DEPARTMENT OF STATISTICS
University of Wisconsin
1300 University Ave.
Madison, WI 53706

TECHNICAL REPORT NO. 1172
September 21, 2012

Using distance correlation and SS-ANOVA to assess associations of familial relationships, lifestyle factors, diseases and mortality

Jing Kong[1]
Department of Statistics
University of Wisconsin, Madison

Barbara E.K. Klein[2], Ronald Klein[2], Kristine E. Lee[3]
Department of Epidemiology and Visual Science
University of Wisconsin, Madison

Grace Wahba[1]
Department of Statistics, Department of Computer Sciences
and Department of Biostatistics and Medical Informatics
University of Wisconsin, Madison

[1]Research supported in part by NIH Grant EY09946 and NSF Grant DMS-0906818
[2]Supported in part by NIH Grant EY06594, and by the Research to Prevent Blindness Senior Scientific Investigator Awards, New York, NY.
[3]Supported in part by NIH Grant EY06594.

Using distance correlation and SS-ANOVA to assess associations of familial relationships, lifestyle factors, diseases and mortality

Jing Kong [*], **Barbara E. K. Klein** [†], **Ronald Klein** [†] **Kristine E. Lee** [†] **and Grace Wahba** [‡]

[*]Department of Statistics, University of Wisconsin, Madison WI 53706,[†]Department of Ophthalmology, University of Wisconsin, Madison WI 53706, and [‡]Departments of Statistics, Biostatistics and Medical Informatics and Computer Sciences, University of Wisconsin, Madison WI 53706

We present a method for examining mortality as it is seen to run in families, and lifestyle factors that are also seen to run in families, in a subpopulation of the Beaver Dam Eye Study that has died by 2011. We observe that pairwise distance between death age in related persons is on average less than pairwise distance in death age between random pairs of unrelated persons. Our goal is to examine the hypothesis that pairwise differences in lifestyle factors correlate with the observed pairwise differences in death age that run in families. Szekely and coworkers have recently developed a method called distance correlation, that is suitable for this task with some enhancements relevant to the particular task at hand. We build a Smoothing Spline ANOVA (SS-ANOVA) model for predicting death age based on four major lifestyle factors generally known to be related to mortality and four of the major diseases contributing to mortality, to develop a lifestyle mortality risk vector and a disease mortality risk vector. We then examine to what extent pairwise differences in these scores correlate with the pairwise differences in mortality as they occur between family members and between unrelated persons. We find significant distance correlations between death ages, lifestyle factors, and family relationships. Considering only sib pairs compared to unrelated persons, distance correlation between siblings and mortality is, not surprisingly, stronger than that between more distantly related family members and mortality. The overall methodological approach here easily adapts to exploring relationships between multiple clusters of variables with observable (real-valued) attributes, and other factors for which only possibly nonmetric pairwise dissimilarities are observed.

pedigrees | lifestyle | mortality | distance correlation | SS-ANOVA | RKE

Introduction

Multiple studies have reported that collectively lifestyle factors, including smoking, low or high body mass index (bmi), low educational attainment and low socio-economic status, are associated with earlier mortality. Diseases, such as diabetes, cardiovascular disease, cancer and chronic kidney diseases, are leading causes of death. Longevity is generally believed to run in families. Furthermore, there is evidence showing that the lifestyle factors all tend to run in families. The goal of this paper is to capture the association of familial relationships, lifestyle factors, diseases and mortality. It is possible that some of the lifestyle variables may be or turn out to be related to genetic factors. Current research interest involves searches for "longevity genes" but this work is not related to that quest. We are not assessing to what extent genetics is involved in longevity.

The Beaver Dam Eye Study (BDES) [8] is an ongoing population-based study of age-related ocular disorders. Subjects at baseline, examined between 1988 and 1990, were a group of 4926 people aged 43-86 years who lived in Beaver Dam, WI. Many group members have relatives in the study, and pedigree information was collected. Mortality information was updated to March 2011. BDES provides an excellent opportunity to attempt to examine and quantify the above associations.

A pair of landmark papers [16, 15] proposed the distance correlation as a measurement of multivariate independence, and others have recently built upon it [14, 9, 6, 12]. The method is extremely general in that it is applicable to random vectors of arbitrary and not necessarily equal dimension and only involves Euclidean pairwise distance. If the two variables are sampled from a bivariate normal distribution, the distance correlation behaves very much like the Pearson's correlation coefficient. Since only Euclidean pairwise distances enter, the method may be applied to inherently unobservable variables with only Euclidean pairwise distances observable. The "genetic distances" defined on pairs of persons representing their familial relationships are generally not Euclidean. However, it is shown that the use of genetic dissimilarity in the distance correlation is still validated since the genetic dissimilarity can be well approximated by Euclidean pairwise distances obtained by embedding the subjects into Euclidean spaces through Regularized Kernel Estimation (RKE) [11, 1].

