ARTICLE IN PRESS

Journal of Statistical Planning and Inference ▮ (▮▮▮▮) ▮▮▮–▮▮▮

Contents lists available at ScienceDirect

Journal of Statistical Planning and Inference

journal homepage: www.elsevier.com/locate/jspi

Encoding dissimilarity data for statistical model building

Grace Wahba

Department of Statistics, University of Wisconsin-Madison, USA

ARTICLE INFO

For the volume in honor of the 80th birthday of distinguished Professor Emmanuel Parzen

Keywords:
Dissimilarity data
Reproducing kernel Hilbert spaces
Regularized kernel estimation
Regularization manifold unfolding
Penalized likelihood
Support vector machines
Radial basis functions

ABSTRACT

We summarize, review and comment upon three papers which discuss the use of discrete, noisy, incomplete, scattered pairwise dissimilarity data in statistical model building. Convex cone optimization codes are used to embed the objects into a Euclidean space which respects the dissimilarity information while controlling the dimension of the space. A "newbie" algorithm is provided for embedding new objects into this space. This allows the dissimilarity information to be incorporated into a smoothing spline ANOVA penalized likelihood model, a support vector machine, or any model that will admit reproducing kernel Hilbert space components, for nonparametric regression, supervised learning, or semisupervised learning. Future work and open questions are discussed. The papers are:

(1) Lu, F., Keles, S., Wright, S., Wahba, G., 2005a. A framework for kernel regularization with application to protein clustering. Proc. Natl. Acad. Sci. 102, 12332–12337.

(2) Corrada Bravo, G., Wahba, G., Lee, K., Klein, B., Klein, R., Iyengar, S., 2009. Examining the relative influence of familial, genetic and environmental covariate information in flexible risk models. Proc. Natl. Acad. Sci. 106, 8128–8133.

(3) Lu, F., Lin, Y., Wahba, G., 2005b. Robust manifold unfolding with kernel regularization. Technical Report 1008, Department of Statistics, University of Wisconsin-Madison.

1. Introduction

In this paper we summarize, review and add commentary to three papers (Lu et al., 2005a, b; Corrada Bravo et al., 2009) which involve machine learning/statistical model building problems where discrete, scattered, noisy, incomplete pairwise dissimilarity information is the main, or at least an important, source of information about objects in the training set.

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 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 may be 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 three papers 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.

The two examples discussed here are the use of BLAST scores to provide a dissimilarity score between pairs of protein sequences, which can be used to visualize and classify proteins from Lu et al. (2005a) (Section 2), and the use of pedigree

E-mail address: wahba@stat.wisc.edu

0378-3758/$ - see front matter © 2010 Elsevier B.V. All rights reserved.
doi:10.1016/j.jspi.2010.04.025

Please cite this article as: Wahba, G., Encoding dissimilarity data for statistical model building. J. Statist. Plann. Inference (2010), doi:10.1016/j.jspi.2010.04.025

(relationship) data in a demographic study of an eye condition in conjunction with other, direct information to build a risk model (Section 3.5) from Corrada Bravo et al. (2009).

The embedding method discussed in Lu et al. (2005a) as well as in Lu et al. (2005b) 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 Corrada Bravo et al. (2009).

In Section 4, we consider a modification of the method in Section 2 from Lu et al. (2005b) 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; it can also have the effect of enhancing clustering by moving near neighbors closer while relaxing the distance on further objects. This task, 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 main content of this review has been liberally extracted from the three papers cited, while we add commentary and discussion of their interrelationships, tuning, and open questions. Gaussian and Matern radial basis functions for incorporating embedded data in learning models are discussed in an Appendix.

2. Dissimilarity information and regularized kernel estimation (RKE)

This section is based on Lu et al. (2005a). 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:

$$\min_{K\in S_N} \sum_{(i,j)\in\Omega} L(w_{ij}, d_{ij}, \hat{d}_{ij}(K)) + \lambda J(K), \tag{1}$$

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 balancing 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 the graph of the pairs of objects in Ω connected by edges be connected. A pair may have repeated observations, which just yield an additional term in (1) for each separate observation. If the pair set induces a connected graph, then the minimizer of (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:

$$\min_{K\succcurlyeq 0} \sum_{(i,j)\in\Omega} w_{ij} |d_{ij} - B_{ij} \cdot K| + \lambda \ \text{trace}(K). \tag{2}$$

Trace was used as a regularizer in Lanckriet et al. (2004) 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) for a set of N objects is augmented by the addition of one more object together with its dissimilarity data. A variant of (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 Lu et al. (2005a).

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

Please cite this article as: Wahba, G., Encoding dissimilarity data for statistical model building. J. Statist. Plann. Inference (2010), doi:10.1016/j.jspi.2010.04.025

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 $x=(x(1),\ldots,x(q))\in\mathcal{R}^q$ 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

$$\min_{X_j,x_i,z}\quad\sum_{j=1}^{n_s}C_j\cdot X_j+\sum_{i=1}^{n_q}c_i\cdot x_i+g\cdot z \tag{3}$$

$$\begin{aligned}\text{s.t.}\quad&\sum_{j=1}^{n_s}A_{rj}\cdot X_j+\sum_{i=1}^{n_q}a_{ri}\cdot x_i+g_r\cdot z=b_r,\quad\forall_r\\&X_j\in S_{s_j}\ \forall_j;\quad x_i\in Q_{q_i}\ \forall_i;\quad z\in P_p.\end{aligned} \tag{4}$$

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 (Tütüncü et al., 2003) and DSDP5 (Benson and Ye, 2004).

2.1.2. RKE with l_1 loss

To convert the problem of Eq. (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 $r=1,2,\ldots,m$. 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 $(i(r),j(r))$, and with some abuse of notation let $i=i(r)$, $j=j(r)$ and $w\in P_{2m}$ with $w(r)=w(r+m)=w_{i(r),j(r)}$, $r=1,\ldots,m$, we can formulate the problem of Eq. (2) as follows:

$$\begin{aligned}\min_{K\succcurlyeq 0,u\geq 0}\quad&w\cdot u+\lambda I_N\cdot K\\\text{s.t.}\quad&d_{ij}-B_{ij}\cdot K+e_{m,r}\cdot u=0,\quad\forall_r,\\&K\in S_N,\quad u\in P_{2m}.\end{aligned} \tag{5}$$

2.2. Embedding

In the example in Lu et al. (2005a) 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. Fig. 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 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 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 $Z_{280\times 3}=\Gamma_{280\times 3}\Lambda_{3\times 3}^{1/2}$ where $\Gamma_{280\times 3}$ is the 280×3 matrix of the three leading vectors of K and Λ_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, $i=1,\ldots,280$. The method automatically centers the collection of the $x(i)$ at 0. Fig. 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) 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

$$\tilde{K}_{N+1}=\begin{bmatrix}K_N & b^T\\ b & c\end{bmatrix}\succcurlyeq 0 \tag{6}$$

Please cite this article as: Wahba, G., Encoding dissimilarity data for statistical model building. J. Statist. Plann. Inference (2010), doi:10.1016/j.jspi.2010.04.025

