AMS/IP Studies in Advanced Mathematics
Volume 51, 2012

Dissimilarity Data in Statistical Model Building and Machine Learning

Grace Wahba

Abstract. We explore three papers concerned with two methods for incorporating discrete, noisy, incomplete dissimilarity data into statistical/machine learning models for supervised, semisupervised or unsupervised machine learning. The two methods are RKE (Regularized Kernel Estimation), and RMU (Regularized Manifold Unfolding). Briefly put, the methods use dissimilarity information between objects in a training set to obtain a nonnegative definite matrix of (usually) relatively low rank, which is then used to embed the objects into a (usually) relatively low dimensional Euclidean space, where their coordinates can then be used as attributes in learning models of various types. Some suggestions for further work are noted.

1. Introduction

We are concerned with a particular aspect of statistical model building and related machine learning methods. Our concern is with building models which ultimately relate predictor variables or inputs, also known as attributes, to outcomes, also known as responses or labels. These models are built from observational data ("training sets") of sets of inputs and their related outputs. "Direct" inputs may be highly multivariate (e. g. as in genetic data, or images), they may be numeric-valued inputs with various kinds of structure. Outputs may be univariate or multivariate, they may be classifiers (two-class or multiclass [**19**]), they may be probabilities of class membership (two-class or multiclass [**40**]). They may be real numbers or vectors of real numbers and their correlations [**39**], correlated Bernoulli $(0, 1)$ outputs [**12**], and so forth. Indirect inputs include noisy observations of the values of bounded linear or nonlinear functionals in some reproducing kernel Hilbert space (not considered further here), and noisy, incomplete pairwise dissimilarity information between objects in the training set. The use of this dissimilarity information is our topic here.

The goal is to provide principled methods for using this dissimilarity information in regression, classification and clustering models. In clustering, there are no labels (unsupervised learning), in classification and regression problems all of the

2010 *Mathematics Subject Classification.* 62G08,68T05,62H20,62H30.

training set may have labels (supervised learning) or only part of the training set may have labels (semisupervised learning). In this latter case, the goal is typically to provide labels for the unlabeled data in the training set (transductive learning), or to provide labels both for the unlabeled training set data and for new objects not in the training set (inductive learning).

The basis for our discussion is three papers [**26**] [**7**] [**27**] which have in common the use of an algorithm for embedding discrete, scattered, noisy, incomplete dissimilarity data into a dimension controlled Euclidean space in such a way that the information can be employed as components in any learning algorithm that can admit reproducing kernel Hilbert space (RKHS)-based components.

Two of these papers apply a new algorithm (called RKE for "Regularized Kernel Estimation") to practical examples. The first [**26**] involves the use of BLAST scores to provide a dissimilarity score between pairs of protein sequences, which can be used to visualize and classify proteins. (Section 2). Following this we briefly mention a potential example of the use of the method in image data (Section 2.5). The second practical example is the use of pedigree (relationship) data in a demographic study of an eye condition in conjunction with other, direct information to build a risk model (Section 3) from [**7**].

The embedding method discussed in [**26**] has the potential for dealing robustly with data that is very much non-Euclidean. For example, consider medical images containing tumors of varying lethality. A panel of experts is to be asked to compare images pairwise to give a possibly crude dissimilarity score (on a scale of 1-4, say very close, close, distant, very distant), and this information is to be used in a learning model. If sufficient "landmark" images labeled with levels of the outcome of interest are available, the results can be used in a semisupervised learning model, and could be combined with other subject/image attribute information and/or objective or other distance measurements in a risk model. The coordinates of the embedded object can then be (implicitly) treated just like other covariates in learning models that have an RKHS component, as is done in [**7**].

Section 4 considers a modification of the method in Section 2 from [**27**], (called RMU for "Robust Manifold Unfolding") where the objects are believed to sit in a low-dimensional (generally nonlinear) manifold where the "effective" distance between objects should be measured along the manifold, and only dissimilarity between nearest neighbors is used. This method can be used to "unroll", or flatten the manifold; RMU can also have the effect of enhancing clustering by moving near neighbors closer while relaxing the distance on further objects. This task, generally called manifold learning and other names in the machine learning community, has become the subject of much recent activity, but we will not attempt a literature survey here. The RMU differs from much of the existing literature in its regularization approach, and is conjectured to have some advantages in certain kinds of network data when the ultimate goal is clustering.

The main content of this review is an overview of the three papers cited, while we add commentary and discussion of their interrelationships, tuning, and open questions. The discussion is based on the work of the author and collaborators, with only occasional references to recent research of others based on dissimilarity information, and represents a modest updating of [**43**] and an earlier Technical Report of the same name.

2. Dissimilarity Information and Regularized Kernel Estimation (RKE)

This Section is based on [**26**]. Given a set of N objects, suppose we have obtained a measure of dissimilarity, d_{ij}, for certain object pairs (i, j). We introduce the class of Regularized Kernel Estimates (RKEs), which we define as solutions to optimization problems of the following form:

$$\min_{K \in S_N} \sum_{(i,j)\in\Omega} L\big(w_{ij}, d_{ij}, \hat{d}_{ij}(K)\big) + \lambda J(K), \tag{2.1}$$

where S_N is the convex cone of all real nonnegative definite matrices of dimension N, Ω is the set of pairs for which we utilize dissimilarity information, and L is some reasonable loss function, $\hat{d}_{ij}$ is the dissimilarity induced by K and L is convex in K, J is a convex kernel penalty (regularizing) functional, and λ is a tuning parameter that balances fit to the data and the penalty on K. The w_{ij} are weights that may, if desired, be associated with particular (i, j) pairs. The natural induced dissimilarity, which is a real squared distance admitting of an inner product, is $\hat{d}_{ij} = K(i,i) + K(j,j) - 2K(i,j) = B_{ij} \cdot K$, where $K(i,j)$ is the (i,j) entry of K, B_{ij} is a symmetric matrix of dimension N with all elements 0 except $B_{ij}(i,i) = B_{ij}(j,j) = 1$, $B_{ij}(i,j) = B_{ij}(j,i) = -1$ and the inner (dot) product of two matrices of the same dimensions is defined as: $A \cdot B = \sum_{i,j} A(i,j) \cdot B(i,j) \equiv \text{trace}(A^T B)$. There are essentially no restrictions on the set of pairs other than requiring that they form a connected set. A pair may have repeated observations, which just yield an additional term in (2.1) for each separate observation. If the pair set induces a connected graph, then the minimizer of (2.1) will have no local minima.

Although it is usually natural to require the observed dissimilarity information $\{d_{ij}\}$ to satisfy $d_{ij} \geq 0$ and $d_{ij} = d_{ji}$, the general formulation above does not require these properties to hold. The observed dissimilarity information may be incomplete (with the restriction noted), it may not satisfy the triangle inequality, or it may be noisy. It also may be crude, as for example when it encodes a small number of coded levels such as “very close”, “close”, “distant”, and “very distant”.

2.1. Numerical Methods for RKE.

In this section, we describe a specific formulation of the approach in Section 2, based on a linearly weighted l_1 loss, and use the trace function in the regularization term to promote dimension reduction. The resulting problem is as follows:

$$\min_{K \succeq 0} \sum_{(i,j)\in\Omega} w_{ij} |d_{ij} - B_{ij} \cdot K| + \lambda \, \text{trace}(K). \tag{2.2}$$

Trace was used as a regularizer in [**17**] in a different approach to obtain K, which limited K to a linear combination of prespecified kernels. We show how the present formulation can be posed as a general convex cone optimization problem and also describe a “newbie” formulation in which the known solution to (2.2) for a set of N objects is augmented by the addition of one more object together with its dissimilarity data. A variant of (2.2), in which a quadratic loss function is used in place of the l_1 loss function, is described in the supplementary material published with [**26**].

2.1.1. *General Convex Cone Problem.* We specify here the general convex cone programming problem. This problem, which is central to modern optimization research, involves some unknowns that are vectors in Euclidean space and others that are symmetric matrices. These unknowns are required to satisfy certain equality constraints and are also required to belong to cones of a certain type. The cones have the common feature that they all admit a self-concordant barrier function, which allows them to be solved by interior-point methods that are efficient in both theory and practice.

To describe the cone programming problem, we define some notation. Let $\mathcal{R}^p$ be Euclidean p-space, and let P_p be the nonnegative orthant in $\mathcal{R}^p$, that is, the set of vectors in $\mathcal{R}^p$ whose components are all nonnegative. We let Q_q be the second-order cone of dimension q, which is the set of vectors $x = \big(x(1), \dots, x(q)\big) \in \mathcal{R}^q$ that satisfy the condition $x(1) \geq [\sum_{i=2}^{q} x(i)^2]^{1/2}$. We define S_s to be the cone of symmetric positive definite $s \times s$ matrices of real numbers. Inner products between two vectors are defined in the usual way and we use the dot notation for consistency with the matrix inner product notation.

The general convex cone problem is then:

$$\min_{X_j, x_i, z} \; \sum_{j=1}^{n_s} C_j \cdot X_j + \sum_{i=1}^{n_q} c_i \cdot x_i + g \cdot z \tag{2.3}$$

$$\begin{aligned} \text{s.t.} \sum_{j=1}^{n_s} A_{rj} \cdot X_j + \sum_{i=1}^{n_q} a_{ri} \cdot x_i + g_r \cdot z = b_r, \;\; \forall_r \\ X_j \in S_{s_j} \; \forall_j; \;\; x_i \in Q_{q_i} \; \forall_i; \;\; z \in P_p. \end{aligned} \tag{2.4}$$

Here, C_j, A_{rj} are real symmetric matrices (not necessarily positive semidefinite) of dimension s_j, c_i, $a_{ri} \in \mathcal{R}^{q_i}$; g, $g_r \in \mathcal{R}^p$; $b_r \in \mathcal{R}^1$.