Smoothing Spline ANOVA (SS-ANOVA) models have a successful history for modeling various aspects of BDES data, two examples are [18, 3]. In this study, we focus on modeling the mortality (death ages) of the form

$$\begin{aligned} death\ age_i =& g_0(baseline\ age_i, gender_i) + \\ & g_1(lifestyle\ factor_i) + g_2(disease_i), \end{aligned}$$

where g_0 is a term involves fixed characteristics, baseline age and gender, for the individuals, g_1 is a term that includes only lifestyle factors and g_2 is a term containing only disease variables, namely diabetes, cancer, cardiovascular disease and chronic kidney disease. In the paper, the fitted values of g_1 and g_2 are treated as scores for the individuals and to be used to assess the association with familial relationships.

Pedigrees

The genetic relationships between pedigree members can be described by Malecot's [13] kinship coefficient φ which defines a pedigree dissimilarity measure. The kinship coefficient φ between individuals i and j in the pedigree is defined as the probability that a randomly selected pair of alleles, one from each individual, is identical by descent, that is, they are derived from a common ancestor. For a parent-offspring pair,

Reserved for Publication Footnotes

$\varphi_{ij} = 0.25$ since there is a 50% chance that the allele inherited from the parent is chosen at random for the offspring, and a 50% chance that the same allele is chosen at random for the parent.

Pedigree Dissimilarity The pedigree dissimilarity between individuals i and j is defined for this study as $d_{ij} = 1 - 2\varphi_{ij}$, where φ is the kinship coefficient. Thus, for $i \neq j$, the pedigree dissimilarity here falls in the interval $[\frac{1}{2}, 1]$. Note that [1] define pedigree dissimilarity for that study as $-log_2(2\varphi)$, which ranges from 1 to ∞ for $i \neq j$, which is not appropriate for the way we will be using pedigree dissimilarity.

In BDES, not all family members are included in the study and not all the subjects have pedigree records.

Smoothing-Spline ANOVA Models

SS-ANOVA models [17, 5, 19] estimate the responses $y_i, i = 1, ..., n$ to be a function of the covariates $f(x_i)$, by assuming that f is a function in a reproducing kernel Hilbert space (RKHS) of the form $\mathcal{H} = \mathcal{H}_0 \bigoplus \mathcal{H}_1$. $\mathcal{H}_0$ is a finite dimensional space spanned by a set of functions $\{\phi_1, ..., \phi_m\}$, and $\mathcal{H}_1$ is an RKHS induced by a given kernel function $k(\cdot, \cdot)$ with the property that $\langle k(x_i, \cdot), k(x_j, \cdot)\rangle_{\mathcal{H}_1} = k(x_i, x_j)$. Thus, the function f has a semiparametric form of

$$f(x) = \sum_{j=1}^{m} d_j \phi_j(x) + g(x),$$

for some coefficients d_j, where the functions ϕ_j's are of parametric linear form and $g \in \mathcal{H}_1$. $\mathcal{H}_1$ is further decomposed by assuming that it is the direct sum of multiple RKHSs. Hence, $g \in \mathcal{H}_1$ is defined to be

$$g(x) = \sum_{\alpha} g_\alpha(x_\alpha) + \sum_{\alpha<\beta} g_{\alpha\beta}(x_\alpha, x_\beta) + \cdots,$$

where $\{g_\alpha\}$ and $\{g_{\alpha\beta}\}$ satisfy side conditions that generalize the standard ANOVA side conditions. Functions g_α are the "main effects" and $g_{\alpha\beta}$ are the "second-order interactions", and so on. The RKHS $\mathcal{H}_\alpha$ is associated with each component in the above sum, along with its corresponding kernel function k_α. In this case, the reproducing kernel function for $\mathcal{H}_1$ is defined to be

$$k(\cdot, \cdot) = \sum_{\alpha} \theta_\alpha k_\alpha(\cdot, \cdot) + \sum_{\alpha<\beta} \theta_{\alpha\beta} k_{\alpha\beta}(\cdot, \cdot) + \cdots,$$

where the coefficients θ's are tuning parameters that weigh the relative importance of each term in the decomposition.

The SS-ANOVA estimates f given data $\{(x_i, y_i), i = 1, ..., n\}$ by the solution of a penalized likelihood problem of the form

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} l(y_i, f(x_i)) + J_{\lambda,\theta}(f), \qquad [1]$$

where $l(y_i, f(x_i)) = (y_i - f(x_i))^2$ and

$$J_{\lambda,\theta}(f) = \lambda[\sum_{\alpha} \theta_\alpha^{-1} \|P_\alpha f\|^2_{\mathcal{H}_\alpha} + \sum_{\alpha<\beta} \theta_{\alpha\beta}^{-1} \|P_{\alpha\beta} f\|^2_{\mathcal{H}_{\alpha\beta}} + \cdots],$$

with $P_\alpha f$ the projection of f into RKHS $\mathcal{H}_\alpha$ and λ a nonnegative regularization parameter. The penalty $J_{\lambda,\theta}(f)$ is a seminorm in RKHS $\mathcal{H}$ and penalizes the complexity of f using the norm of RKHS $\mathcal{H}_1$ to avoid overfitting f to the training data.