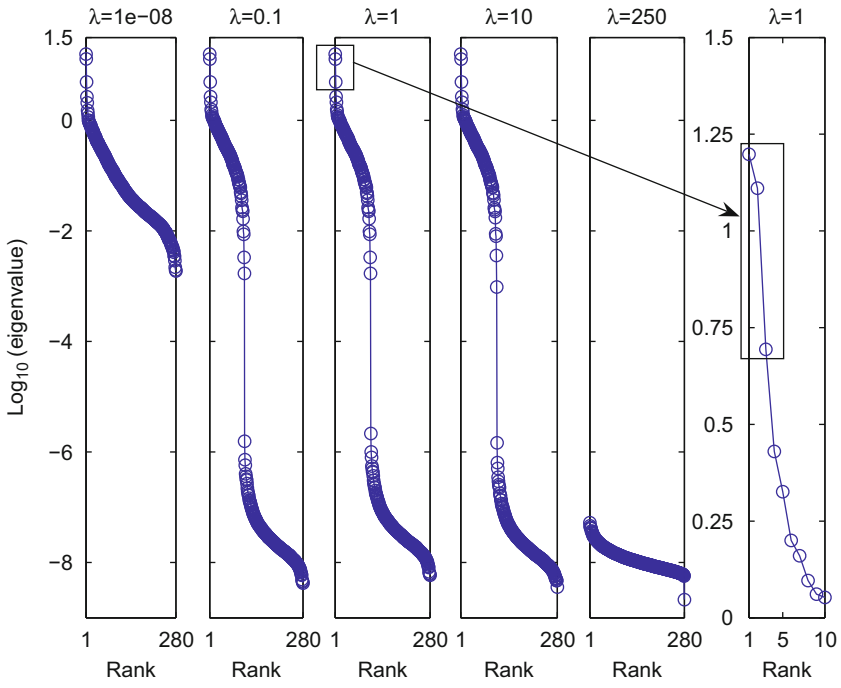

Fig. 1. Left five panels: log scale eigensequence plots for five values of λ. As λ increases, smaller eigenvalues begin to shrink. Right panel: first 10 eigenvalues of the $\lambda = 1$ case displayed on a larger scale.

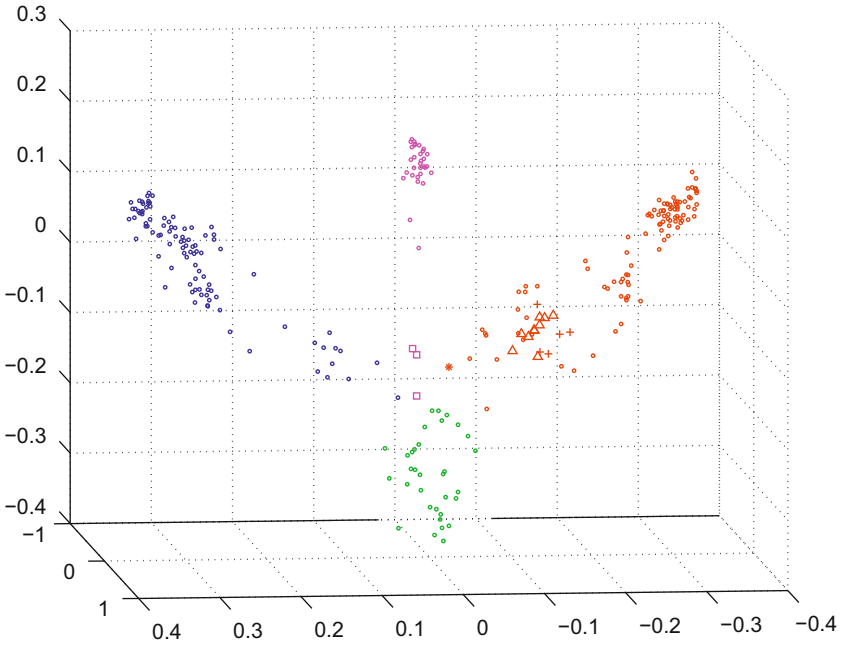

Fig. 2. 3D representation of the sequence space for 280 proteins from the globin family. Different subfamilies are encoded with different colors: Red symbols are alpha-globin subfamily, blue symbols are beta-globins, purple symbols represent myoglobin subfamily, and green symbols, scattered in the middle, 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 symbol △.

Please cite this article as: Wahba, G., Encoding dissimilarity data for statistical model building. J. Statist. Plann. Inference (2010), doi:10.1016/j.jspi.2010.04.025

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

$$\min_{c \geq 0, b} \quad \sum_{i \in \Psi} w_i |d_{i,N+1} - B_{i,N+1} \cdot K_{N+1}|$$
$$\text{s.t.} \quad b \in \text{Range}(K_N), \quad c - b^T K_N^+ b \geq 0, \tag{7}$$

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

$$Z \stackrel{\text{def}}{=} \begin{bmatrix} 1 & \tilde{c} \\ \tilde{c} & c \end{bmatrix} \in S_2. \tag{8}$$

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

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

Further we define the $N \times (p+1)$ matrix $\Sigma \stackrel{\text{def}}{=} [0_N : 2\Gamma\Lambda^{1/2}]$, where 0_N is the zero vector of length N, and 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 iith 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 (7) by the following equivalent convex cone program:

$$\min_{Z \succcurlyeq 0, u \geq 0, x} \quad w \cdot u$$
$$\text{s.t.} \quad \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} \cdot Z = 1,$$
$$\begin{bmatrix} 0 & 0.5 \\ 0.5 & 0 \end{bmatrix} \cdot Z - \begin{bmatrix} 1 \\ 0_p \end{bmatrix} \cdot x = 0,$$

$$d_{i,N+1} - K_N(i,i) - \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} \cdot Z + \Sigma_{i\cdot} \cdot x + e_{t,r} \cdot u = 0, \quad \forall_{r = 1,2,\ldots,t}, \tag{9}$$

$$Z \in S_2, \quad x \in Q_{p+1}, \quad u \in P_{2t}, \tag{10}$$

where $i = i(r)$ as before. Note that the constraints on Z ensure that it has the form (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). Fig. 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 Lu et al. (2005a). 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 leghemoglobins was used.

2.4. *Classification overlay: the multicategory support vector machine*

In examining Fig. 2 it is clear that if a sufficient number of labels were given, a fairly successful classification algorithm could be built on these data, especially if a "none of the above" category is allowed. The multicategory support vector machine (MSVM) (Lee et al., 2004) 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 of building an MSVM on embedded dissimilarity data. See Lee et al. (2004) 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

Please cite this article as: Wahba, G., Encoding dissimilarity data for statistical model building. J. Statist. Plann. Inference (2010), doi:10.1016/j.jspi.2010.04.025

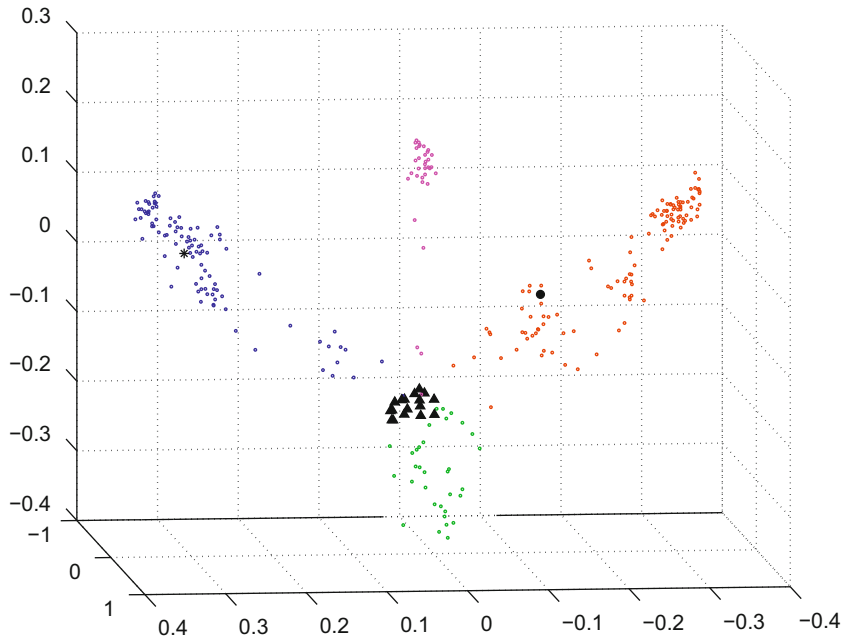

Fig. 3. Positioning test globin sequences in the coordinate system of 280 training sequences from the globin family. The newbie algorithm is used to locate one hemoglobin zeta chain (black circle), one hemoglobin theta chain (black star), and 17 leghemoglobins (black triangles) into the coordinate system of the training globin sequence data.