The solution of a general convex cone problem can be obtained numerically using publicly available software such as SDPT3 [**37**] and DSDP5 [**3**].

2.1.2. *RKE with l_1 Loss.* To convert the problem of equation (2.2) into a convex cone programming problem, we may, without loss of generality, let Ω contain m distinct (i, j) pairs, which we index with $r = 1, 2, \dots, m$. Define I_N to be the N-dimensional identity matrix and $e_{m,r}$ to be vector of length $2m$ consisting of all zeros except for the rth element being 1 and $(m + r)$th element being -1. If we denote the rth element of Ω as $\big(i(r), j(r)\big)$, and with some abuse of notation let $i = i(r)$, $j = j(r)$ and $w \in P_{2m}$ with $w(r) = w(r + m) = w_{i(r), j(r)}$, $r = 1, \dots, m$, we can formulate the problem of equation (2.2) as follows:

$$\begin{aligned} &\min\nolimits_{K \succeq 0, u \geq 0} \; w \cdot u + \lambda I_N \cdot K \\ &\text{s.t.} \;\; d_{ij} - B_{ij} \cdot K + e_{m,r} \cdot u = 0, \;\; \forall_r, \\ &\quad K \in S_N, \;\; u \in P_{2m}. \end{aligned} \tag{2.5}$$

2.2. Embedding. In the example in [**26**] there are $N = 280$ (labeled) proteins from four different members of the globin family, and the d_{ij} were from a subset of the $\binom{N}{2}$ pairs, the pairs chosen so that each protein was paired with about 55 of the others. The d_{ij} were obtained from BLAST scores. Figure 1 gives plots of the log eigenvalues of K for λ over several orders of magnitude. It can be seen that there is very little difference between $\lambda = 0.1$ and $\lambda = 10$. It can also be seen that the first three or at most four eigenvectors will contain a very large fraction of the trace of K. This is convenient in this example because it means that the result can

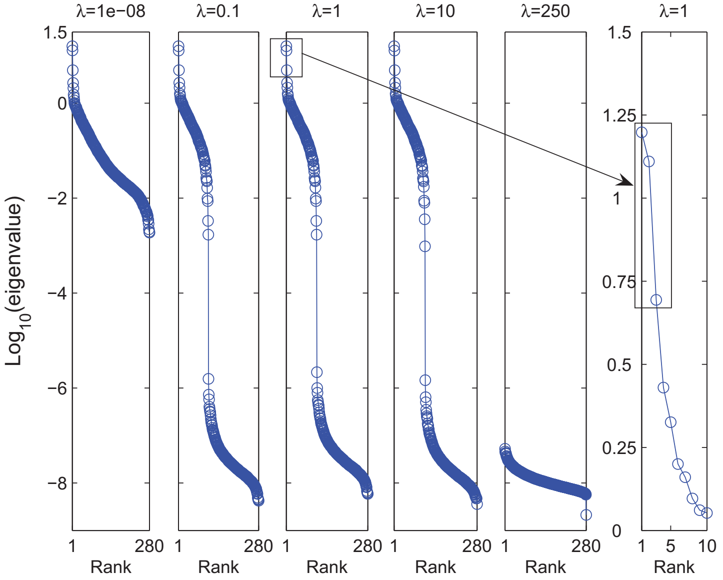


FIGURE 1. Left five panels: log scale eigensequence plots for five values of λ. As λ increases, smaller eigenvalues begin to shrink. Right panel: first ten eigenvalues of the $\lambda = 1$ case displayed on a larger scale.

be visualized readily. For this example λ was taken as 1. Truncating all but the first three eigenvalues in K determines an embedding in Euclidean three-space, which, however, is only determined up to a rotation, since only the distances between objects are relevant. A convenient choice for the embedding goes as follows: Let $Z_{280\times 3} = \Gamma_{280\times 3}\Lambda^{1/2}_{3\times 3}$ where $\Gamma_{280\times 3}$ is the 280×3 matrix of the three leading vectors of K and $\Lambda_{3\times 3}$ the 3×3 diagonal matrix with the three leading eigenvalues in the diagonal. The ith row of Z then gives the three coordinates $z(i) = (z_1(i), z_2(i), z_3(i))$ of the ith object, $i = 1, \dots, 280$. The method automatically centers the collection of the $x(i)$ at 0. Figure 2 gives a plot of the embedding of the 280 proteins. In this example the four colors represent four subfamilies within the globin family; the labels alpha-globin, beta-globin, myoglobin and a heterogenous subfamily are known. It can be seen that these globins could be clustered or if some members of this population were not labeled, they could be identified fairly accurately by any one of several methods.

2.3. 'Newbie' Formulation. Consider the situation in which a solution K_N of (2.2) is known for some set of N objects. We wish to augment the optimal kernel (by one row and column), without changing any of its existing elements, to account for a new object. That is, we wish to find a new "pseudo-optimal" kernel $\tilde{K}_{N+1}$ of the form

$$\tilde{K}_{N+1} = \begin{bmatrix} K_N & b^T \\ b & c \end{bmatrix} \succeq 0, \tag{2.6}$$

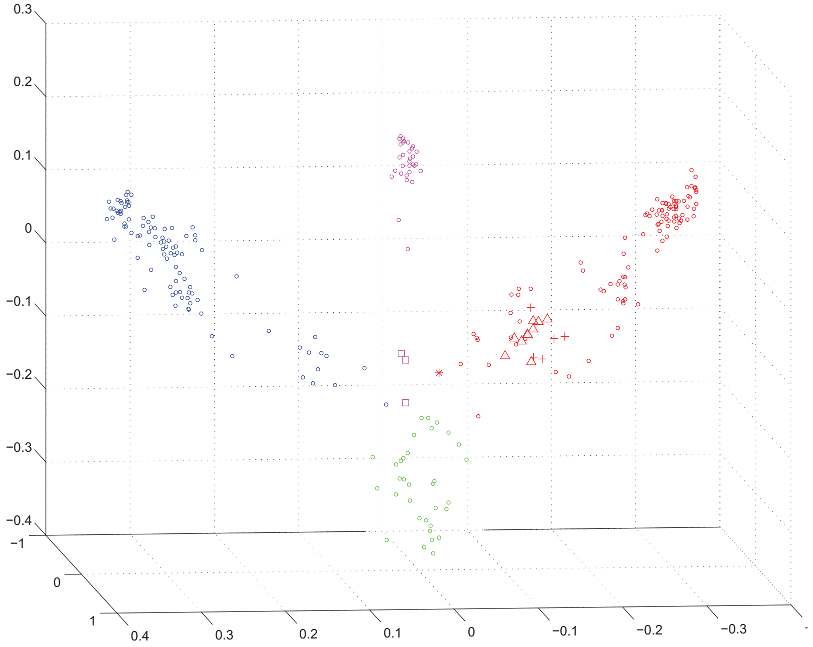


FIGURE 2. *3D representation of the sequence space for 280 proteins from the globin family.* Different subfamilies are encoded with different colors: Red symbols are alpha-globin subfamily, blue symbols are beta-globins, purple symbols represent myoglobin subfamily, and green symbols, scattered in the middle, are a heterogeneous group encompassing proteins from other small subfamilies within the globin family. Here, hemoglobin zeta chains are represented by the symbol +, fish myoglobins are marked by the symbol □, and the diverged alpha-globin HBAM_RANCA is shown by the symbol *. Hemoglobin alpha-D chains, embedded within the alpha-globin cluster, are highlighted using the the symbol △.

(where $b \in \mathcal{R}^N$ and c is a scalar) that solves the following optimization problem:

$$\begin{aligned} &\min_{c\geq 0,b} \textstyle\sum_{i\in\Psi} w_i \, |d_{i,N+1} - B_{i,N+1} \cdot K_{N+1}| \\ &\text{s.t. } \; b \in \text{Range}(K_N), \;\; c - b^T K_N^+ b \geq 0, \end{aligned} \tag{2.7}$$

where K_N^+ is the pseudo-inverse of K_N and Ψ is a subset of $\{1, 2, \ldots, N\}$ of size t. The quantities w_i, $i \in \Psi$ are the weights assigned to the dissimilarity data for the new point. The constraints in this problem are the necessary and sufficient conditions for $\tilde{K}_{N+1}$ to be positive semidefinite.

Suppose that K_N has rank $p < N$ and let $K_N = \Gamma\Lambda\Gamma^T$, where $\Gamma_{N\times p}$ is the orthogonal matrix of non-zero eigenvectors and Λ is the $p \times p$ matrix of positive eigenvalues of K_N. By introducing the variable $\tilde{b}$ and setting $b = \Gamma\Lambda^{1/2}\tilde{b}$, we can ensure that the requirement $b \in \text{Range}(K_N)$ is satisfied. We also introduce the scalar variable $\tilde{c}$, and enforce $c \geq \tilde{c}^2$ by requiring that

$$Z \overset{\text{def}}{=} \begin{bmatrix} 1 & \tilde{c} \\ \tilde{c} & c \end{bmatrix} \in S_2. \tag{2.8}$$

Using these changes of variable, the condition $c - b^T K_N^+ b \geq 0$ is implied by the second order cone condition:

$$x \overset{\text{def}}{=} [\tilde{c} \;\; \tilde{b}^T]^T \in Q_{p+1}.$$

Further we define the $N \times (p+1)$ matrix $\Sigma \overset{\text{def}}{=} [0_N : 2\Gamma\Lambda^{1/2}]$, where 0_N is the zero vector of length N, and we let $\Sigma_{i\cdot}$ be the row vector consisting of the $p+1$ elements of row i of Σ. We use $K_N(i,i)$ to denote the (i,i)th entry of K_N and define the weight vector $w \in P_{2t}$ with components $w(r) = w(t+r) = w_{i(r)},\ r = 1, \ldots, t$. We then replace problem (2.7) by the following equivalent convex cone program:

$$\min_{Z \succeq 0, u \geq 0, x} \quad w \cdot u$$

$$\text{s.t.} \quad \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \cdot Z = 1,$$

$$\begin{bmatrix} 0 & 0.5 \\ 0.5 & 0 \end{bmatrix} \cdot Z - \begin{bmatrix} 1 \\ 0_p \end{bmatrix} \cdot x = 0,$$

$$d_{i,N+1} - K_N(i,i) - \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \cdot Z + \Sigma_{i\cdot} \cdot x + e_{t,r} \cdot u = 0, \ \ \forall_{r=1,2,\ldots,t}, \tag{2.9}$$

$$Z \in S_2, \ \ x \in Q_{p+1}, \ \ u \in P_{2t}, \tag{2.10}$$

where $i = i(r)$ as before. Note that the constraints on Z ensure that it has the form (2.8). The $\hat{d}_{i,N+1}$ are given by $\hat{d}_{i,N+1} = B_{i,N+1} \cdot K_{N+1}$ and are used to insert the newbie in the original embedding coordinate system.