According to [7], the minimizer of the problem in equation [1] has a finite representation taking the form of

$$f(\cdot) = \sum_{j=1}^{m} d_j \phi_j(\cdot) + \sum_{i=1}^{n} c_i k(x_i, \cdot),$$

where $\|P_1 f\|^2_{\mathcal{H}_1} = c^T K c$ for kernel matrix K with $K_{ij} = k(x_i, x_j)$. Therefore, for a given value of the regularization parameter λ, the minimizer f_λ can be estimated by solving the following convex optimization problem:

$$\min_{c \in R^n, d \in R^m} \sum_{i=1}^{n} (y_i - f(x_i))^2 + n\lambda c^T K c, \qquad [2]$$

where $f = [f(x_1), ..., f(x_n)]^T = Td + Kc$ with $T_{ij} = \phi_j(x_i)$. The hyperparameters, λ and θ's, are to be chosen by the generalized cross validation (GCV) [4, 2] method.

Distance Correlation

For a random sample $(X, Y) = \{(X_k, Y_k) : k = 1, ..., n\}$ of n i.i.d random vectors (X, Y) from the joint distribution of random vectors X in R^p and Y in R^q, the Euclidean distance matrices $(a_{ij}) = (|X_i - X_j|_p)$ and $(b_{ij}) = (|Y_i - Y_j|_q)$ are computed. Define the double centering distance matrices

$$A_{ij} = a_{ij} - \overline{a}_{i\cdot} - \overline{a}_{\cdot j} + \overline{a}_{\cdot\cdot}, \quad i, j = 1, ..., n,$$

where

$$\overline{a}_{i\cdot} = \frac{1}{n} \sum_{j=1}^{n} a_{ij}, \quad \overline{a}_{\cdot j} = \frac{1}{n} \sum_{i=1}^{n} a_{ij}, \quad \overline{a}_{\cdot\cdot} = \frac{1}{n^2} \sum_{i,j=1}^{n} a_{ij},$$

similarly for $B_{ij} = b_{ij} - \overline{b}_{i\cdot} - \overline{b}_{\cdot j} + \overline{b}_{\cdot\cdot}, \quad i, j = 1, ..., n.$

Sample Distance Covariance The sample distance covariance $\mathcal{V}_n(X, Y)$ is defined by

$$\mathcal{V}_n^2(X, Y) = \frac{1}{n^2} \sum_{i,j=1}^{n} A_{ij} B_{ij}.$$

Sample Distance Correlation The sample distance correlation $\mathcal{R}_n(X, Y)$ is defined by

$$\mathcal{R}_n^2(X, Y) = \begin{cases} \dfrac{\mathcal{V}_n^2(X, Y)}{\sqrt{\mathcal{V}_n^2(X)\mathcal{V}_n^2(Y)}}, & \mathcal{V}_n^2(X)\mathcal{V}_n^2(Y) > 0; \\ 0, & \mathcal{V}_n^2(X)\mathcal{V}_n^2(Y) = 0, \end{cases}$$

where the sample distance variance is defined by

$$\mathcal{V}_n^2(X) = \mathcal{V}_n^2(X, X) = \frac{1}{n^2} \sum_{i,j=1}^{n} A_{ij}^2.$$

The nonnegativity of $\mathcal{V}_n^2$ and $\mathcal{R}_n^2$ is guaranteed, see [15]. The theory in [15] is based on dissimilarities being actual distances between objects embedded in a Euclidean space, although it is mentioned in the rejoinder to the discussion there that the results hold in certain other metric spaces, see also [12]. The pedigree dissimilarity (d_{ij}) cannot be considered as coming from some metric space, however, since, at least in our study, it does not satisfy the triangle inequality. But we could still treat the pedigree dissimilarity as though it were a distance, since we will see that it can be well approximated by a Euclidean distance obtained by RKE, which we discuss in the next section.

Regularized Kernel Estimation

The Regularized Kernel Estimation (RKE) framework was introduced in [11] as a robust method for estimating dissimilarity measures between objects from noisy, incomplete, inconsistent, and repetitious dissimilarity data. RKE is useful in settings where object classification or clustering is desired but objects do not easily admit description by fixed-length feature vectors, but instead, there is access to a source of noisy and incomplete dissimilarity information between objects. It estimates a symmetric positive semidefinite kernel matrix K which induces a real squared distance admitting of an inner product $d_{ij}^2 = K_{ii} + K_{jj} - 2K_{ij}$.