$f(x)=h(x)+b$ with $h \in \mathcal{H}_K$, an RKHS with reproducing kernel (RK) $K(\cdot,\cdot)$ and b, a constant minimizing

$$\frac{1}{n}\sum_{i=1}^{n}(1-y_i f(x(i)))_+ + \lambda\|h\|^2_{\mathcal{H}_K}, \tag{11}$$

where $(x)_+ = \max(x,0)$ and $\|h\|^2_{\mathcal{H}_K}$ denotes the square norm of h in $\mathcal{H}_K$. According to Kimeldorf and Wahba (1971), the minimizer h is of the form $h(x)=\sum_{i=1}^{n} c_i K(x,x(i))$ for some $c=(c_1,\ldots,c_n)$.. $(1-\tau)_+$ 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 (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 Wahba et al., 2002, 2003.) 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 $-1/(k-1)$ in the other positions. For instance, if example i falls into class 1, $y_i=(1,-1/(k-1),\ldots,-1/(k-1))$. Thus the components of each y_i are required to sum to zero. Accordingly, we define a k-tuple of separating functions $f(x)=(f^1(x),\ldots,f^k(x))$ 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 $f_\lambda=(f^1_\lambda,\ldots,f^k_\lambda)$, with each h^k in $\mathcal{H}_K$ satisfying the sum-to-zero constraint, which minimizes

$$\frac{1}{n}\sum_{i=1}^{n}\sum_{r\neq cat(i)}\left(f^r(x(i))+\frac{1}{k-1}\right)_+ + \lambda\sum_{j=1}^{k}\|h^j\|^2_{\mathcal{H}_K}, \tag{12}$$

where $cat(i)$ is the category of y_i. It is not hard to show that the $k=2$ case reduces to the usual 2-category SVM. The target for the MSVM is $f_\lambda(x)=(f^1_\lambda(x),\ldots,f^k_\lambda(x))$ 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 $f^j(x)=-1/(k-1)$ otherwise. Simulations in Lee et al. (2004) 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 Fig. 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

Please cite this article as: Wahba, G., Encoding dissimilarity data for statistical model building. J. Statist. Plann. Inference (2010), doi:10.1016/j.jspi.2010.04.025

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 RBFs. 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, Shi et al. (2009) give a novel take on clustering with the distance matrix corresponding to K_{RKE}.

3. Incorporating dissimilarity data into an SS-ANOVA model

This section is primarily based on Corrada Bravo et al. (2009). We begin with smoothing spline ANOVA (SS-ANOVA) models (Wahba et al., 1995; Gu, 2002; Lin et al., 2000; Wahba, 1990) which are a well known approach to penalized likelihood regression given heterogenous attribute variables, with the ability to model their various interactions. In Gao et al. (2001) an SS-ANOVA model was built to model 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 Lu et al. (2005a). Recently genetic markers have been found that are associated with a risk of AMD. See Kanda et al. (2007), Magnusson et al. (2005) and other references cited in Corrada Bravo et al. (2009). 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 Lu et al. (2005a), 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 Lu et al. (2005a), 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 Lu et al. (2005a), 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) $f(x) = p(x)/(1-p(x))$ and recover $p(x)$ from $p(x) = e^{f(x)}/(1+e^{f(x)})$. Given y_i, $x(i)$, $i = 1,2,\ldots,n$, $y \in \{0,1\}$, $x=(x_1,x_2,\ldots,x_d)$ the negative log likelihood in the Bernoulli case is given by

$$\mathcal{L}(y,f) = \sum_{i=1}^{n} -y_i f(x(i)) + \log(1+e^{f(x(i))}) \tag{13}$$

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

$$I(f) = \mathcal{L}(y,f) + \lambda J(f), \tag{14}$$

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 Wahba (2002). In the two category case, if the data are 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 Eq. (11). See Wahba (2002). They have quite different properties. Recently Liu and Zhang (2009) 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 reproducing kernel Hilbert space (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, $t_1,\ldots,t_k \in \mathcal{T}$, $a_1,\ldots,a_k \sum_{r,s=1}^{k} a_r a_s K(t_r,t_s) \geq 0$. The Moore–Aronszajn Theorem (Aronszajn, 1950) tells us that to every positive definite function $K(\cdot,\cdot)$ there corresponds a unique RKHS $\mathcal{H}_K$ and vice versa

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

Please cite this article as: Wahba, G., Encoding dissimilarity data for statistical model building. J. Statist. Plann. Inference (2010), doi:10.1016/j.jspi.2010.04.025

$\sum_r c_r K(\cdot,t_r) \in \mathcal{H}_K$,
$f \in \mathcal{H}_K \Rightarrow \langle f(\cdot),K(\cdot,t*)\rangle = f(t*),\ \forall t* \in \mathcal{T}$,
$\| \sum c_r K(\cdot,t_r)\|^2_{\mathcal{H}_K} = \sum_{rs} c_r c_s K(t_r,t_s)$.

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 (Wahba et al., 1995; Gu, 2002; Lin et al., 2000; Wahba, 1990) are based on ANOVA decompositions of functions of several variables. We describe the functional ANOVA decomposition is some generality. Let

$$x = (x_1,\ldots,x_d) \in \mathcal{X} \equiv \mathcal{X}^{(1)} \otimes \cdots \otimes \mathcal{X}^{(d)} \tag{15}$$

and

$$f(x) = f(x_1,\ldots,x_d), \quad x_\alpha \in \mathcal{X}^{(\alpha)}. \tag{16}$$

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

$$(\mathcal{E}_\alpha f)(x) = \int_{\mathcal{X}^{(\alpha)}} f(x_1,\ldots,x_d)\, d\mu_\alpha(x_\alpha). \tag{17}$$

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

$$f(x_1,\ldots,x_d) = \mu + \sum_\alpha f_\alpha(x_\alpha) + \sum_{\alpha\beta} f_{\alpha\beta}(x_\alpha,x_\beta) + \cdots, \tag{18}$$

where

$$\mu = \prod_\alpha \mathcal{E}_\alpha f = \int \cdots \int f(x_1,\ldots,x_d)\, d\mu_1(x_1)\cdots d\mu_d(x_d),$$

$$f_\alpha = (I-\mathcal{E}_\alpha)\prod_{\beta\neq\alpha} \mathcal{E}_\beta f,$$

$$f_{\alpha\beta} = (I-\mathcal{E}_\alpha)(I-\mathcal{E}_\beta)\prod_{\gamma\neq\alpha,\beta} \mathcal{E}_\gamma f,$$

$$\vdots$$

$$\mathcal{E}_\alpha f_\alpha = 0, \quad \mathcal{E}_\alpha \mathcal{E}_\beta f_{\alpha\beta} = 0 \quad \text{etc.}$$

The series in (18) is truncated at some point. Terms satisfy ANOVA-like side conditions which insure identifiability. An SS-ANOVA representation with weights on kernels looks like

$$f(\cdot) = \sum_{j=1}^m d_j \phi_j(\cdot) + \sum_{i=1}^n c_i K_\theta(\cdot,x(i)), \tag{19}$$

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

$$K_\theta(\cdot,\cdot) = \sum_{\alpha=1}^d \theta_\alpha K_\alpha(\cdot,\cdot), + \sum_{\alpha\leq\beta} \theta_{\alpha\beta} K_{\alpha\beta}(\cdot,\cdot) + \cdots\ . \tag{20}$$

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:

$$J(f) = \sum_{i,j=1}^n c_i c_j \left[\sum_{\alpha=1}^d \theta_\alpha^{-1} K_\alpha(x(i),x(j)) + \sum_{\alpha\leq\beta} \theta_{\alpha\beta}^{-1} K_{\alpha\beta}(x(i),x(j)) + \cdots\right], \tag{21}$$

where it is understood that only the components of $x(i)$ indicated by the subscripts on the kernel actually enter. The θ's are tuning parameters along with λ and with an identifiability constraint. For each trial set of tuning parameters, the c_i are to 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 are 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) (Xiang and Wahba, 1996) for Bernoulli data models with RKHS penalties is used in Corrada Bravo et al. (2009).