2.3.1. *Embedding of new protein sequences.* We next illustrate how the newbie algorithm worked to visualize unlabeled protein sequences in the coordinate space of training data obtained by RKE. We used the following protein sequences as our test data: (1) Hemoglobin zeta chain (black circle), (2) Hemoglobin theta chain (black star). Figure 3 displays the positions of these two test protein sequences with respect to 280 training sequences. We observe that the black circle clusters nicely with the rest of the hemoglobin zeta chains, whereas the black star, is located closer to beta-globins. Additionally, 17 Leghemoglobins (black triangles) were used as test data and were found to cluster tightly within the heterogeneous globin group. More details, including the scientific implications of the clustering are found in [**26**]. In this example one striking result here is the fact that a simple 3D plot is sufficient for visual identification of the subfamily information. Also, note that the Leghemoglobins cluster tightly together despite the fact that no dissimilarity information between pairs of Legemoglobins was used.

2.4. Classification Overlay: The Multicategory Support Vector Machine. In examining Figure 2 it is clear that if a sufficient number of labels were given, a fairly successful classification algorithm could be built on this data, especially if a "none of the above" category is allowed. The Multicategory Support Vector Machine (MSVM) [**19**] is a good way of doing this. We first very briefly describe the two category SVM and then the MSVM in the general case, where x represents an attribute vector in some space $\mathcal{X}$. Then we return to the application

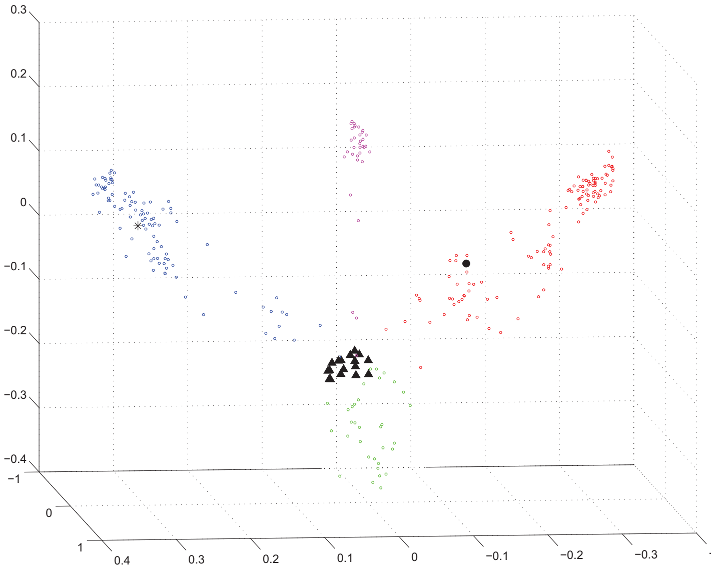


FIGURE 3. *Positioning test globin sequences in the coordinate system of 280 training sequences from the globin family.* The newbie algorithm is used to locate one Hemoglobin zeta chain (black circle), one Hemoglobin theta chain (black star), and seventeen Leghemoglobins (black triangles) into the coordinate system of the training globin sequence data.

of building an MSVM on embedded dissimilarity data. See [**19**] for further information and the properties of the MSVM, and a good place to look for the properties of the SVM as well as the MSVM.

The class labels y_i are either 1 or -1 in the two class SVM setting. Similar to penalized likelihood estimators, the SVM is obtained as the solution to an optimization problem in a reproducing kernel Hilbert space (RKHS). The reader unfamiliar with RKHS may want to skip forward to Section 3.2 and return here later. The SVM methodology seeks a function $f(x) = h(x) + b$ with $h \in \mathcal{H}_K$, an RKHS with reproducing kernel (RK) $K(\cdot,\cdot)$ and b, a constant minimizing

$$\frac{1}{n}\sum_{i=1}^{n}(1 - y_i f(x(i)))_+ + \lambda\|h\|^2_{\mathcal{H}_K}, \tag{2.11}$$

where $(x)_+ = \max(x, 0)$ and $\|h\|^2_{\mathcal{H}_K}$ denotes the square norm of h in $\mathcal{H}_K$. According to [**15**], the minimizer h is of the form $h(x) = \sum_{i=1}^{n} c_i K(x, x(i))$ for some $c = (c_1, \cdots, c_n)$. $(1 - \tau)_+$ is known as the hinge function. If $\mathcal{H}_K$ is the d-dimensional space of homogeneous linear functions $h(x) = w \cdot x$ with $\|h\|^2_{\mathcal{H}_K} = \|w\|^2$, then (2.11) reduces to the linear SVM. $\lambda = \lambda_{SVM}$ is a tuning parameter. The classification rule $\phi(x)$ induced by $f(x)$ is $\phi(x) = sign(f(x))$.

For ease of exposition, assume that all misclassification costs are equal and there is no sampling bias in the training data set, and consider the k-category classification problem. (For the general case see [**44**] [**45**].) In the MSVM, the observation y_i is coded into a k dimensional vector with 1 in the j position if object i is in class j and $\frac{-1}{(k-1)}$ in the other positions. For instance, if example i falls into class 1, $y_i = (1, \frac{-1}{(k-1)}, \cdots, \frac{-1}{(k-1))})$. Thus the components of each y_i are required to sum to zero. Accordingly, we define a k-tuple of separating functions $f(x) = (f^1(x), \cdots, f^k(x))$ with the sum-to-zero constraint, $\sum_{j=1}^k f^j(x) = 0$ for any x in Euclidean d space. Each component $f^j(x)$ can be expressed as $h^j(x) + b^j$ with $h^j \in \mathcal{H}_{Kj}$. For expository purposes we assume they are in the same RKHS denoted by $\mathcal{H}_K$.

The MSVM is defined as the vector of functions $f_\lambda = (f_\lambda^1, \cdots, f_\lambda^k)$, with each h^k in $\mathcal{H}_K$ satisfying the sum-to-zero constraint, which minimizes

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{r \neq cat(i)}(f^r(x(i)) + \frac{1}{k-1})_+ + \lambda\sum_{j=1}^{k}\|h^j\|_{\mathcal{H}_K}^2, \tag{2.12}$$

where $cat(i)$ is the category of y_i. It is not hard to show that the $k = 2$ case reduces to the usual two category SVM. The target for the MSVM is $f_\lambda(x) = (f_\lambda^1(x), \cdots, f_\lambda^k(x))$ with $f^j(x) = 1$ if $p_j(x)$, the probability that an object with attribute vector x is in category j, is bigger than the other $p_l(x)$ and $f^j(x) = -\frac{1}{k-1}$ otherwise. Simulations in [**19**] and elsewhere demonstrate how well this target can be hit. Each $f^r(x)$ has a representation $\sum_{i=1}^n c_{ir}K(x, x(i)) + b^r$, and class r is assigned if $f^r(x) > f^j(x), j \neq r$.

We return to application to embedded dissimilarity data (z's). If we let the reproducing kernel for $\mathcal{H}_K$ be $K_{\lambda_{RKE}}$, the embedding kernel, we have (from Section 2.2) that $K(z, z(i)) = K_{\lambda_{RKE}}(z, z(i)) = z \cdot z(i)$, so that the $f^r(z)$ are hyperplanes in the embedding coordinate system. Note that classification based on hyperplanes will be invariant under rotations of the coordinate system, as it should be. For the embedded data in Figure 2 it is likely that hyperplanes would provide a reasonable classifier. In general, hyperplanes may not provide a reasonable classifier, and in that case it would be desirable to build a nonparametric MSVM on the embedded coordinates. To insure that the resulting classification does not depend on the orientation of the embedding system, it is sufficient to choose an RK based on a radial basis function (RBF), in which case $K(z, z(i)) = r(\|z - z(i)\|)$, for an appropriate r. See Section 3 and the Appendix for more on RBF's. Note that if we begin with dissimilarity data for labeled, or partly labeled data, embed the observations in Euclidean d-space and then apply the MSVM to make an automatic classifier, there are two tuning parameters, λ_{RKE}, and λ_{SVM}, for the penalty functional in the RKHS determined by the RBF. See Section 3 for more on tuning. Recently, [**34**] give a novel take on clustering with the distance matrix corresponding to K_{RKE}.