Assume dissimilarity information is given for a subset Ω of the $\binom{n}{2}$ possible pairs occurring in a training set of n objects, with the dissimilarity between objects i and j denoted as $d_{ij} \in \Omega$. RKE estimates an $n \times n$ symmetric positive semidefinite kernel matrix K of size n such that the fitted squared distance between objects induced by K, $\hat{d}_{ij}^2 = K_{ii} + K_{jj} - 2K_{ij}$, is as close as possible to the square of the observed dissimilarities $d_{ij} \in \Omega$. RKE solves the following optimization problem with semidefinite constraints:

$$\min_{K \succeq 0} \sum_{d_{ij} \in \Omega} w_{ij} |d_{ij}^2 - \hat{d}_{ij}^2| + \lambda_{rke} trace(K). \qquad [3]$$

The parameter $\lambda_{rke} \geq 0$ is a regularization parameter that trades off fit of the dissimilarity data, as given by absolute deviation, and a penalty, trace(K), on the complexity of K. The trace may be seen as a proxy for the rank of K. Thus, RKE is regularized by penalizing high dimensionality of the space spanned by K. RKE requires that Ω satisfies a connectivity constraint that the undirected graph consisting of objects as nodes and edges between them, such that an edge between nodes i and j is included if $d_{ij} \in \Omega$, is connected. Additionally, optional weights w_{ij} may be associated with each $d_{ij} \in \Omega$. A method for choosing the regularization parameter λ_{rke} is required. In this work λ_{rke} is fixed at 1. Unlike in many regularization models, results in the RKE tend to be remarkably insensitive to λ_{rke} over a wide range of values, as can be seen in Figure 1 of [11].

The solution to the RKE problem is a symmetric positive semidefinite matrix K from which an embedding $Z \in R^{n \times r}$ in $r-$dimensional Euclidean space is obtained by decomposing K as $K = ZZ^T$ with $Z = \Gamma_r \Lambda_r^{\frac{1}{2}}$, where the $n \times r$ matrix Γ_r and the $r \times r$ diagonal matrix Λ_r contains the r leading eigenvalues and eigenvectors of K respectively. The ith row of Z is regarded as the vector of "pseudo" coordinates $z(i)$ for subject i. A method for choosing r is required.

The fact that RKE operates on inconsistent dissimilarity data, rather than distances, fits into pedigree studies significantly where the distance correlation depends on Euclidean distances. The pedigree dissimilarity defined above does not satisfy the triangle inequality for general pedigrees, thus is not Euclidean distance. The Euclidean distances induced by the embedding resulting from RKE provides an approximation of the pedigree dissimilarities in our case. This allows us to validate our result of involving the non-metric pedigree dissimilarity in distance correlation by comparing with that obtained by using the embedded Euclidean distances.

Beaver Dam Eye Study

The Beaver Dam Eye Study (BDES) is an ongoing population-based study of age-related ocular disorders. Subjects at baseline, examined between 1988 and 1990, were a group of 4926 people aged 43-86 years. Pedigree information was available for 2356 of the subjects. Although we will only use data from the baseline study for our experiments, five, ten, fifteen and twenty year follow-ups were also obtained. Familial relationships of participants were ascertained and pedigrees of different sizes were constructed for the subset of 1004 subjects who were dead prior to March 2011 with death ages ranging from 46 to 101 years.

Our goal is to use the data to study the association of familial relationships, lifestyle factors, diseases and mortality. The strategy is to first estimate the effects of lifestyle factors and diseases on mortality, i.e. death ages, based on the 1004 subjects using an SS-ANOVA model. The distance correlation is then applied to capture the associations with the estimated effects for a subgroup of 843 people coming from pedigrees containing two or more members. This results in 222 pedigrees in the data set, with sizes ranging from 2 to 23 subjects. Note that it is possible for two persons in one pedigree to be genetically unrelated. They become relatives because of their relationships with other members in the pedigree. The pedigree dissimilarity for such a pair is 1 as previously defined.

It is necessary to notice that the covariates can be continuous, binary and of different magnitude. In addition, the effects of the variables may not be linear in mortality, in which case a large pairwise distance of the covariates values may not result in a large pairwise distance of the death ages. Body mass index (bmi) is such an example in that both underweight and obesity are unhealthy and risky to longevity. In this case, the distance of bmi for two individuals, one with low value and the other with high value, is quite large, however, their death age distance may be small. Thus, instead of the original covariates, the estimated effects are preferred in the calculation of distance correlation because the fitted values are naturally assigned with weights and transformations.