Please cite this article as: Wahba, G., Encoding dissimilarity data for statistical model building. J. Statist. Plann. Inference (2010), doi:10.1016/j.jspi.2010.04.025

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 Lin et al. (2000). Six hundred and eighty-four 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

$$\begin{aligned} f(t) = \mu &+ f_1(\mathtt{sys}) + f_2(\mathtt{chol}) + f_{12}(\mathtt{sys},\mathtt{chol}) \\ &+ d_{\mathtt{age}} \cdot \mathtt{age} + d_{\mathtt{bmi}} \cdot \mathtt{bmi} \\ &+ d_{\mathtt{horm}} \cdot I_1(\mathtt{horm}) + d_{\mathtt{drin}} \cdot I_2(\mathtt{drin}) + d_{\mathtt{smoke}} \cdot I_3(\mathtt{smoke}). \end{aligned} \tag{22}$$

This is the same model that was fitted in Lin et al. (2000) with the exception that `smoke` was not included there. In this model, $f_1 f_2$ and f_{12} are splines. Next we go on to add genetic markers and pedigree information to this model.

3.5. Modeling E/C, genetic and pedigree data in an SS-ANOVA model

In the study under discussion, logit has the representation

$$\begin{aligned} f(t) = \mu &+ d_{\mathrm{SNP1,1}} \cdot I(X_1 = 12) + d_{\mathrm{SNP1,2}} \cdot I(X_1 = 22) \\ &+ d_{\mathrm{SNP2,1}} \cdot I(X_2 = 12) d_{\mathrm{SNP2,2}} \cdot I(X_2 = 22) \\ &+ f_1(\mathrm{sysbp}) + f_2(\mathrm{chol}) + f_{12}(\mathrm{sysbp},\mathrm{chol}) \\ &+ d_{\mathrm{age}} \cdot \mathrm{age} + d_{\mathrm{bmi}} \cdot \mathrm{bmi} \\ &+ d_{\mathrm{horm}} \cdot I_1(\mathrm{horm}) + d_{\mathrm{drin}} \cdot I_2(\mathrm{drin}) + d_{\mathrm{smoke}} \cdot I_3(\mathrm{smoke}) + f_{ped}(z). \end{aligned} \tag{23}$$

The first two lines in (23) 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 know to be related to AMD. See Kanda et al. (2007), Magnusson et al. (2005) and references there. The next three lines are E/C variables, and the last line contains pedigree/relationship data to be explained shortly. Fig. 4(a) gives an example of a pedigree from BDES and Fig. 4(b) gives the relationship graph for five members of this pedigree. In Fig. 4(a) it can be seen that persons 35 and 26 are siblings, and are assigned a dissimilarity of 1 in Fig. 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 (Malecot, 1948; Lloyd and Mallows, 1973), 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 no known common ancestor are indicated with dashed edges in Fig. 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 cone optimization problem as in Section 2.1:

$$\min_{R \succeq 0} \sum_{(i,j)\in\Omega} |d_{ij} - B_{ij}(R)| + \lambda_{RKE}\mathrm{trace}(R). \tag{24}$$

$R_{\lambda_{RKE}}(i,j)$ then gives a (unique up to rotation) embedding $z(i)$, $i = 1,\ldots,n$ 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. Fig. 5 shows the embedding of the five people in the relationship graph of Fig. 4(b). These five people 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 (23). The horizontal axis (z_3) of this plot is orders of magnitude 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

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

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`	mmHg	Systolic blood pressure
`chol`	mg/dL	Serum cholesterol
`smoke`	Yes/no	History of smoking

Please cite this article as: Wahba, G., Encoding dissimilarity data for statistical model building. J. Statist. Plann. Inference (2010), doi:10.1016/j.jspi.2010.04.025

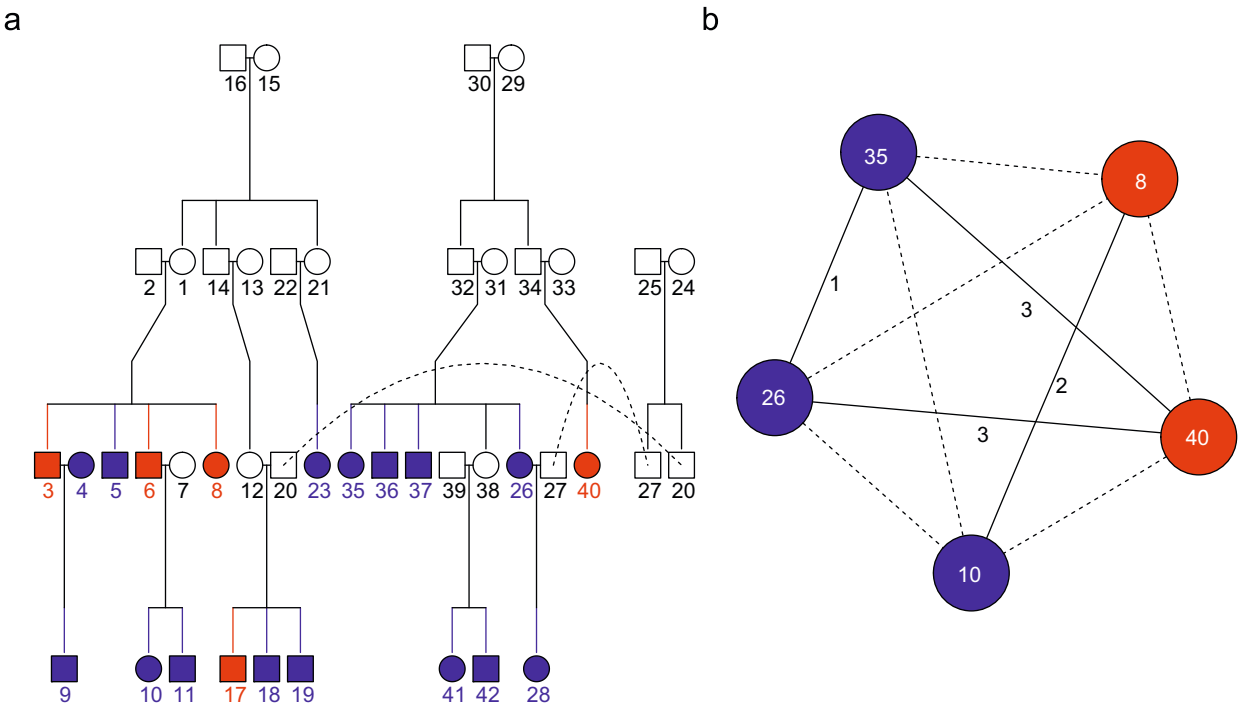

Fig. 4. An example pedigree from the BDES and a relationship graph for five subjects. Colored nodes are subjects assessed for retinal pigmentary abnormalities (red encodes a positive result). Circles are females and rectangles are males. (a) Example pedigree. (b) Relationship graph. Edge labels are dissimilarities defined by the kinship coefficient (sibling/parental=1, avuncular= 2, first cousins=3, etc.). Dotted edges indicate unrelated pairs.