2.5. Image Similarity and Dissimilarity. Consider the problem of comparing shapes of images, for example the shape of some image of a region of the brain (after registration) with the goal of, say, classifying normal subjects and those with some condition. The κ index [**48**] provides a measure of similarity. With some abuse of notation let κ be the matrix with ijth entry $\kappa_{ij} = \frac{2S_i \cap S_j}{S_i + S_j}$ where the S_i and S_j are the areas or volumes of region i and region j respectively and $S_i \cap S_j$ is the volume of their intersection, $\kappa_{ij} \in [0, 1]$. Let $K = \kappa\kappa$, and let K_d be obtained by setting all but d eigenvalues of K to 0. Then K_d may be used as in Section 2.2 to embed the n images into a d-dimensional Euclidean space. [**18**]. Nonnegative definite similarity matrices have in recent years been frequently used in statistical model building for images; for a recent example see [**31**]. A recent discussion of the use of noisy (indefinite) similarity matrices and methods for transforming them to kernel matrices can be found in [**5**]. If there is missing data, but the available data is accurate, matrix completion procedures [**4**] could be used to fill in the missing values. Alternatively suppose that the set of κ_{ij} are noisy and incomplete, one could set d_{ij} to $1 - \kappa_{ij}$ and use RKE to get a Kernel matrix. The RKE method is computer intensive but deals with both noisy and incomplete data simultaneously in a transparent way, and appears to be quite robust to missing data - in theory, at least, only requiring that the set of pairs with observations be a connected set. It makes it easy to choose the eigenvalue cutoff point either by visual inspection, by a requirement to keep a certain fraction of the trace of K, or, by tuning methods, particularly ones that respect the ultimate use of the embedded data in further analysis via regression methods. However, comparative experiments, either theoretical or on observational quality data sets, are yet to be done.

A different take on dissimilarities between functonal brain network images based on FDG-PET scans is found in [**18**].

3. Incorporating Dissimilarity Data into an SS-ANOVA Model

This section is primarily based on [**7**]. We begin with Smoothing Spline ANOVA (SS-ANOVA) models [**46**] [**13**] [**21**] [**38**] which are are a well known approach to penalized likelihood regression given heterogenous attribute variables, with the ability to model their various interactions. In [**12**] an SS-ANOVA model was built to estimate the probability that a member of a study cohort in the Beaver Dam Eye Study (BDES) has a particular eye condition (retinal pigmentary abnormalities, a precursor to age-related macular degeneration, AMD) as a function of several risk factors. In the BDES a large fraction of people in the study had relatives in the study, and it is known that AMD tends to run in families. The pedigree (familial relationship) structure has been carefully documented in BDES, and this provided an incomparable opportunity to use a measure of genetic distance to assign pairwise distances between people in pedigrees, and to develop and demonstrate an approach to incorporating this information into an SS-ANOVA model with the use of the RKE of [**26**]. Recently genetic markers have been found that are associated with a risk of AMD. See [**14**] [**29**] and other references cited in [**7**]. A set of two genetic markers relevant to AMD were also available and are easily incorporated into an SS-ANOVA model, so that the relative influence of the original covariates, the genetic markers and the pedigree information could be assessed. The embedding structure of the pedigree data is quite different than what was seen in [**26**], but the method of incorporation of dissimilarity data here is applicable to a wide variety of circumstances, while at the same time raising issues for further work.

3.1. Penalized Log Likelihood for Bernoulli Responses. For the protein classification problem of [**26**], the SVM is ideal – it returns an estimated class label accurately when classes are easily separable, and concentrates the calculational work on identifying the separation boundary - it does not estimate a probability of class membership and it is not sensitive to outliers. If classes are easily separable, as in [**26**], the log odds ratio will be $\pm\infty$ leading to numerical instabilities in estimating the log odds ratio. In various kinds of medical problems, it is desired to estimate the probability of class membership, such as some phenotype, when attribute vectors and relationships influence response, but by no means guarantee it. We will discuss the Bernoulli case where there are two classes, and it is desired to estimate $p(x) = Prob(y|x = 1)$ using a penalized log likelihood model. We estimate instead the log odds ratio (a.k.a. logit) $f(x) = \log \frac{p(x)}{1-p(x)}$ and recover $p(x)$ from $p(x) = e^{f(x)}/(1 + e^{f(x)})$. Given $y_i, x(i), i = 1, 2, \cdots, n,\ y \in \{0, 1\}$, $x = (x_1, x_2, \cdots, x_d)$ the negative log likelihood in the Bernoulli case is given by

$$\mathcal{L}(y, f) = \sum_{i=1}^{n} -y_i f(x(i)) + \log(1 + e^{f(x(i))}) \tag{3.1}$$

and the penalized log likelihood estimate of f is obtained by finding f in some prescribed function space to minimize

$$I(f) = \mathcal{L}(y, f) + \lambda J(f) \tag{3.2}$$

where $J(f)$ is a penalty functional on f and $\lambda = \lambda_{MAIN}$ is a (main) tuning parameter which balances fit to the data and complexity/wiggliness of f, or signal-to-noise ratio, in the Bernoulli case. The multicategory penalized likelihood case is discussed in [**41**]. In the two category case, if the data is coded as ±1 (as opposed to $\{0, 1\}$), then the negative log likelihood becomes $\log(1 + e^{-yf})$ and may be directly compared to the hinge function $(1 - yf)_+$ of Equation (2.11). See [**41**]. The negative log likelihood and the hinge function have quite different properties. Recently [**23**] have proposed a family of so-called large margin classifiers called large-margin unified machines (LUMs), which cover a broad range of classifiers including both the SVM and penalized likelihood, to allow "interpolation" between their properties.

3.2. Reproducing Kernel Hilbert Spaces (RKHS).

It will be seen that RKHS methods provide a convenient and natural approach to include dissimilarity data in regression and classification models.

We briefly review some facts concerning RKHS. Let $K(s, t)$ be a positive definite function on $\mathcal{T} \otimes \mathcal{T}$. This means for any $k,\ t_1, \cdots, t_k \in \mathcal{T},\ a_1, \cdots, a_k$ $\sum_{r,s=1}^{k} a_r a_s K(t_r, t_s) \geq 0$. The Moore-Aronszajn Theorem [**1**] tells us that to every positive definite function $K(\cdot, \cdot)$ there corresponds a unique RKHS $\mathcal{H}_K$ and vice versa.

$K(\cdot, t*) \in \mathcal{H}_K\ \forall t* \in \mathcal{T}$,

$\sum_r c_r K(\cdot, t_r) \in \mathcal{H}_K$,

$f \in \mathcal{H}_K \Rightarrow < f(\cdot), K(\cdot, t*) >= f(t*)\ \forall t* \in \mathcal{T}$,

$\| \sum c_r K(\cdot, t_r)\|^2_{\mathcal{H}_K} = \sum_{rs} c_r c_s K(t_r, t_s)$.

The closure of the span of the $K(\cdot, t_r),\ t_r \in \mathcal{T}$ in the above norm completes $\mathcal{H}_K$. It is important to note that $\mathcal{T}$ can be any domain whatsoever on which it is possible to define a positive definite function. In particular, tensor sums and products of positive definite functions are also positive definite. It is also good to know that positive definite functions (a.k.a. Reproducing Kernels) are available that only depend on the Euclidean distance between the two arguments.

3.3. Smoothing Spline ANOVA (SS-ANOVA) Models. SS-ANOVA models [**46**] [**13**] [**21**] [**38**]. are based on ANOVA decompositions of functions of several variables. We describe the functional ANOVA decomposition is some generality. Let

$$x = (x_1, \cdots, x_d) \in \mathcal{X} \equiv \mathcal{X}^{(1)} \otimes \cdots \otimes \mathcal{X}^{(d)} \tag{3.3}$$

and

$$f(x) = f(x_1, \cdots, x_d),\ x_\alpha \in \mathcal{X}^{(\alpha)} \tag{3.4}$$

Let $d\mu_\alpha$ be a probability measure on $\mathcal{X}^{(\alpha)}$ and define the averaging operator $\mathcal{E}_\alpha$ on $\mathcal{X}$ by

$$(\mathcal{E}_\alpha f)(x) = \int_{\mathcal{X}^{(\alpha)}} f(x_1, \cdots, x_d) d\mu_\alpha(x_\alpha). \tag{3.5}$$

The averaging operators $\mathcal{E}_\alpha$ give a (unique) ANOVA decomposition of f:

$$f(x_1, \cdots, x_d) = \mu + \sum_\alpha f_\alpha(x_\alpha) + \sum_{\alpha\beta} f_{\alpha\beta}(x_\alpha, x_\beta) + \cdots \tag{3.6}$$

where

$$\begin{array}{rcl} \mu & = & \prod_\alpha \mathcal{E}_\alpha f = \int \cdots \int f(x_1, \cdots, x_d) d\mu_1(x_1) \cdots d\mu_d(x_d) \\ f_\alpha & = & (I - \mathcal{E}_\alpha) \prod_{\beta \neq \alpha} \mathcal{E}_\beta f \\ f_{\alpha\beta} & = & (I - \mathcal{E}_\alpha)(I - \mathcal{E}_\beta) \prod_{\gamma \neq \alpha,\beta} \mathcal{E}_\gamma f \\ \vdots & & \vdots\ \mathcal{E}_\alpha f_\alpha = 0,\ \mathcal{E}_\alpha \mathcal{E}_\beta f_{\alpha\beta} = 0, etc. \end{array}$$

The series in (3.6) is truncated at some point. Terms satisfy ANOVA-like side conditions (identifiable). An SS-ANOVA representation with weights on kernels looks like

$$f(\cdot) = \sum_{j=1}^m d_j \phi_j(\cdot) + \sum_{i=1}^n c_i K_\theta(\cdot, x(i)) \tag{3.7}$$

where the ϕ_j are a small set of unpenalized components (parametric part), and

$$K_\theta(\cdot,\cdot) = \sum_{\alpha=1}^d \theta_\alpha K_\alpha(\cdot,\cdot), + \sum_{\alpha \le \beta} \theta_{\alpha\beta} K_{\alpha\beta}(\cdot,\cdot) + \cdots \tag{3.8}$$