For the above purpose, we fit an SS-ANOVA model of the form

$$\begin{aligned} deathage =& \mu + f_1(baseage) + \beta_{gender} I_{\{gender=F\}} && \Big\} fixed \\ &+ f_2(edu) + f_{12}(baseage : edu) + f_3(bmi) && \Big\} lifestyle \\ &+ \beta_{smoke} I_{\{smoke=no\}} + \beta_{inc} I_{\{inc>20T\}} && \\ &+ \beta_{diabetes} I_{\{diabetes=no\}} + \beta_{cancer} I_{\{cancer=no\}} && \Big\} disease \\ &+ + \beta_{heart} I_{\{heart=no\}} + \beta_{kidney} I_{\{kidney=no\}} && \end{aligned}$$

Table 1: *Variable description in the SS-ANOVA model*

variable	units	description
deathage	years	death age
baseage	years	age at baseline
gender	F/M	gender
edu	years	highest year school/college completed
bmi	kg/m^2	body mass index
smoke	yes/no	history of smoking
inc	yes/no	household personal income $>$ 20T
diabetes	yes/no	history of diabetes
cancer	yes/no	history of cancer
heart	yes/no	history of cardiovascular disease
kidney	yes/no	history of chronic kidney disease

with variables being described in *Table 1* based on 1004 people. The terms in lines one, two to three, and four to five of the above equation are the fixed characteristics, lifestyle factors and disease variables respectively. Functions f_1, f_2 and f_3 are cubic splines and f_{12} uses the tensor product construction. The remaining covariates are unpenalized and modeled as linear terms with $I_{\{\cdot\}}$ as indicator functions. The fitted ef-

fects for *edu* and *bmi* are shown in *Figure 1*. The fitted effects of the linear terms are listed in *Table 2*.

TABLE 2: *Fitted effects of linear terms in the SS-ANOVA model*

gender = F	smoke = no	inc > 20T	
1.141	1.349	0.546	
diabetes = no	cancer = no	heart = no	kidney = no
2.000	0.888	1.131	1.303

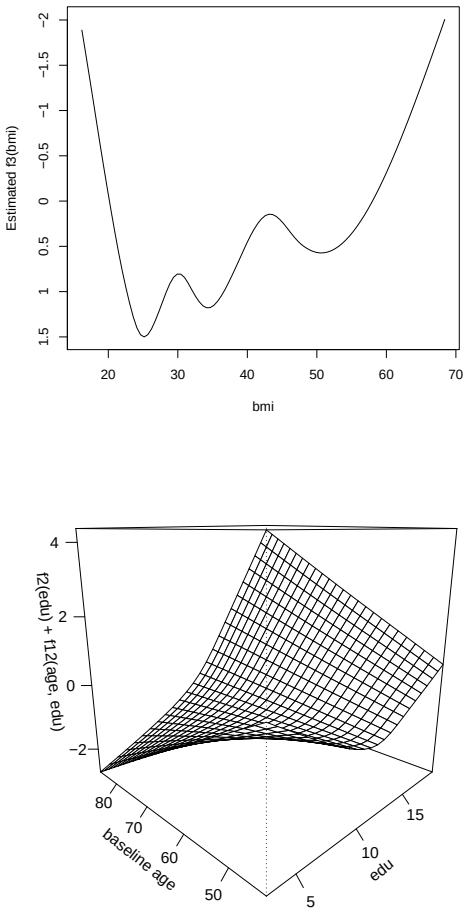

Fig. 1. $f_3(bmi)$ (flipped y-axis) (top), and $f_2(edu) + f_{12}(baseage, edu)$ (bottom) are the fitted effects for bmi and education.

Distance correlation, relying on pairwise distances, is the tool for measuring the association among the lifestyle factors, disease variables, mortality and pedigree. The cohort was restricted to the subgroup of 843 people coming from pedigrees with two or more members. Up to now, the pedigree dissimilarities and Euclidean pairwise death age distances are ready for the calculation of the distance correlation. Lifestyle factors and disease variables get involved as the form of lifestyle factor scores and disease scores. The lifestyle factor score for an individual is the vector of the fitted effects for *smoke*, *bmi*, *edu* and *inc*. Similarly, the disease score is defined to be the vector of the fitted effects for the four disease variables. The Euclidean pairwise distances of the lifestyle factor scores and disease scores are constructed as the input information for lifestyle factors and disease variables in the distance correlation. Permutation tests are implemented to obtain the p-values of the distance correlations. The network in *Figure 2* summarizes the results. Both mortality and lifestyle factors are associated with familial relationships significantly. Heart disease and some cancers are known to run in families. However, the relationship between pedigree and disease variables in this part of the study is not significant at level 0.05. Included here are some pairs of relatives as distant as second cousins, which may be the cause of the weak signal. However, lifestyle factors, disease variables and mortality are closely associated with each other.

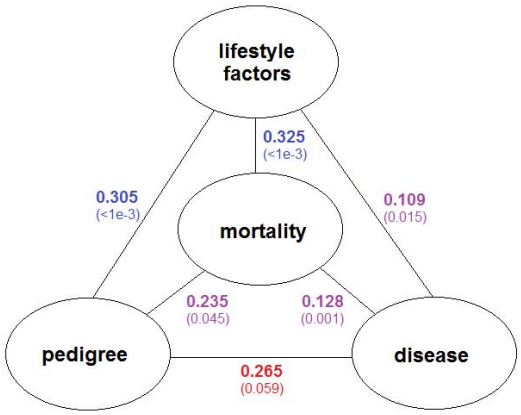

Fig. 2. The network of lifestyle factors, disease variables, mortality and pedigree with distance correlations. The p-values obtained from permutation tests with 1000 replicates are presented in parenthesis. The significance level is distinguished by color: blue for p-value < 0.001, purple for p-value in $(0.001, 0.05)$, red for p-value > 0.05.