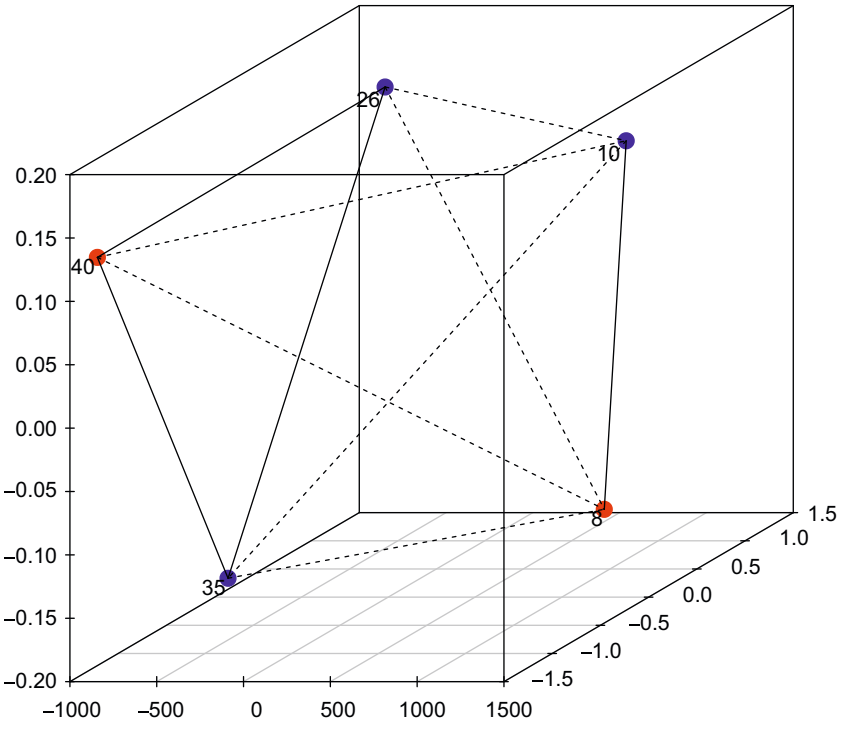

Fig. 5. Embedding of relationship graph in Fig. 4 by RKE. Note that the horizontal axis of this plot is orders of magnitude larger than the other two axes.

Please cite this article as: Wahba, G., Encoding dissimilarity data for statistical model building. J. Statist. Plann. Inference (2010), doi:10.1016/j.jspi.2010.04.025

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). The Matern family of RBFs and other RBFs are discussed in Appendix A; m and α are tuning parameters to be chosen. In the present work a Matern kernel of order $m=3$ was chosen. It is

$$K_z(z^*, z') = r_3(\|z^*-z'\|), \tag{25}$$

where

$$r_3(\tau) = \frac{1}{\alpha^7}\exp\{-\alpha\tau\}[15+15\alpha\tau+6\alpha^2\tau^2+\alpha^3\tau^3]. \tag{26}$$

If a newbie is not in a pedigree, then $K_z(z_{newbie},z(j))$ will be very small or 0 for all j. Eq. (20) becomes

$$K_\theta(\cdot,\cdot) = \sum_{\alpha=1}^{d}\theta_\alpha K_\alpha(\cdot,\cdot), + \sum_{\alpha\le\beta}\theta_{\alpha\beta}K_{\alpha\beta}(\cdot,\cdot)+\cdots+\theta_z K_z(\cdot,\cdot) \tag{27}$$

and $K_\theta(\cdot,x(j))$ of Eq. (19) becomes $K_\theta(\cdot,x(j):z(j))$, that is, $K_\theta(x,x(j)))$ becomes $K_\theta(x:z,x(j):z(j)$.

3.6. Tuning

We have the following tuning parameters:

- λ_{MAIN} of Eq. (14) 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 Eq. (27) subject to a single side condition so that they are identifiable in the presence of λ_{MAIN}.
- λ_{RKE} of Eq. (24) 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 (Xiang and Wahba, 1996; Lin et al., 2000) 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.

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

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 de los Campos et al. (2009) and de los Campos (2009) 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.

Please cite this article as: Wahba, G., Encoding dissimilarity data for statistical model building. J. Statist. Plann. Inference (2010), doi:10.1016/j.jspi.2010.04.025

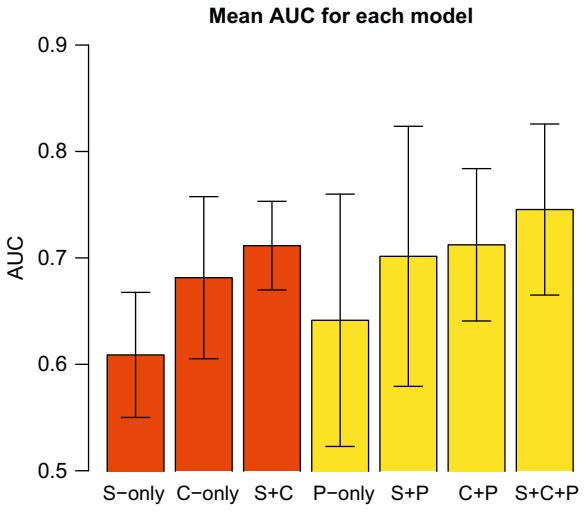

Fig. 6. AUC comparison of seven models.

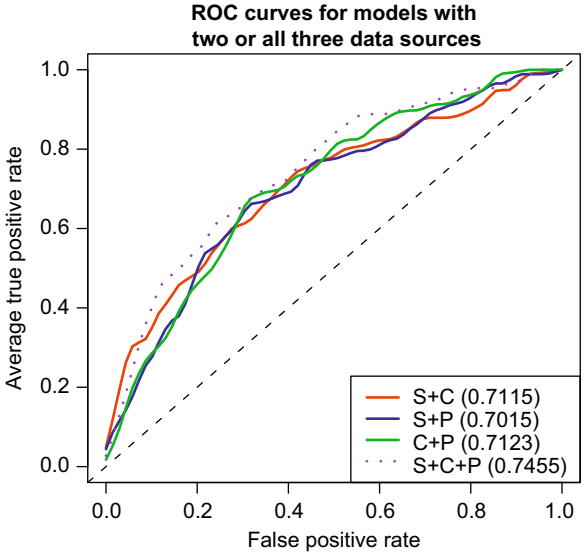

Fig. 7. ROC curves for two and three source models.

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. Fig. 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. Fig. 8 was constructed by "rolling up" the two dimensional data in Fig. 9. So, simply put, given the data in Fig. 8 contaminated by noise, can you recover (unroll, flatten) to get an estimate of the unrolled data in Fig. 9?