The kernels depending only on x_α satisfy $\mathcal{E}_\alpha K_\alpha(\cdot, x_\alpha) = 0$ where the averaging operator acts on $(\cdot)$ and the higher order kernels are usually tensor products of such kernels, which will then satisfy the ANOVA side conditions.. Since $\|f\|^2_{\mathcal{H}_{\theta K}} = \theta^{-1}\|f\|^2_{\mathcal{H}_K}$, the SS-ANOVA penalty functional has the form:

$$J(f) = \sum_{i,j=1}^n c_i c_j \left[\sum_{\alpha=1}^d \theta_\alpha^{-1} K_\alpha(x(i), x(j)) + \sum_{\alpha \le \beta} \theta_{\alpha\beta}^{-1} K_{\alpha\beta}(x(i), x(j)) + \cdots \right] \tag{3.9}$$

where it is understood that only the components of $x(i)$ indicated by the subscripts on the kernel actually enter. The θs are tuning parameters along with λ and with an identifiability constraint. For each trial set of tuning parameters, the c_i are to

code	units	description
`horm`	yes/no	current usage of hormone replacement therapy
`hist`	yes/no	history of heavy drinking
`bmi`	kg/m^2	body mass index
`age`	years	age at baseline
`sysbp`	$mmmHg$	systolic blood pressure
`chol`	mg/dL	serum cholesterol
`smoke`	yes/no	history of smoking

Table 1. E/C covariates for BDES pigmentary abnormalities SS-ANOVA model

be fitted. Calling the fitted result $f_{\lambda\theta}$, the fitted $f_{\lambda\theta}$ are evaluated for the best set of tuning parameters via a tuning criterion. When data is copious, it can be separated into train, tune and test groups and tuned on the tuning set, but when the sample size is moderate an internal tuning criterion is appropriate. The Generalized Approximate Cross Validation (GACV) [**47**] for Bernoulli data models with RKHS penalties is used in [**7**].

3.4. SS-ANOVA Model in the Beaver Dam Eye Study. The Beaver Dam Eye Study (BDES) is an ongoing population-based study of age related ocular disorders, begun in 1988. An SS-ANOVA model for association of a number of environmental/clinical (E/C) variables based on 2585 women with complete E/C data appears in [**21**]. 684 women have at least one relative also in the study with complete E/C data, and this provides an opportunity to make use of this relationship (pedigree) data. The predictor variables of present interest are in Table 1:

The fitted E/C model that is used in the study under discussion is

$$\begin{aligned} f(t) = \mu \quad &+ \quad f_1(\mathtt{sys}) + f_2(\mathtt{chol}) + f_{12}(\mathtt{sys},\mathtt{chol}) \\ &+ \quad d_{\mathtt{age}} \cdot \mathtt{age} + d_{\mathtt{bmi}} \cdot \mathtt{bmi} \\ &+ \quad d_{\mathtt{horm}} \cdot I_1(\mathtt{horm}) + d_{\mathtt{drin}} \cdot I_2(\mathtt{drin}) + d_{\mathtt{smoke}} \cdot I_3(\mathtt{smoke}) \end{aligned} \tag{3.10}$$

This is the same model that was fitted in [**21**] with the exception that `smoke` was not included there. In this model, f_1, f_2 and f_{12} are splines.

3.5. Modeling E/C, Genetic and Pedigree Data in an SS-ANOVA Model. In the study under discussion, logit has the representation

$$\begin{aligned} f(t) = \mu \quad &+ \quad d_{\mathrm{SNP1,1}} \cdot I(X_1 = \mathtt{12}) + d_{\mathrm{SNP1,2}} \cdot I(X_1 = \mathtt{22}) \\ &+ \quad d_{\mathrm{SNP2,1}} \cdot I(X_2 = \mathtt{12}) d_{\mathrm{SNP2,2}} \cdot I(X_2 = \mathtt{22}) \\ &+ \quad f_1(\mathrm{sysbp}) + f_2(\mathrm{chol}) + f_{12}(\mathrm{sysbp},\mathrm{chol}) \\ &+ \quad d_{\mathrm{age}} \cdot \mathrm{age} + d_{\mathrm{bmi}} \cdot \mathrm{bmi} \\ &+ \quad d_{\mathrm{horm}} \cdot I_1(\mathrm{horm}) + d_{\mathrm{drin}} \cdot I_2(\mathrm{drin}) + d_{\mathrm{smoke}} \cdot I_3(\mathrm{smoke}) \\ &+ \quad f_{ped}(z). \end{aligned} \tag{3.11}$$

The first two lines in (3.11) are Genetic (SNP) data. There are two SNPS each with three levels, (1,1), (1,2), (2,2). They are markers for ARMS2 (rs10490924) and CFH1, two genetic locations that are known to be related to AMD. See ([**14**]

[**29**] and references there. The next three lines are E/C variables, and the last line contains pedigree/relationship data to be explained shortly. Figure 4(a) gives an example of a pedigree from BDES and 4(b) gives the relationship graph for five members of this pedigree. In Figure 4(a) it can be seen that persons 35 and 26 are siblings, and are assigned a dissimilarity of 1 in Figure 4(b), persons 8 and 10 are aunt and niece and are assigned a dissimilarity of 2, persons 35 and 40 are first cousins and are assigned a dissimilarity of 3, and persons 26 and 40 are also first cousins and assigned a dissimilarity of 3. These numbers are monotone functions of Malecot's kinship (coancestry) coefficient ψ [**30**] [**24**], a measure of genetic similarity of two individuals with a common ancestor - the score is $\log_2(2\psi)$. Relationship scores go up to 5 in this study. Pairs with with no known common ancestor are indicated with dashed edges in 4(b), and these edges will be coded with a large, arbitrary constant, L. Since there are many disconnected pedigrees, in order to have a connected graph for input to the RKE, a large number of unrelated pairs are coded as L. An embedding matrix R for the subjects is obtained by solving the same convex convex cone optimization problem as in Section 2.1:

$$\min_{R\succeq 0} \sum_{(i,j)\in\Omega} |d_{ij} - B_{ij}(R)| + \lambda_{RKE} trace(R). \tag{3.12}$$

$R_{\lambda_{RKE}}(i,j)$ then gives a (unique up to rotation) embedding $z(i), i = 1, \cdots, n$ of the subjects, as in Section 2.2. Tuning of λ_{RKE} will be described later. For each trial value of λ_{RKE}, 95% of the trace is retained while small eigenvalues are deleted. Figure 5 shows the embedding of the five persons in the relationship graph of Figure 4(b). These five persons can be embedded in three dimensions but not all five person subgraphs have this property. These embeddings will go into $f_{ped}(z)$ in the extended SS-ANOVA model of (3.11). The horizontal axis (z_3) of this plot is order of magnitudes larger than the other two axes. The unrelated edges in the relationship graph occur along this dimension, while the other two dimensions encode the relationship distance. Unlike in Section 2.4 it is fairly clear that we do not want to build a linear model on the embedded points $z(i)$. Since only the distances $\|z(i) - z(j)\|$ are relevant, we can "kernelize" a function defined on the embedding space using any RK that only depends on the distances, that is, any radial basis function (RBF). The Matern family of RBF's is a convenient two parameter family with m, an order parameter, and α, a scale parameter (not to be confused with variable subscripts α); m and α are tuning parameters to be chosen. In the present work a Matern kernel of order $m = 3$ was chosen. It is

$$K_z(z^*, z') = r_3(\|z^* - z'\|) \tag{3.13}$$

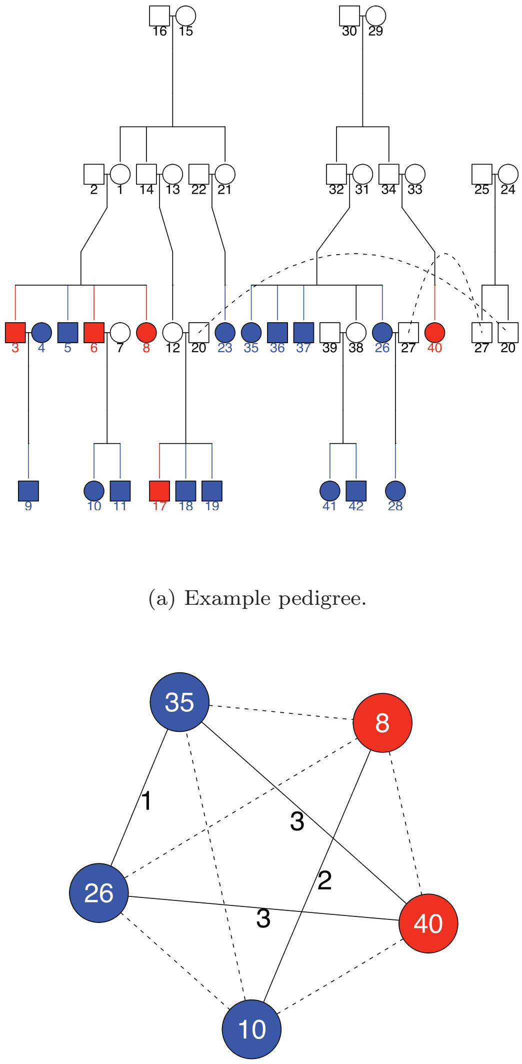
where

$$r_3(\tau) = \frac{1}{\alpha^7}\exp\{-\alpha\tau\}[15 + 15\alpha\tau + 6\alpha^2\tau^2 + \alpha^3\tau^3]. \tag{3.14}$$

If a newbie is not in a pedigree, then $K_z(z_{newbie}, z(j))$ will be very small or 0 for all j. Equation (3.8) becomes

$$K_\theta(\cdot,\cdot) = \sum_{\alpha=1}^{d} \theta_\alpha K_\alpha(\cdot,\cdot), + \sum_{\alpha\le\beta} \theta_{\alpha\beta} K_{\alpha\beta}(\cdot,\cdot) + \cdots + \theta_z K_z(\cdot,\cdot). \tag{3.15}$$

and $K_\theta(\cdot, x(j))$ of Equation(3.7) becomes $K_\theta(\cdot, x(j) : z(j))$, that is, $K_\theta(x, x(j)))$ becomes $K_\theta(x : z, x(j) : z(j)$.



