Available online at www.sciencedirect.com
ScienceDirect
Computational Statistics and Data Analysis 52 (2008) 4643–4657
COMPUTATIONAL STATISTICS & DATA ANALYSIS
www.elsevier.com/locate/csda

Semisupervised learning from dissimilarity data

Michael W. Trosset[a,*], Carey E. Priebe[b], Youngser Park[c], Michael I. Miller[c]

[a] *Department of Statistics, Indiana University, Bloomington, IN 47405, USA*
[b] *Department of Applied Mathematics & Statistics, Johns Hopkins University, Baltimore, MD 21218, USA*
[c] *Center for Imaging Science, Johns Hopkins University, Baltimore, MD 21218, USA*

Received 3 February 2007; received in revised form 23 February 2008; accepted 27 February 2008
Available online 4 March 2008

Abstract

The following two-stage approach to learning from dissimilarity data is described: (1) embed both labeled and unlabeled objects in a Euclidean space; then (2) train a classifier on the labeled objects. The use of linear discriminant analysis for (2), which naturally invites the use of classical multidimensional scaling for (1), is emphasized. The choice of the dimension of the Euclidean space in (1) is a model selection problem; too few or too many dimensions can degrade classifier performance. The question of how the inclusion of unlabeled objects in (1) affects classifier performance is investigated. In the case of spherical covariances, including unlabeled objects in (1) is demonstrably superior. Several examples are presented.

1. Introduction

Let $(\Omega, \mathcal{F}, \mathcal{P})$ denote a probability space, i.e., Ω is a sample space, $\mathcal{F}$ is a sigma-field, and $\mathcal{P}$ is a probability measure. Suppose that $\mathcal{P}$ is a mixture, i.e.,

$$\mathcal{P} = \sum_{i=1}^{k} \alpha_i \mathcal{P}_i, \tag{1}$$

where $\alpha_i \geq 0$ and $\sum_{i=1}^{k} \alpha_i = 1$. When $\omega \in \Omega$ is drawn from $\mathcal{P}$, it is drawn from one of the $\mathcal{P}_i$. If $\omega \sim \mathcal{P}_i$, then we say that ω belongs to class i. If we know the class to which ω belongs, then we say that ω is labeled; otherwise, ω is unlabeled. Our goal is to construct a classifier, i.e., a function that assigns a label $i \in \{1, \ldots, k\}$ to an unlabeled $\omega \in \Omega$.

Without supposing that $(\Omega, \mathcal{F})$ is a Euclidean space, we intend to construct a labeling function by Euclidean methods, e.g., linear discriminant analysis (LDA). Accordingly, we are concerned with classifiers of the form $\mathbf{L} \circ \mathbf{M}$, where $\mathbf{M} : \Omega \to \mathbf{X}$ and $\mathbf{L} : \mathbf{X} \to \{1, \ldots, k\}$. The *representation space* $\mathbf{X}$ is a measurable metric space, typically $(\Re^d, \mathcal{B})$ where $\mathcal{B}$ is the Borel sigma-field, with a metric induced by the Euclidean inner product. The map $\mathbf{M}$ is the

* Corresponding address: P.O. Box 6424, Bloomington, IN 47407, USA. Tel.: +1 812 856 7824.
E-mail addresses: mtrosset@indiana.edu (M.W. Trosset), cep@jhu.edu (C.E. Priebe), youngser@jhu.edu (Y. Park), mim@cis.jhu.edu (M.I. Miller).

0167-9473/$ - see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.csda.2008.02.030

embedding function, and the map $\mathbf{L}$ is the *labeling function*. Only labeled outcomes are used to construct the labeling function, while both labeled and unlabeled outcomes are used to construct the embedding function. This paradigm exemplifies *semisupervised learning*, in which one has access to both labeled and unlabeled outcomes.

We further suppose that we are unable to observe $\omega \in \Omega$ directly, but that we have access to a function $\delta : \Omega \times \Omega \to \Re$ that behaves in such a way that we have agreed to interpret $\delta(\omega_1, \omega_2)$ as the "dissimilarity" of outcomes ω_1 and ω_2. We assume that $\delta(\omega_1, \omega_2) \geq 0$, that $\delta(\omega, \omega) = 0$, and that $\delta(\omega_1, \omega_2) = \delta(\omega_2, \omega_1)$. If we draw $\omega_1, \ldots, \omega_n \sim \mathcal{P}$, then the data available for semisupervised learning comprise the dissimilarity matrix, $\Delta = \left[\delta\left(\omega_i, \omega_j\right)\right]$, plus whatever labels of the ω_i may be known. The challenge is to construct a classifer from these data.

