second local principal components. (***Right***) Signal representation $\mathbf{z}^{(\ell)}$ at different levels. The s- and d-coordinates represent projections along scaling and detail functions in a multi-scale treelet decomposition. Each such representation is associated with an orthogonal basis in $\mathbb{R}^N$ that captures the local eigenstructure of the data. (This figure has been adopted from Lee, Nadler and Wasserman (2008).)*

Fig. 7 (left) shows an example of a treelet construction for a "signal" of length $N = 5$, with the data representations $\mathbf{z}^{(\ell)}$ at the different levels of the tree shown on the right. The s-components (projections in the main principal directions) represent coarse-grained "sums". We associate these variables to the nodes in the cluster tree. Similarly, the d-components (projections in the orthogonal directions) represent "differences" between node representations at two consecutive levels in the tree. For example, in the figure, $d_1\psi_1 = (s_{0,1}\phi_{0,1} + s_{0,2}\phi_{0,2}) - s_1\phi_{1,1}$.

The output of the algorithm can be compactly summarized in terms of an ordered set of rotations and pairs of indices $\{\theta_\ell, \alpha_\ell, \beta_\ell\}_{\ell=1}^{L}$, where $L \leq N-1$ is the height of the associated hierarchical tree.

In terms of computational complexity, a naive implementation of the treelet algorithm with an exhaustive nearest neighbor search corresponds to a total of $O(N^3)$ operations. By using a fast nearest neighbor search that is linear in time, one can further reduce the computational cost to $O(N^2)$.