(a) Example pedigree.

(b) Relationship graph. Edge labels are dissimilarities defined by the kinship coefficient (sibling/parental=1, avuncular=2, first cousins=3,...). Dotted edges indicate unrelated pairs.

FIGURE 4. An example pedigree from the BDES and a relationship graph for five subjects. Colored nodes are subjects assessed for retinal pigmentary abnormalities (red encodes a positive result). Circles are females and rectangles are males.

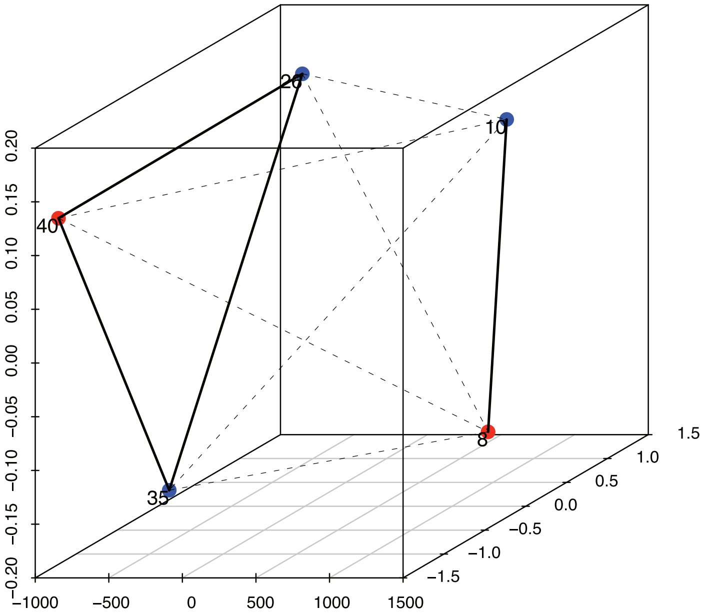


FIGURE 5. Embedding of relationship graph in Figure 4 by RKE. The horizontal axis of this plot is order of magnitudes larger than the other two axes. The unrelated edges in the relationship graph occur along this dimension, while the other two dimensions encode the relationship distance.

3.6. Tuning. We have the following tuning parameters:

- λ_{MAIN} of of Equation (3.2) which controls the tradeoff between the goodness of fit and the size of the penalty functional in a penalized likelihood model. This governs the signal-to-noise ratio given the other parameters in J
- $\theta_\alpha, \theta_{\alpha\beta} \ldots$ and θ_z of Equation (3.15) subject to a single side condition so that they are identifiable in the presence of λ_{MAIN}
- λ_{RKE} of Equation (3.12) used to get the positive definite function providing the embedding of the dissimilarity information.
- Parameter(s) in the RBF $r(z)$ that will be used to build the regression on the embedding coordinates. If a member of the Matern family is used, those parameters are the scale α and the order m

The embedding tends to be fairly insensitive to λ_{RKE} over several orders of magnitude, so generally only a small number of values of $\log \lambda_{RKE}$ need to be considered, similarly if a member of the Matern family is to be used, only a small number of order parameters m need to be tried. The results are invariably most sensitive to λ_{MAIN}, and can be very sensitive to scale factors in kernels, such as the Matern parameter α, and so these need to be chosen carefully. In this work, the GACV tuning method for Bernoulli data with RKHS penalty [**47**] [**22**] was used to choose these parameters. The GACV is a prediction oriented method targeted to minimize the Kullback-Liebler distance between the fit and the "true" but unknown model, derived from a leaving-out-one argument, but much easier to compute.

3.7. Qualitative Results. An important goal of the study was to explore the relative contribution of each source of data. Since there are three sources of information: (S=SNPS, P=Pedigrees,C= Environmental/Clinical) there were seven models to consider:

- S = SNPS (genetic data) only
- C = Environmental/Clinical (E/C) data only
- S + C
- P = Pedigrees only
- S + P
- C + P
- S + C + P

Figure 7 gives the ROC Curves for the S + C + P model and the three models with two sources of information. Figure 6 plots the area under the ROC curve (AUC) for all seven models.

We can see the relative importance of clinical/environmental variables, certain genetic information, and pedigree information in modeling risk of pigmentary abnormalities in the BDES. The approach has promise for many other applications where relationship or dissimilarity information is available along with covariate information. Recently [**10**] [**8**] [**9**] have approached the same problem of including pedigree relationship information along with other covariate information in breeding data sets in a model with Gaussian outcomes which has many similarities and some differences with the present approach. Their approach directly uses a measure of genetic distance which is actually a Euclidean distance and thus there is no step analogous to RKE.

4. Dissimilarity Data and Regularized Manifold Unrolling

Within the last few years there has been much interest in data that is believed to lie in a low dimensional possibly nonlinear manifold in a high dimensional space. Figure 8 gives a picture of the (in)famous Swiss roll, which is a highly stylized depiction of this situation and quoted by many authors. The import of the figure is that determining Euclidean distances or dissimilarities between the data points would make points that are far apart when measured along the manifold (or a corresponding graph) appear wrongly close if measured in Euclidean coordinates. Rather, distances or dissimilarities should be measured along the manifold. Figure 8 was constructed by "rolling up" the two dimensional data in Figure 9. So, simply put, given the data in Figure 8 contaminated by noise, can you recover (unroll, flatten) to get an estimate of the unrolled data in Figure 9?

See the references in [**27**] for various approaches [**36, 33, 2, 11, 16**] to unrolling the Swiss roll, and many real applications. More recently, [**49**] discuss manifold unrolling and give further references. In [**27**] we show that small modifications to the RKE of Section 2 can be used to efficiently "unroll" the Swiss roll. Let Ω_k be the set of pairs of points that are neighbors according to some criterion indexed by k, for example, k-nearest neighbors, although other criteria can be used. The goal is to embed the data in such a way that pairs that are not in Ω_k are as far apart as possible while the end product embedding respects the dissimilarity information

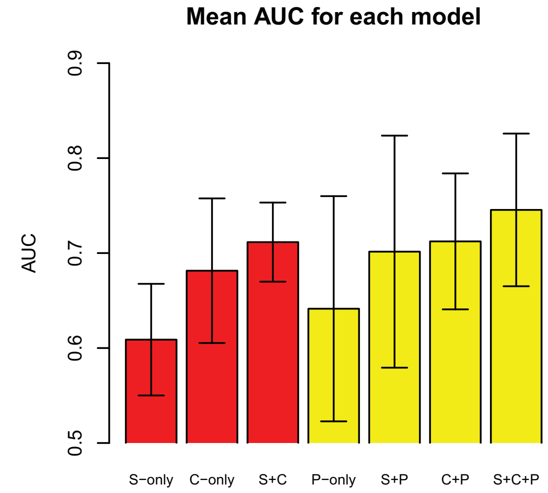


FIGURE 6. AUC Comparison of Seven Models

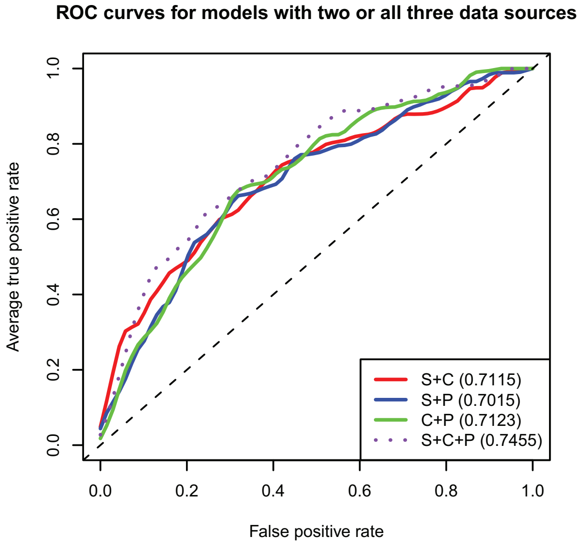


FIGURE 7. ROC Curves for Two and Three Source Models

for pairs in Ω_k. Several equivalent formulations of the solution to this problem are given; here we only describe one. The optimization problem is:

$$\min_{R\succeq 0} \sum_{(i,j)\in\Omega_k} w_{ij}|d_{ij} - B_{ij}\cdot R| - 2\lambda\,\text{trace}(R). \tag{4.1}$$

subject to $E \cdot R = 0$, where E is the $N \times N$ matrix with all entries as 1. Given $R = R_{\lambda_{RMU}}$ and the neighbor index k the embedding proceeds as in the previous sections, and a newbie algorithm proceeds similarly, except that the newbie is embedded using only nearest neighbors according to the criterion determined by k. Given the embedding, supervised (and semisupervised) learning algorithms