There is an extensive literature on semisupervised learning, although most researchers have studied experiments in which $\omega \in \Omega$ is observed directly. The importance of semisupervised learning derives from the fact that it is often much easier to observe the features of objects (or the dissimilarities between objects) than it is to acquire class labels for objects. When this is the case, as it often is in image recognition and classification, natural language processing, hypertext categorization, remote sensing, etc., a data set may comprise many unlabeled outcomes and relatively few labeled outcomes. Rather than discard the unlabeled outcomes during training, semisupervised learning methods attempt to extract information from the entire data set.

Comprehensive surveys of semisupervised learning include technical reports by Seeger (2000) and Zhu (2006), and a forthcoming book by Chapelle et al. (2006). Popular approaches to semisupervised learning include imputation and co-training. The former approach uses the EM algorithm (Dempster et al., 1977) to impute missing labels during training, as in McCallum et al. (2000). The latter approach, proposed by Blum and Mitchell (1998), assumes that two distinct "views" of an object can be distinguished, e.g., words occurring on a web page and words occurring in hyperlinks that point to that page. Separate classifiers are trained on each view, then used to enlarge the training set of the other. Perhaps unsurprisingly, semisupervised learning methods tend to outperform traditional supervised learning methods that ignore unlabeled observations.

2. Embedding

Suppose that the representation space is $\Re^d$. Embedding an $n \times n$ dissimilarity matrix $\Delta = [\delta_{ij}]$ in $\Re^d$ means constructing a configuration of points, $x_1, \ldots, x_n \in \Re^d$, in such a way that the interpoint distances, $\|x_i - x_j\|$, approximate the dissimilarities, δ_{ij}. In psychometrics and statistics, techniques for embedding are called *multidimensional scaling* (MDS). The vast literature on MDS includes monographs, e.g., Cox and Cox (1994), Borg and Groenen (1997) and Everitt and Rabe-Hesketh (1997); expositions in multivariate statistics texts, e.g., Mardia et al. (1979, Chapter 14), Seber (1984, Section 5.5), Everitt and Dunn (1991, Chapter 5), and Krzanowski and Marriott (1994, Chapter 5); and surveys, e.g., Kruskal (1977), Carroll and Arabie (1980), de Leeuw and Heiser (1982), Trosset (1997), and Carroll and Arabie (1998).

Standard MDS methods produce finite configurations of points, not maps between spaces. However, the case for semisupervised learning is clarified by imagining an embedding methodology that does produce a map from Ω to $\Re^d$. Let $\mathcal{M} = \{M_\theta : \theta \in \Theta\}$ denote a parametric family of possible embedding maps. (The question of how to choose an appropriate and tractable family is crucial, but does not concern us here.) Let

$$T(\theta; P) = \int_\Omega \int_\Omega \left[\|M_\theta(\omega_1) - M_\theta(\omega_2)\| - \delta(\omega_1, \omega_2)\right]^2 P(\mathrm{d}\omega_1) P(\mathrm{d}\omega_1),$$

an error criterion inspired by the raw stress criterion. The optimal embedding map chooses $\theta^*(P)$ to minimize $T(\theta; P)$. If we could observe $\omega_1, \ldots, \omega_n \sim P$, then we could form the empirical distribution, $\hat{P}_n$, and estimate $\theta^*(P)$ by minimizing $T(\theta; \hat{P}_n)$. Then, under suitable regularity conditions, it should be the case that $\theta^*(\hat{P}_n) \to \theta^*(P)$ as $n \to \infty$.

The *parametric embedding* techniques described above are currently under development and are not yet available for analyzing dissimilarity data. Nevertheless, the case for semisupervised learning is the argument that one can construct a better representation space by using all of the dissimilarities instead of only the dissimilarities between labeled subjects. One must exercise considerable caution in making this case. To whatever extent an embedding is optimal, it is optimal only in the sense of how faithfully $x_1, \ldots, x_n \in \Re^d$ approximate $\omega_1, \ldots, \omega_n \in \Omega$—there is no guarantee that the best representation of Ω is the representation that is most useful for classification. This is

precisely the distinction, to which we now turn, between principal components and discriminant coordinates: the first d principal components are optimal for summarizing the data, but they may be very different from the first d discriminant coordinates, which are optimal for classifying the data.

For simplicity – and for other reasons that will become apparent – we will rely on classical multidimensional scaling (CMDS) to embed Δ in $\Re^d$. CMDS was introduced by Torgerson (1952) and subsequently analyzed by Gower (1966), who noted its intimate relation to principal component analysis (PCA).