The theory of distance correlation is based on Euclidean pairwise distance. However, three of the above six distance correlations involve the non-Euclidean pedigree dissimilarity. The strategy is to validate the results by showing that the pedigree dissimilarity can be well approximated by Euclidean distances through embedding the subjects in Euclidean spaces by RKE. It is possible to establish the embedding effectively in the RKE framework for a moderate sample size of subjects. However, it is too time consuming to solve the RKE semidefinite problem with the full dissimilarity information for 843 people in our case.

Alternatively, we break down the embedding into two steps. The first step only takes care of the within-pedigree dissimilarity. That is, we feed the familywise pedigree dissimilarities to RKE family by family so that it embeds the subjects into Euclidean spaces pedigree by pedigree. The kernel matrices obtained from RKE are then truncated to those leading eigenvalues that account for 95% of the matrix trace to create the "pseudo"-attribute embedding. The resulted familywise coordinates are put together in a way that each pedigree is assigned its own subspace which is orthogonal to the others. This ends up with a coordinate matrix being a horizontal concatenation of the familywise coordinates. The second step is to take into account of the out-pedigree dissimilarity, which requires pedigree specific variables. We assign

one extra dimension to the coordinate matrix for each pedigree. The entries of this extra dimension are the pedigree specific variable for the family members and 0 for the rest of the subjects. This leads to a coordinate matrix being a function of the pedigree specific variables. Thus, the augmented coordinate matrix for the rth member in the pth pedigree takes the form of $(0, ..., 0, v^p, x^p_{r1}, ..., x^p_{rq}, 0, ..., 0)$, where v^p is the pedigree specific variable for the pth pedigree and q is the dimension of the subspace for the pth pedigree. The way to choose the pedigree specific variables is to maximize the Pearson's correlation between the vector form of the double centered pedigree dissimilarities and the vector form of the Euclidean pairwise distances resulting from the above coordinate matrix. The optimal value of the Pearson's correlation is 0.9907. *Figure 3* shows a comparison of the embedded Euclidean pairwise distances and the pedigree dissimilarities for a subset of 100 subjects. It turns out that the non-Euclidean pedigree dissimilarities are well approximated by the embedded Euclidean distances.

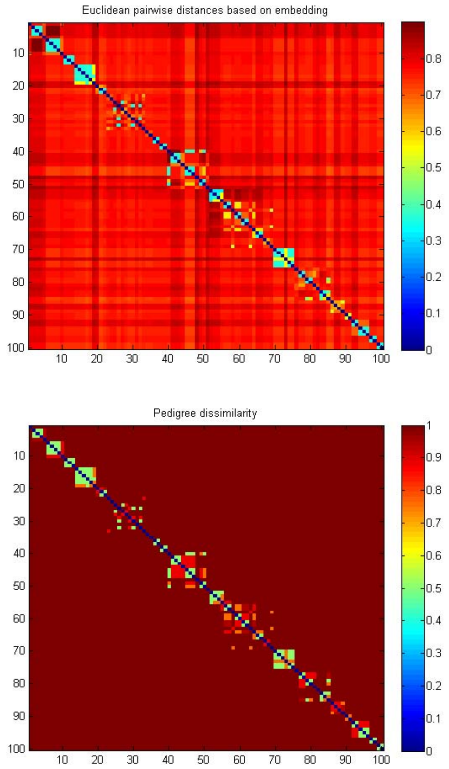

Fig. 3. The comparison of the Euclidean pairwise distances by embedding and the pedigree dissimilarity for a subset of 100 subjects.

We could establish the distance correlations among the lifestyle factors, disease variables, mortality and pedigree based on the embedded Euclidean pairwise distances. The results are presented in *Figure 4* where the p-values are also obtained through permutation tests with 1000 replicates. Both the values of the distance correlation and the p-values are similar to those from the pedigree dissimilarity in *Figure 2*. The embedded results are slightly weaker than the original ones due to the shrinkage of RKE by penalizing high dimensionality of the space spanned by the kernel.