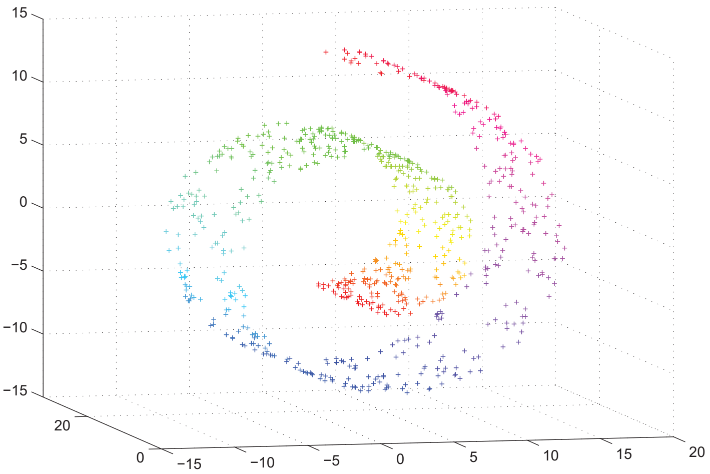


Figure 8. Swiss Roll

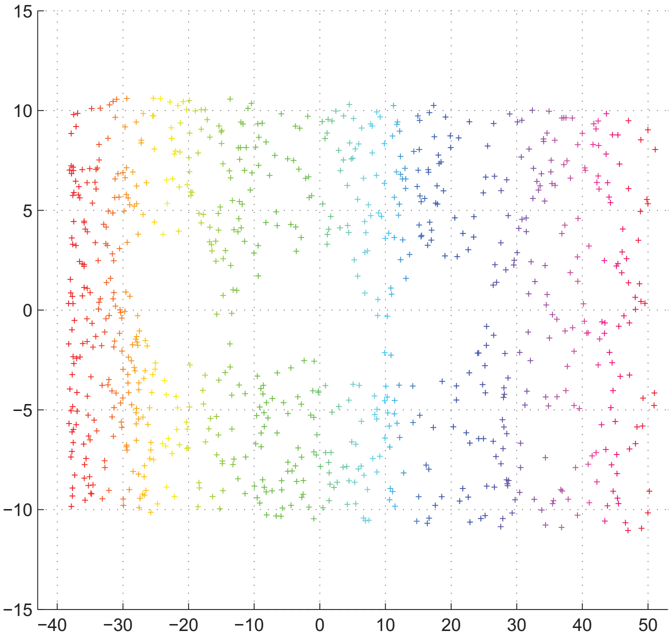


Figure 9. True Swiss Roll Unrolled

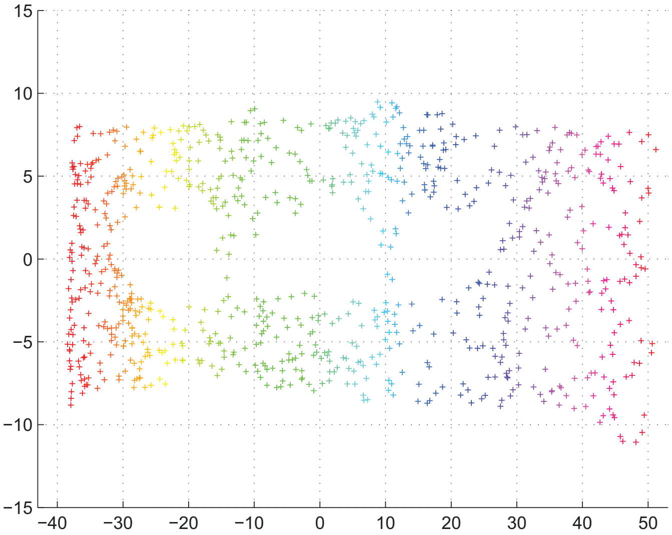


Figure 10. Swiss Roll with Noisy Data Unrolled

including the Support Vector Machine and SS-ANOVA models can be built using the embedded coordinates and newbies in conjunction with an RBF kernel for the embedded coordinates. The same tuning issues exist as we have seen so far, with the addition of the neighbor index k. This program has not been carried out to our knowledge, but it would be interesting to see how it might work on problems for which the RMU is more appropriate than the RKE for embedding. Certainly the two approaches can be compared on the same data set.

Here we just show plots of the embedding for the unrolled noisy Swiss roll. Noisy data was added to the Swiss roll by modifying 20% of the pairwise distances by a uniform random number between .85 and 1.15; the results of the unrolling are given in Figure 10. The k index was taken as k-nearest neighbor and chosen subjectively, as was λ here. The eigenvalues of the resulting R_{RMU} are plotted on a log scale in Figure 11. Two large eigenvalues make clear that that the unrolled figure sits in a two-dimensional space. The hanging eigenvalue is computational zero and reflects the constraint $E \cdot R = 0$.

4.1. Regularized Manifold Unfolding and Social Network Graphs.

A graph G is a vertex or node set and an edge set which contains the pair (i, j) if node i is connected to node j. Edges may have weights, but for this discussion we will assume that weights are 1. Recently a large literature has arisen with the goal of clustering the nodes, based on what amounts to embedding the nodes in a d-dimensional Euclidean space, and then using a standard clustering algorithm such as k-means. Let W be the $n \times n$ matrix with i, jth entry 1 if (i, j) is in the edge set and 0 otherwise, and let D be the diagonal matrix with (i, i)th entry $D_{ii} = \sum_k W_{ik}$. The normalized graph Laplacian L is defined as $L = I - D^{-1/2} W D^{-1/2}$. Let Z be the $n \times d$ matrix with columns the eigenvectors corresponding to the d largest absolute eigenvalues. The ith node is assigned the d Euclidean coordinates the ith row of Z.

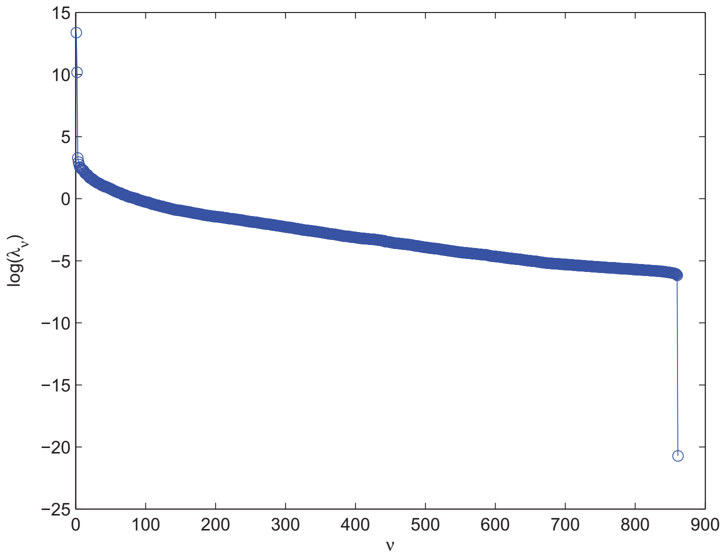


FIGURE 11. Eigenvalues of the embedding kernel

See, for example [**32**]. (Note that only the eigenvectors, and not the eigenvalues are used in the embedding there.) In these comments, we take a different view, which may be more appropriate for social networks, where the edges represent connections between people. Bearing in mind the "Six degrees of separation" of people, which posits that everyone on the planet is connected to everyone else through a path of at most 6 edges, we adopt the point of view that in order to cluster people, we only need to consider a very few degrees of separation, say, persons who are connected through at most, say $k = 2$ or $k = 3$ edges, and, if people are "further apart" than k edges we want to move them even further apart. Thus d_{ij} in (4.1) is either 1 or 2 if $k = 2$, and omitted otherwise; analogously if k is chosen to be 3. Alternatively the d_{ij} could be normalized to take account of all distinct paths of length less or equal to k joining node i and node j. It can be argued that Regularized Manifold Unfolding using R_{RMU} for embedding, as discussed here and in more detail in [**27**], provides an appealing alternative to the usual spectral clustering in the case where more than k edges is considered very far away, and pairs of points which are far away want to be pushed even more far away.

5. Conclusions, Further Work

The three papers discussed here together provide an approach to statistical model building/machine learning which incorporates scattered, noisy dissimilarity information and other information into nonparametric regression/classification models based on positive definite functions and reproducing kernel Hilbert spaces. In recent years there has been a huge growth in the literature in the use of dissimilarity information. As of this writing google provides over 4000 hits for the phrase "dissimilarity data" with the words "machine" and "learning, and the surface has only begun to be scratched.

Combinations of medical image modalities along with patient attribute information including clinical/environmental and genetic variables to build models to predict outcomes, including multiple correlated outcomes, is a rich area for exploration as data becomes available. Modern genetic data sets can include hundreds of thousands of genetic markers or gene expression information, providing serious challenges to figure out patterns or clusters of important variables [**35**] which when added to other information, can best be used to predict medical associations or outcomes of interest. Interacting effects of various sources of data often need to be realistically included. Analysis of network data along with attributes of various kinds assigned to nodes, which could be time-dependent, scattered or noisy or have censored or missing information [**28**] present interesting challenges in model building. An important key to success with observational dissimilarity data is to bring to the analysis definitions of what constitutes "dissimilarity" that are meaningful for the context at hand.

We have only discussed penalties in an RKHS, (other than [**35**]) but various kinds of quadratic, ℓ_1 and other penalties appear in the literature, designed for particular model structures. These models generally require multiple tuning parameters which balance goodness of fit to constraints on the model.

For unsupervised data, the tuning parameter(s) in RKE may be chosen by CV2, a cross validation approach involving leaving out pairs [**6**] Section A.2, [**25**] Section 3.5, or [**42**]. The pattern of eigenvalues appearing in [**27**] is relatively insensitive to λ over several orders of magnitude near the minimum. This may be a result of the fact that there are four very distinct groups of tightly clustered points and they can be embedded well in just three or four dimensions. However, if the RKE is part of a supervised or semisupervised model, it may be advisable to tune it along with the other tunable parameters of the model according to the target of the model, and in this case the estimated target of the model can be sensitive to the RKE tuning parameters. The problem of tuning complex models with differing objectives has some open questions. One of them is related to the issue of tuning to optimize prediction and tuning to optimize variable selection - not always the same, see [**20**].

Statistical model building and machine learning areas are experiencing a rapidly growing literature, spurred on by the availability of large training sets and increasing computer power. Nevertheless, many different data structures and models remain to be carefully studied. Theoretical properties should be obtained where possible and efficient computational algorithms must be developed. New models need to be tested on realistic simulated data, and finally applied to extract information from observational training sets of interest and importance.

Acknowledgments

We thank David Callan, Sijian Wang, Karl Rohe and Ben Hecht for helpful comments. This work has been partly supported by NSF Grant DMS0906818, NIH Grant EY09946, and ONR Grant N00014-09-1-0655.

References

[1] N. Aronszajn, *Theory of reproducing kernels*, Trans. Am. Math. Soc. **68** (1950), 337–404.