See the references in Lu et al. (2005b) for various approaches (Tenenbaum et al., 2000; Roweis and Saul, 2000; Belkin and Niyogi, 2005; Donoho and Grimes, 2003; Weinberger et al., 2004) to unrolling the Swiss roll, and many real applications. More recently, Zhu and Goldberg (2009) discuss manifold unrolling and give further references. In Lu et al. (2005b) 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 for pairs in Ω_k. Several equivalent

Please cite this article as: Wahba, G., Encoding dissimilarity data for statistical model building. J. Statist. Plann. Inference (2010), doi:10.1016/j.jspi.2010.04.025

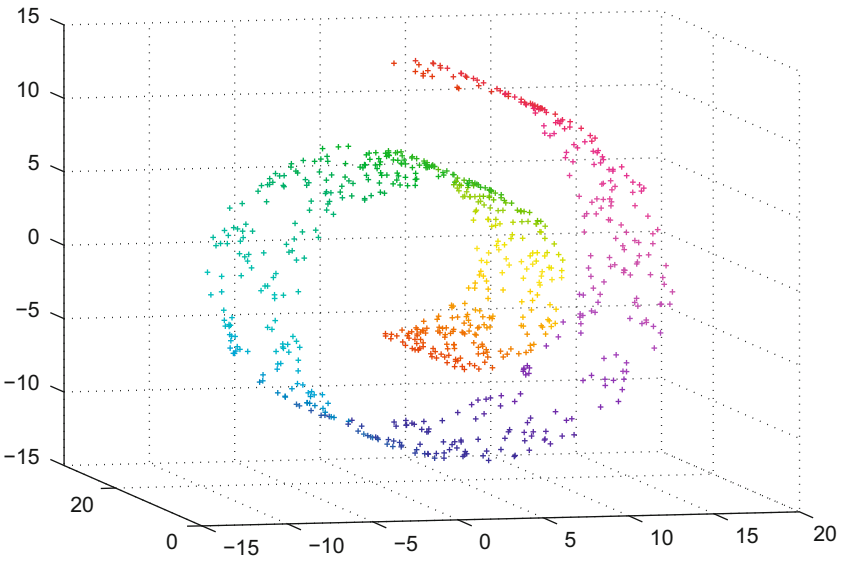

Fig. 8. Swiss roll.

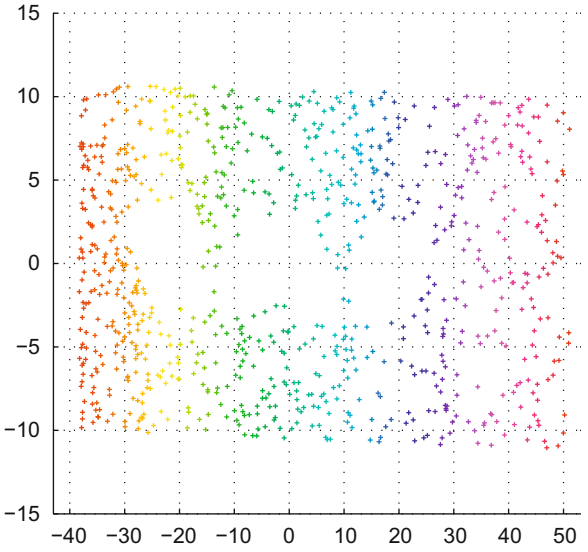

Fig. 9. True Swiss roll unrolled.

formulations of the solution to this problem are given; here we only describe one. The optimization problem is

$$\min_{R\succeq 0} \sum_{(i,j)\in\Omega_k} w_{ij}|d_{ij}-B_{ij}\cdot R|-2\lambda \text{ trace}(R) \tag{28}$$

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 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 were added to the Swiss roll by modifying 20% of the pairwise distances by a uniform random number between 0.85 and 1.15; the results of the unrolling are given in Fig. 10. The k index was taken as k-nearest neighbor and chosen subjectively, as was λ here. The eigenvalues of

Please cite this article as: Wahba, G., Encoding dissimilarity data for statistical model building. J. Statist. Plann. Inference (2010), doi:10.1016/j.jspi.2010.04.025

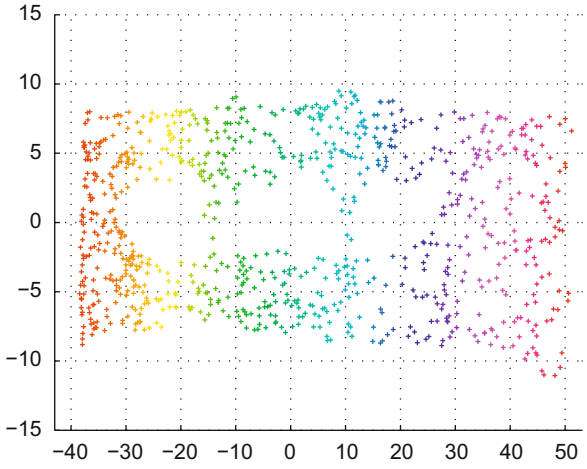

Fig. 10. Swiss roll with noisy data unrolled.

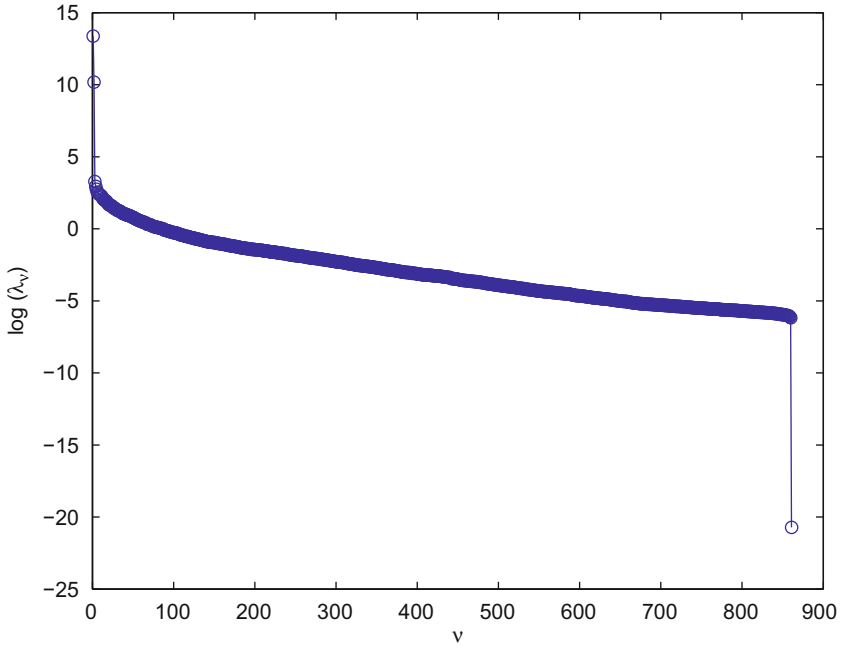

Fig. 11. Eigenvalues of the embedding kernel.

the resulting R_{RMU} are plotted on a log scale in Fig. 11. Two large eigenvalues make clear that the unrolled figure sits in a two-dimensional space. The hanging eigenvalue is computational zero and reflects the constraint $E \cdot R = 0$.

5. Conclusions, further work

The three papers discussed here provide an overall scheme for including discrete, heterogenous, scattered, noisy dissimilarity data into nonparametric classification and regression models, in the context of reproducing kernel Hilbert space methods. These models, which comprise tradeoffs between fit to the data, however defined, and complexity/wiggliness/degrees of freedom of the solution are typically obtained via the solution of an optimization problem. They comprise a large subset of the approaches known as statistical learning and statistical model building, but of course only scratch the surface. As attendees at the 2009 Joint Statistical Meetings may have observed, this class of methods is engaging a large number of Statistics researchers in many fields of application. Application to indirectly sensed observations (Wahba, 1977) go back to the 1970s, as do solving the optimization problems with inequality constraints (Wahba, 1990) and other side conditions, and these models could be brought up to date by merging them with

Please cite this article as: Wahba, G., Encoding dissimilarity data for statistical model building. J. Statist. Plann. Inference (2010), doi:10.1016/j.jspi.2010.04.025

dissimilarity observations via the methods discussed here. More recently, data sets with extremely large numbers of candidate attribute variables, as occur in genetics data, have given rise to many problems and techniques for variable selection, in particular with penalties involving absolute value norms and other norms which induce sparsity, and these can be merged with the methods described here. Methods for dealing with missing or noisy components in the direct attributes (Ma et al., 2010) have the potential for being included.

