Journal of Statistical Planning and Inference 140 (2010) 3580–3596

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
Available online 8 May 2010

*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

(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 $N$, $\Omega$ 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 $\lambda$ 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 $\Omega$ 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:

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

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