In addition to the study of all relatives, the analysis focusing on the full siblings shows that the signal of running in families gets stronger as the familial relationships become closer. The cohort are further restricted to 462 subjects who had at least one full sibling in the group of 843 people. To simplify the procedure, we change the pedigree dissimilarity for the full sibling pairs, which is shown to be Euclidean. The pedigree dissimilarity is assigned to be 0 for two full siblings and 1 for two unrelated persons. Suppose the subjects who are full siblings to each other are collected to different clusters and there are in total m such clusters. The members in the ith full sibling cluster are assigned the coordinates of length m, $(0, ..., 0, \frac{1}{\sqrt{2}}, 0, ..., 0)$, where the ith element is $\frac{1}{\sqrt{2}}$ and the rest are 0. The corresponding Euclidean pairwise distances are unchanged with the above pedigree dissimilarity being defined for full siblings. The distance correlations and p-values are summarized in *Figure 5* for the full siblings study. The three distance correlation values and related p-values involving familial relationships are strengthened compared to the all relatives study, indicating that the signal of running in families is getting stronger as the subjects are closer. The other three associations are weaker due to the shrinkage of the sample size.

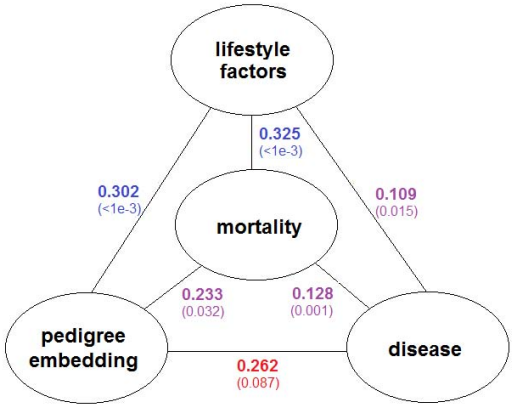

Fig. 4. The network of lifestyle factors, disease variables, mortality and pedigree with distance correlations using the embedded Euclidean distances. The p-values obtained from permutation tests with 1000 replicates are presented in parenthesis.

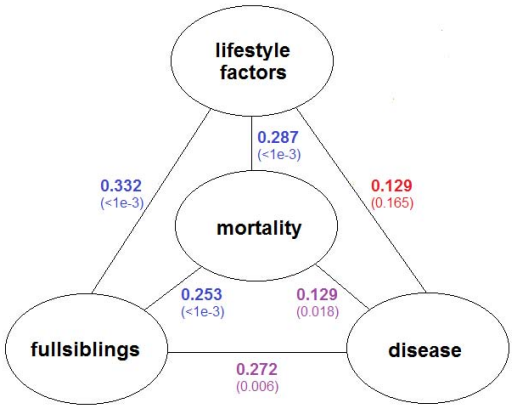

Fig. 5. The distance correlations for full siblings study. The p-values obtained from permutation tests with 1000 replicates are presented in parenthesis.

For the full siblings study, the pairwise distances for mortality could be separated into two groups, group 0 collecting all the pairwise death age distances of full sibling pairs and group 1 for the unrelated pairs. This allows us to compare the difference between the mean of group 1 and the mean of group 0 and construct 95% Bootstrap percentile confidence interval for the test statistic with 10000 replicates. In the case of mortality, the average death age distance of full sibling pairs is 1.571 years less compared to that of two unrelated persons in the cohort. The corresponding 95% Bootstrap percentile confidence interval (CI) for the difference between the mean of group 1 and the mean of group 0 is $(0.919, 2.211)$. We could establish the analysis for the pairwise distances of lifestyle factors and disease variables in the same fashion. The observed test statistics and corresponding confidence intervals are summarized in *Table 3*. All the three mean differences between group 1 and group 0 are positive and the confidence intervals do not overlap 0, which means that the full siblings are significantly closer than unrelated people in terms of death age distances, lifestyle factor scores and disease scores.

TABLE 3: *Bootstrap percentile confidence intervals for the mean differences in the full siblings study*

variable	mortality	lifestyle	disease
group 0 mean	8.091	1.405	1.119
group 1 mean	9.662	1.654	1.229
difference	1.571	0.249	0.110
95% CI	(0.919, 2.211)	(0.167, 0.331)	(0.020, 0.202)

Discussion

The Beaver Dam Eye Study, which began collecting data from a population aged 43 and older in 1988, and continues to the present, provides an ideal opportunity to apply some emerging statistical tools to examine questions regarding relationships between various kinds of information collected at the start of the study and mortality. Since the study contains a large number of people with relatives in the study, this provided an ideal opportunity to examine the correlations between familial relationships, lifestyle factors, disease and mortality. The methodological approach we have proposed here is easily adaptable to other studies for exploring relations between attributes of subjects with multiple clusters of observable attributes, simultaneously with other factors for which pairwise dissimilarities are observed. Some caveats with respect to the mortality data here are worth mentioning. The mortality data is censored at both ends, that is, we do not see cohorts of the oldest subjects who have died before the study began, and, at the other end, we have access to death ages only to those in the study who have died by March 2011. The left censoring is, to some extent accounted for in the presence of baseage in the SS-ANOVA model for deathage–note that there is an interaction term for baseage and education, since it was observed that the oldest cohort in the study clearly had fewer years of formal education than younger members. This study does not use the subjects who would otherwise be included who do not have a recorded death age prior to March 2011. This is, of course a possible source of bias in the conclusions, and we hope to continue following this group as time goes on. Further research concerning residual lifetimes is ongoing, and the results may be able to utilize in addition the partial information contributed by subjects that are known to be alive past a particular time. Other information that is not used here includes attributes collected in the followup examinations. We cannot in this study exclude possible genetic effects behind the lifestyle factors - we only observe that our lifestyle factors significantly run in families, exactly why is beyond the scope of this project. We have shown that pairwise differences in lifestyle factors that run in families correlate well with pairwise differences in death age that also run in families, partially accounting for the familial death age effect. This leads to new questions to be asked about the complex relations between genetics, family structure, lifestyle factors, and other variables. We provide here an overall metholological approach which shows promise to help in answering these questions.