In some examples it is important to be able to consider interactions between the dissimilarity information and direct information, and this appears to be straightforward. When dealing with ordered dissimilarity information, it may be appropriate to tune the scale, for example scores like very close, close, distant, very distant; the four scores could be coded as 1, $1+\delta_1$, $1+\delta_1+\delta_2$, $1+\delta_1+\delta_2+\delta_3$, where the δ's are positive numbers to be chosen (tuned).

The RKE method is a relative of multidimensional scaling (MDS), another very old method (Kruskal, 1964). In MDS, dissimilarity data are given, a desired dimension d for Euclidean space is pre-selected and an algorithm finds a set of position coordinates in Euclidean d space that best fits the dissimilarity data according to a given criterion, for example least squares. The choice of dimension of the embedding space is essentially the tuning parameter, since the higher the dimension, the better the fit on training data, while too low a dimension may "oversmooth", or rather "overproject" the data with a bias towards reducing fitted dissimilarity. The RKE has the ability to "smooth" over higher dimensions if that is warranted, rather than slicing them off, but if the data firmly sits in a low dimensional space, that will be evident by examining the eigenvalue plot. For unsupervised data, the tuning parameter(s) in RKE may be chosen by CV2, a cross validation approach involving leaving out pairs (Corrada Bravo, 2008, Section A.2; Lu, 2006, Section 3.5; Wahba, 2004). The actual resulting pattern of eigenvalues appears to be relatively insensitive to λ over several orders of magnitude near the minimum. However, if the RKE is part of a supervised or semisupervised model, it is 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. Applications based on MDS abound, for example multidimensional unfolding (deLeeuw, 2004). In multidimensional unfolding, the graph of the connected pairs has a special structure; there are n_1 objects of type A and n_2 objects of type B, dissimilarities are observed between all pairs consisting of one member of type A and the other member of type B. The goal is to estimate "distances" between members within the two types. Garten et al. (2009) gives an interesting application of multidimensional unfolding, which the authors call antigenic maps, which estimates differences among influenza viruses (type A) based on the results of binding assay data points (type B). They provide a two dimensional map of the results. It would be interesting to compare multidimensional unfolding examples with tuned RKE.

Many extensions are ripe for application.

Acknowledgments

This work has been partly supported by NSF Grants DMS0604572 and DMS0906818, NIH Grant EY09946, and ONR Grant N00014-09-1-0655. The author thanks David Callan for assistance and Yuedong Wang for bringing the multidimensional unfolding work to her attention. It is appropriate to acknowledge a great debt to Professor Parzen from whom I learned about Reproducing Kernel Hilbert Spaces as a student: key references are Parzen (1962a, b, 1960, 1963, 1970).

Appendix A. Radial basis functions and the Matern class

Since the coordinates obtained via the RKE or RMU are based on distances, and are unique only up to rotation, only radial basis functions (RBFs), which depend only on distances between the two arguments of the RK are appropriate. According to Skorokhod and Yadrenko (1973) any $r(\tau)$ of the form

$$r(\tau)=\int_0^\infty \frac{J_{(d-2)/2}(\omega\tau)}{(\omega\tau)^{(d-2)/2}}h(\omega)\,d\omega, \tag{29}$$

where J_ν is the Bessel function of first kind of ν th order, generates an RBF kernel on Euclidean d space according to $R(s,t)=r(\tau)$ with $\tau=\|s-t\|$. The Gaussian RBF $r(\tau)=e^{-(1/2\sigma^2)\tau^2}$ is probably the most popular or familiar RBF in machine learning, with the single scale parameter σ^2, but it is not the only good choice. The Matern family of RBFs (Matern, 1986; Stein, 1999) provides a two parameter family of RBFs, with a scale parameter and a parameter which can be thought of as controlling the number of derivatives at the origin, equivalently the rate of decay of the Fourier transform of the RBF. Table 2 gives five Matern RBFs and their Fourier transforms ($h(\omega)$). The Fourier transforms may be compared to that of the Gaussian: $r(\tau)=e^{-\alpha\tau^2}\sim\sqrt{(\pi/\alpha)}e^{-\omega^2/4\alpha}$. Results can be quite sensitive to the scale parameter, either in the Gaussian (σ^2) or a Matern kernel (α). The Gaussian kernel is infinitely differentiable at the origin, and representers obtained from it will share that property, while the members of the Matern family have an increasing number of derivatives at the origin as the order goes up, beginning with no derivatives for the $m=0$ case. In the pedigree study (Corrada Bravo et al., 2009), some experimentation showed that a tuned Gaussian kernel was inferior to the third order Matern kernel that was used in the paper.

Please cite this article as: Wahba, G., Encoding dissimilarity data for statistical model building. J. Statist. Plann. Inference (2010), doi:10.1016/j.jspi.2010.04.025

Table 2
Matern RBF's and their Fourier transforms.

$$r(\tau)\sim h(\omega)\frac{1}{\alpha}e^{-\alpha\tau}\sim\kappa_d(\omega^2+\alpha^2)^{-((d+1)/2)},\quad m=0$$
$$\frac{1}{\alpha^3}e^{-\alpha\tau}[1+\alpha\tau]\sim\kappa_d(d+1)(\omega^2+\alpha^2)^{-((d+3)/2)},\quad m=1$$
$$\frac{1}{\alpha^5}e^{-\alpha\tau}[3+3\alpha\tau+\alpha^2\tau^2]\sim\kappa_d(d+1)(d+3)(\omega^2+\alpha^2)^{-((d+5)/2)},\quad m=2$$
$$\frac{1}{\alpha^7}e^{-\alpha\tau}[15+15\alpha\tau+6\alpha^2\tau^2+\alpha^3\tau^3]\sim\kappa_d(d+1)(d+3)(d+5)(\omega^2+\alpha^2)^{-((d+7)/2)},\quad m=3$$
$$\cdots\sim\cdots$$