[2] M. Belkin & P. Niyogi, *Laplacian eigenmaps for dimensionality reduction and data representation*, Neural Computation **15(6)** (2005), 1373–1396.

[3] S. Benson & Y. Ye, *DSDP5: A software package implementing the dual-scaling algorithm for semidefinite programming*, Tech. Report ANL/MCS-TM-255, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, June 2004.

[4] E. Candes & B. Recht, *Exact matrix completion via convex programming*, Found. Comput. Math **9** (2008), 717–772.

[5] Y. Chen, M. Gupta & B. Recht, *Learning kernels from indefinite similarities*, Proceedings 26th International Conference on Machine Learning (Montreal), 2009.

[6] H. Corrada Bravo, *Graph-based data analysis: Tree-structured covariance estimation, prediction by regularized kernel estimation and aggregate database query processing for probabilistic inference*, Ph.D. thesis, Department of Statistics, University of Wisconsin, Madison WI, 2008, Technical Report 1145.

[7] H. Corrada Bravo, G. Wahba, K.E. Lee, B.E.K. Klein, R. Klein & 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** (2009), 8128–8133, PMCID: PMC 2677979.

[8] G. de los Campos, *Semi-parametric methods with applications to quantitative genetics and production economics*, Ph.D. thesis, Department of Animal Science, University of Wisconsin-Madison, Madison WI, 2009.

[9] G. de los Campos, D. Gianola, G. Rosa, K. Weigel & J. Crossa, *Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert space methods*, Genetics Research **92** (2010), 295–308.

[10] G. de los Campos, H. Naya, D. Gianola, J. Crossa, A. Legarra, E. Manfredi, K. Weigel & J. Cotes, *Predicting quantitative traits with regression models for dense molecular markers and pedigree*, Genetics **182** (2009), 375–385.

[11] D. Donoho & C. Grimes, *Hessian eigenmaps: locally linear embedding techniques for high dimensional data*, Proceedings of the National Academy of Arts and Sciences **100** (2003), 5591–5596.

[12] F. Gao, G. Wahba, R. Klein & B. Klein, *Smoothing spline ANOVA for multivariate Bernoulli observations, with applications to ophthalmology data, with discussion*, J. Amer. Statist. Assoc. **96** (2001), 127–160.

[13] C. Gu, *Smoothing spline anova models*, Springer, 2002.

[14] A. Kanda, W. Chen, M. Othman, K. Branham, M. Brooks, R. Khanna, S.He, R. Lyons, G. Abecasis & A. Swaroop, *A variant of mitochondrial protein LOC387715/ARMS2, not HTRA1 is strongly associated with age-related macular degeneration*, Proc. Natl. Acad. Sci. **104** (2007), 16227–16232.

[15] G. Kimeldorf & G. Wahba, *Some results on Tchebycheffian spline functions*, J. Math. Anal. Applic. **33** (1971), 82–95.

[16] K. Weinberger, F. Sha & L. Saul, *Learning a kernel matrix for nonlinear dimensionality reduction*, ICML '04: Proceedings of the twenty-first international conference on Machine learning (New York, NY, USA), ACM Press, 2004, p. 106.

[17] G. Lanckriet, N. Cristianini, P. Bartlett, L. ElGhoui & M. Jordan, *Learning the kernel matrix with semidefinite programming*, J. Mach. Learn. Res. **5** (2004), 27–72.

[18] H. Lee, M. Chung, H. Kang, B. Kim & D. Lee, *Computing the shape of brain networks using graph filtration and Gromov-Hausdorff metric*, Tech. Report 215, Department of Statistics, University of Wisconsin, Madison WI, 2011.

[19] Y. Lee, Y. Lin & G. Wahba, *Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data*, J. Amer. Statist. Assoc. **99** (2004), 67–81.

[20] C. Leng, Y. Lin & G. Wahba, *A note on the LASSO and related procedures in model selection*, Statistica Sinica **16** (2006), 1273–1284.

[21] X. Lin, G. Wahba, D. Xiang, F. Gao, R. Klein & B. Klein, *Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV*, Ann. Statist. **28** (2000), 1570–1600.

[22] ———, *Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV*, Ann. Statist. **28** (2000), 1570–1600.

[23] Y. Liu & H.H. Zhang, *Hard or soft classification?. Large margin unified machines (LUMs)*, Manuscript, 2009, talk at JSM 2009 and personal communication.

[24] S. Lloyd & C. Mallows, *An index of genealogical relatedness derived from a genetic model*, Ann. Prob. **1** (1973), 758–771.

[25] F. Lu, *Regularized nonparametric logistic regression and kernel regularization*, Ph.D. thesis, Department of Statistics, University of Wisconsin, Madison, 2006, Technical Report 1124.

[26] F. Lu, S. Keles, S. Wright & G. Wahba, *A framework for kernel regularization with application to protein clustering*, Proceedings of the National Academy of Sciences **102** (2005), 12332–12337, Open Source at www.pnas.org/content/102/35/12332, PMCID: PMC118947.

[27] F. Lu, Y. Lin & G. Wahba, *Robust manifold unfolding with kernel regularization*, Tech. Report 1008, Department of Statistics, University of Wisconsin, Madison WI, 2005.

[28] X. Ma, B. Dai, R. Klein, B. Klein, K. Lee & G. Wahba, *Pealized likelihood regression in reproducing kernel Hilbert spaces with randomized covariate data*, Tech. Report 1158, Department of Statistics, University of Wisconsin, Madison WI, 2010, under revision.

[29] K. Magnusson, S. Duan, H. Sigurdsson, H. Petursson, Z. Yang, Y. Zhao, P. Bernstein, J. Ge, F. Jonasson & E. Stefansson, *CFH Y402H confers similar risk of soft drusen and both forms of advanced AMD*, PLoS Med **3** (2005), e5.

[30] G. Malecot, *Les mathematiques de L'Heridite*, Masson et Cie, 1948.

[31] D. Pachauri, C. Hinrichs, M. Chung, S. Johnson & V. Singh, *Topology based kernels with application to inference problems in Alzheimer's disease*, manuscript, 2011.

[32] K. Rohe, S. Chatterjee & B. Yu, *Spectral clustering and the high-dimensional stochastic block model*, Tech. Report 791, Statistics Department, UC Berkeley, Berkeley, CA, 2010.

[33] S. Roweis & L. Saul, *Nonlinear dimensionality reduction by locally linear embedding*, Science **290** (2000), 2323–2326.

[34] T. Shi, M. Belkin & B. Yu, *Data spectroscopy: Eigenspaces of convolution operators and clustering*, Ann. Statist. **to appear** (2009), xx–xx.

[35] W. Shi, G. Wahba, S. Wright, K. Lee, B. Klein & R. Klein, *LASSO Pattern search algorithm with applications to ophthalmology and genomic data*, Statistics and Its Interface **1** (2008), 137–153, SII-1-1-A12-Shi.pdf, PMCID:PMC2566544.

[36] J. Tenenbaum, V. de Silva & J. Langford, *G global geometric framework for nonlinear dimensionality reduction*, Science **290** (2000), 2391–2323.

[37] R.H. Tütüncü, K.C. Toh & M.J. Todd, *Solving semidefinite-quadratic-linear programs using SDPT3*, Mathematical Programming **95** (2003), no. 2, 189–217.

[38] G. Wahba, *Spline models for observational data*, SIAM, 1990, CBMS-NSF Regional Conference Series in Applied Mathematics, v. 59.

[39] ———, *Multivariate function and operator estimation, based on smoothing splines and reproducing kernels*, Nonlinear Modeling and Forecasting, SFI Studies in the Sciences of Complexity, Proc. Vol XII (M. Casdagli and S. Eubank, eds.), Addison-Wesley, 1992, pp. 95–112.

[40] ———, *Smoothing splines in nonparametric regression*, Encyclopedia of Environmetrics (A. El-Shaarawi & W. Piegorsch, eds.), vol. 4, Wiley, 2001, pp. 2099–2112.

[41] ———, *Soft and hard classification by reproducing kernel Hilbert space methods*, Proceedings of the National Academy of Sciences **99** (2002), 16524–16530, Open Source at www.pnas.org/content/99/26/16524, PMCID: PMC125262.

[42] ———, *Dissimilarity data and regularized kernel estimation in classification and clustering*, Talk, Duke University, March 31, 2004, 2004, Available via the TALKS link at `http:www.stat.wisc.edu/ wahba`.

[43] ———, *Encoding dissimilarity data for statistical model building*, J. Statistical Planning and Inference (2010), 3580–3596, PMCID: PMC2929577 [available on 2011/12/1].

[44] G. Wahba, Y. Lin, Y. Lee & H. Zhang, *Optimal properties and adaptive tuning of standard and nonstandard support vector machines*, Nonlinear Estimation and Classification (D. Denison, M. Hansen, C. Holmes, B. Mallick, and B. Yu, eds.), Springer, 2002, pp. 129–148.

[45] G. Wahba, Y. Lin, Y. Lee, H. Zhang, D. Nychka & W. Wong, *The 2003 Wald lectures, with discussion*, Tech. Report 1080, Department of Statistics, University of Wisconsin, Madison WI, 2003.

[46] G. Wahba, Y. Wang, C. Gu, R. Klein & B. Klein, *Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy*, Ann. Statist. **23** (1995), 1865–1895, Neyman Lecture.

[47] D. Xiang & G. Wahba, *A generalized approximate cross validation for smoothing splines with non-Gaussian data*, Statistica Sinica **6** (1996), 675–692.

[48] X. Xie, M. Chung & G. Wahba, *Magnetic resonance image segmentation with thin plate spline thresholding*, Tech. Report 1105, Department of Statistics, University of Wisconsin, Madison WI, 2006.

[49] X. Xu & A. Goldberg, *Introduction to semi-supervised learning*, Morgan Claypool, 2009.

Department of Statistics, University of Wisconsin-Madison
E-mail address: wahba@stat.wisc.edu