Materials and Methods

The package `gss` in R (`www.r-project.org`) by Chong Gu was used for the SS-ANOVA calculations. The R package `energy` by Gabor Szekely was used for the the dcor calculations. Further information regarding RKE calculations can be found in [11], and Matlab code found in Appendix B of the thesis [10].

ACKNOWLEDGMENTS. GW acknowledges mathematical and editorial help from David Callan. This work was partially supported by National Institutes of Health (NIH) Grant EY09946 and National Science Foundation Grant DMS-0906818 (J.K. and G.W.), NIH Grant EY06594 (R.K., K.L. and B.K.) and by the Research to Prevent Blindness Senior Scientist-Investigator Awards, New York, NY (R.K. and B.K).

1. H. Corrada Bravo, G. Wahba, K. E. Lee, B. E. K. Klein, R. Klein, and S. K. Iyengar. Examining the relative influence of familial, genetic and environmental covariate information in flexible risk models. Proceedings of the National Academy of Sciences, 106:8128–8133, 2009. PMCID: PMC 2677979.
2. P. Craven and G. Wahba. Smoothing noisy data with spline functions: estimating the correct degree of smoothing by the method of generalized cross-validation. Numer. Math., 31:377–403, 1979.
3. F. Gao, G. Wahba, R. Klein, and B. Klein. Smoothing spline ANOVA for multivariate Bernoulli observations, with applications to ophthalmology data, with discussion. J. Amer. Statist. Assoc., 96:127–160, 2001.
4. G. Golub, M. Heath, and G. Wahba. Generalized cross validation as a method for choosing a good ridge parameter. Technometrics, 21:215–224, 1979.
5. C. Gu. Smoothing Spline ANOVA Models. Springer, 2002.
6. E. Khoshgnauz. Learning markov network structure using brownian distance covariance. arXiv:1206.6361v1, 2012.
7. G. Kimeldorf and G. Wahba. Some results on Tchebycheffian spline functions. J. Math. Anal. Applic., 33:82–95, 1971.
8. R. Klein, B. E. K. Klein, K. Linton, and D. DeMets. The Beaver Dam eye study: Visual acuity. Ophthalmology, 98:1310–1315, 1991.
9. R. Li, W. Zhong, and L. Zhu. Feature screening via distance correlation. J. Amer.Statist. Assoc., xx:xx, 2012. To appear, DOI:10.1080/01621459.2012.695654.
10. F. Lu. Regularized Nonparametric Logistic Regression and Kernel Regularization. PhD thesis, Department of Statistics, University of Wisconsin, Madison, 2006. Technical Report 1124.
11. F. Lu, S. Keles, S. Wright, and G. Wahba. A framework for kernel regularization with application to protein clustering. Proceedings of the National Academy of Sciences, 102:12332–12337, 2005. Open Source at www.pnas.org/content/102/35/12332, PMCID: PMC118947.
12. R. Lyons. Distance covariance in metric spaces. arXiv:1106.5758v3, 2011. to appear Ann. Probab.
13. G. Malecot. Les mathematiques de L'Heridite. Masson et Cie, 1948.
14. D. Nott, M. Tran, and R. Kohn. Simultaneous variable selection and component selection for regression density estimation with mixtures of heteroscedastic experts. Electronic Journal of Statistics, 6:1170–1199, 2012.
15. G. Szekely and M. Rizzo. Brownian distance covariance. Ann. Appl. Statist., 3:1236–1265, 2009.
16. G. Szekely, M. Rizzo, and N Bakirov. Measuring and testing independence by correlation of distances. Ann. Statist., 35:2769–2794, 2007.
17. G. Wahba. Spline Models for Observational Data. SIAM, 1990. CBMS-NSF Regional Conference Series in Applied Mathematics, v. 59.
18. G. Wahba, Y. Wang, C. Gu, R. Klein, and B. Klein. Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. Ann. Statist., 23:1865–1895, 1995. Neyman Lecture.
19. Y. Wang. Smoothing Splines: Methods and Applications. Chapoman & Hall/CRC Monographs on Statistics & Applied Probability, 2011.