References

Aronszajn, N., 1950. Theory of reproducing kernels. Trans. Amer. Math. Soc. 68, 337–404.
Belkin, M., Niyogi, P., 2005. Laplacian eigenmaps for dimensionality reduction and data representation. Neural Comput. 15 (6), 1373–1396.
Benson, S., Ye, Y., 2004. DSDP5: a software package implementing the dual-scaling algorithm for semidefinite programming. Technical Report ANL/MCS-TM-255, Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, June ⟨http://www-unix.mcs.anl.gov/∼benson/dsdp/dsdp5userguide.pdf⟩.
Corrada Bravo, H., 2008. 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, Technical Report 1145.
Corrada Bravo, H., Wahba, G., Lee, K.E., Klein, B.E.K., Klein, R., Iyengar, S.K., 2009. Examining the relative influence of familial, genetic and environmental covariate information in flexible risk models. Proc. Natl. Acad. Sci. 106, 8128–8133, doi: 10.1073/pnas.0902906106.
de los Campos, G., 2009. Semi-parametric methods with applications to quantitative genetics and production economics. Ph.D. Thesis, Department of Animal Science, University of Wisconsin-Madison, Madison, WI.
de los Campos, G., Naya, H., Gianola, D., Crossa, J., Legarra, A., Manfredi, E., Weigel, K., Cotes, J., 2009. Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182, 375–385.
deLeeuw, J., 2004. Multidimensional unfolding. Department of Statistics, UCLA ⟨http://repositories.cdlib.org/uclastat/papers/2004030301⟩.
Donoho, D., Grimes, C., 2003. Hessian eigenmaps: locally linear embedding techniques for high dimensional data. Proc. Natl. Acad. Arts Sci. 100, 5591–5596.
Gao, F., Wahba, G., Klein, R., Klein, B., 2001. Smoothing spline ANOVA for multivariate Bernoulli observations with applications to ophthalmology data with discussion. J. Amer. Statist. Assoc. 96, 127–160.
Garten, R., et al., 2009. Antigenic and genetic characteristics of swine=origin 2009 A(H1N1) influenza viruses circulating in humans. Science 325, 197–201.
Gu, C., 2002. Smoothing Spline ANOVA Models. Springer, New York.
Kanda, A., Chen, W., Othman, M., Branham, K., Brooks, M., Khanna, R., He, S., Lyons, R., Abecassis, G., Swaroop, A., 2007. A variant of mitochondrial protein LOC387715/ARMS2, not HTRA1 is strongly associated with age-related macular degeneration. Proc. Natl. Acad. Sci. 104, 16227–16232.
Kimeldorf, G., Wahba, G., 1971. Some results on Tchebycheffian spline functions. J. Math. Anal. Appl. 33, 82–95.
Kruskal, J., 1964. Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis. Psychometrika 29, 1–27.
Lanckriet, G., Cristianini, N., Bartlett, P., ElGhoui, L., Jordan, M., 2004. Learning the kernel matrix with semidefinite programming. J. Mach. Learn. Res. 5, 27–72.
Lee, Y., Lin, Y., Wahba, G., 2004. Multicategory support vector machines, theory, and application to the classification of microarray data and satellite radiance data. J. Amer. Statist. Assoc. 99, 67–81.
Lin, X., Wahba, G., Xiang, D., Gao, F., Klein, R., Klein, B., 2000. Smoothing spline ANOVA models for large data sets with Bernoulli observations and the randomized GACV. Ann. Statist. 28, 1570–1600.
Liu, Y., Zhang, H.H., 2009. Hard or soft classification? Large margin unified machines (LUMs). Manuscript, talk at JSM 2009 and personal communication.
Lloyd, S., Mallows, C., 1973. An index of genealogical relatedness derived from a genetic model. Ann. Probab. 1, 758–771.
Lu, F., 2006. Regularized nonparametric logistic regression and kernel regularization. Ph.D. Thesis, Department of Statistics, University of Wisconsin, Madison, Technical Report 1124.
Lu, F., Keles, S., Wright, S., Wahba, G., 2005a. A framework for kernel regularization with application to protein clustering. Proc. Natl. Acad. Sci. 102, 12332–12337 Open source at ⟨www.pnas.org/content/102/35/12332⟩, PMCID: PMC118947.
Lu, F., Lin, Y., Wahba, G., 2005b. Robust manifold unfolding with kernel regularization. Technical Report 1008, Department of Statistics, University of Wisconsin, Madison, WI.
Ma, X., Dai, B., Klein, R., Klein, B., Lee, K., Wahba, G., 2010. Penalized likelihood regression in reproducing kernel Hilbert spaces with randomized covariate data. University of Wisconsin-Madison Statistics Department, TR1158, submitted.
Magnusson, K.P., Duan, S., Sigurdsson, H., Petursson, H., Yang, Z., Zhao, Y., Bernstein, P.S., Ge, J., Jonasson, F., Stefansson, E., Helgadottir, G., Zabriskie, N.A., Jonsson, T., Bjórnsson, A., Thorlacius, T., Jonsson, P.V., Thorleifsson, G., Kong, A., Stefansson, H., Zhang, K., Stefansson, K., Gulcher, J.R., 2005. CFH Y402H confers similar risk of soft drusen and both forms of advanced AMD. PLoS Med. 3, e5.
Malecot, G., 1948. Les mathematiques de L'Heridite. Masson et Cie, Paris.
Matern, B., 1986. Spatial Variation, in: Lecture Notes in Statistics, vol. 36, Springer.
Parzen, E., 1960. Regression analysis of continuous parameter time series. In: Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability. University of California Press, Berkeley, CA, pp. 469–489.
Parzen, E., 1962a. An approach to time series analysis. Ann. Math. Statist. 32, 951–989.
Parzen, E., 1962b. Stochastic Processes. Holden-Day, San Francisco.
Parzen, E., 1963. Probability density functionals and reproducing kernel Hilbert spaces. In: Rosenblatt, M. (Ed.), Proceedings of the Symposium on Time Series Analysis. Wiley, New York, pp. 155–169.
Parzen, E., 1970. Statistical inference on time series by RKHS methods. In: Pyke, R. (Ed.), Proceedings 12th Biennial Seminar. Canadian Mathematical Congress, Montreal, pp. 1–37.
Roweis, S., Saul, L., 2000. Nonlinear dimensionality reduction by locally linear embedding. Science 290, 2323–2326.
Shi, T., Belkin, M., Yu, B., 2009. Data spectroscopy: eigenspaces of convolution operators and clustering. Ann. Statist. 37, 3960–3984.
Skorokhod, A., Yadrenko, M., 1973. On absolute continuity of measures corresponding to homogeneous Gaussian fields. Theory Probab. Appl. XVIII, 27–40.

Please cite this article as: Wahba, G., Encoding dissimilarity data for statistical model building. J. Statist. Plann. Inference (2010), doi:10.1016/j.jspi.2010.04.025

Stein, M., 1999. Interpolation of Spatial Data: Some Theory for Kriging. Springer-Verlag, New York.
Tenenbaum, J., de Silva, V., Langford, J., 2000. A global geometric framework for nonlinear dimensionality reduction. Science 290, 2323–2391.
Tütüncü, R.H., Toh, K.C., Todd, M.J., 2003. Solving semidefinite-quadratic-linear programs using SDPT3. Math. Programming 95 (2), 189–217.
Wahba, G., 1977. Practical approximate solutions to linear operator equations when the data are noisy. SIAM J. Numer. Anal. 14, 651–667.
Wahba, G., 1990. Spline models for observational data. In: SIAM, cBMS-NSF Regional Conference Series in Applied Mathematics, vol. 59.
Wahba, G., 2002. Soft and hard classification by reproducing kernel Hilbert space methods. Proc. Natl. Acad. Sci. 99, 16524–16530 Open Source at 〈www.pnas.org/content/99/26/16524〉, PMCID: PMC125262.
Wahba, G., 2004. Dissimilarity data and regularized kernel estimation in classification and clustering. Talk, Duke University, March 31, 2004, available via the TALKS link at 〈http://www.stat.wisc.edu/wahba〉.
Wahba, G., Lin, Y., Lee, Y., Zhang, H., 2002. Optimal properties and adaptive tuning of standard and nonstandard support vector machines. In: Denison, D., Hansen, M., Holmes, C., Mallick, B., Yu, B. (Eds.), Nonlinear Estimation and Classification. Springer, New York, pp. 129–148.
Wahba, G., Lin, Y., Lee, Y., Zhang, H., Nychka, D., Wong, W., 2003. The 2003 Wald lectures, with discussion. Technical Report 1080, Department of Statistics, University of Wisconsin, Madison, WI.
Wahba, G., Wang, Y., Gu, C., Klein, R., Klein, B., 1995. Smoothing spline ANOVA for exponential families with application to the Wisconsin epidemiological study of diabetic retinopathy. Ann. Statist. 23, 1865–1895 (Neyman Lecture).
Weinberger, K., Sha, F., Saul, L., 2004. Learning a kernel matrix for nonlinear dimensionality reduction. In: ICML Proceedings of the Twenty-first International Conference on Machine Learning. ACM Press, New York, NY, USA, pp. 106.
Xiang, D., Wahba, G., 1996. A generalized approximate cross validation for smoothing splines with non-Gaussian data. Statist. Sinica 6, 675–692.
Zhu, X., Goldberg, A., 2009. Introduction to Semi-Supervised Learning. Morgan Claypool, Princeton.

Please cite this article as: Wahba, G., Encoding dissimilarity data for statistical model building. J. Statist. Plann. Inference (2010), doi:10.1016/j.jspi.2010.04.025