2.1. Principal component analysis

We briefly review PCA. Let Y denote an $n \times q$ data matrix, in which each row corresponds to a subject and each column corresponds to a measurement variable. The subject profiles, i.e., the n rows of Y, are n points in $\Re^q$; the variable profiles, i.e., the q columns of Y, are q points in $\Re^n$. To center the data, i.e., to translate the subject profiles so that their mean lies at the origin of $\Re^q$, let $e = (1, \ldots, 1)^t \in \Re^n$, let I denote the $n \times n$ identity matrix, and let $P = I - ee^t/n$. Then P is a projection matrix and the centered data matrix is $\tilde{Y} = PY$. The column sums of $\tilde{Y}$ vanish; hence, after centering, the variable profiles lie in the $(n-1)$-dimensional subspace $e^{\perp} \subset \Re^n$.

Consider two matrices of inner products:

(1) $\tilde{Y}^t\tilde{Y}$ is the variable inner product matrix, i.e., the $q \times q$ matrix of inner products between columns of $\tilde{Y}$. This matrix is sometimes called the total sum-of-squares matrix. Upon dividing it by $n-1$ or (depending on the author) n, one obtains the sample covariance matrix.
(2) $\tilde{Y}\tilde{Y}^t$ is the subject inner product matrix, i.e., the $n \times n$ matrix of inner products between rows of $\tilde{Y}$.

It is well-known but, perhaps, insufficiently emphasized that the principal component representation of Y can be extracted from the subject inner product matrix.

Let

$$\tilde{Y} = U \left[\begin{array}{c|c} \Sigma & 0 \\ \hline 0 & 0 \end{array}\right] V^t$$

denote the singular value decomposition of the centered data matrix, where $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$ and $\sigma_1 \geq \cdots \geq \sigma_r > 0$ are the singular values of $\tilde{Y}$. Let $\Sigma_d = \mathrm{diag}(\sigma_1, \ldots, \sigma_d)$ and let $U = [U_d|\cdot]$. Then

$$B = \tilde{Y}\tilde{Y}^t = U \left[\begin{array}{c|c} \Sigma^2 & 0 \\ \hline 0 & 0 \end{array}\right] U^t$$

is the spectral decomposition of the subject inner product matrix and the inner product matrix

$$\bar{B} = [U_d|\cdot] \left[\begin{array}{c|c} \Sigma_d^2 & 0 \\ \hline 0 & 0 \end{array}\right] \left[\begin{array}{c} U_d^t \\ \hline \cdot \end{array}\right] = [U_d\,\Sigma_d|0]\,[U_d\,\Sigma_d|0]^t$$

is the best (in the sense of Frobenius norm, i.e., squared error) rank-d approximation of B. The rows of the $n \times d$ data matrix $U_d\Sigma_d$ are the d-dimensional principal component scores; thus, the d-dimensional principal component representation of the data can be extracted from the pairwise inner products between the subjects.

2.2. Classical multidimensional scaling

Having noted that the principal component representation of Y can be extracted from the subject inner product matrix $\tilde{Y}\tilde{Y}^t$, CMDS is easily described. Let $D(Y)$ denote the $n \times n$ matrix of pairwise Euclidean distances between the rows of Y, let $D_2(Y)$ denote the corresponding matrix of squared Euclidean distances, and define a linear transformation τ by $\tau(A) = -PAP/2$. It is easily checked that $\tau(D_2(Y)) = \tilde{Y}\tilde{Y}^t$, i.e., τ converts squared Euclidean distances to Euclidean inner products.

Now let Δ denote the $n \times n$ matrix of pairwise dissimilarities and let Δ_2 denote the corresponding matrix of squared dissimilarities. If the dissimilarities are Euclidean distances, then $\tau(\Delta_2)$ is an inner product matrix B. Regarding B as the subject inner product matrix, CMDS extracts the principal component representation. More generally, when $\tau(\Delta_2)$ is not an inner product matrix, CMDS approximates $\tau(\Delta_2)$ with the nearest inner product matrix of rank d, say $\bar{B}$, treats $\bar{B}$ as the subject inner product matrix, and extracts the principal component representation from $\bar{B}$.

3. Learning

Our concern is with supervised learning from dissimilarity data. Presumably, therefore, we prefer discriminant coordinates to principal components. Following Trosset (2004), we consider how to compute inner products of points represented in discriminant coordinates. If these inner products can be derived from the pairwise Euclidean distances between the subject profiles, then we can extend the construction of discriminant coordinates to the case of dissimilarity data.

3.1. Discriminant coordinates

Let X_i denote an $n_i \times d$ data matrix for class i and write

$$X^t = \left(X_1^t \mid \cdots \mid X_k^t\right).$$

Let $\bar{x} \in \Re^d$ denote the grand mean of all subject profiles, let $\bar{x}_i \in \Re^d$ denote the group mean for class i, and let $\tilde{X}_i$ denote X_i centered at $\bar{x}_i$. Then two $d \times d$ matrices are crucial for LDA:

$$W = \sum_{i=1}^{k} \tilde{X}_i^t \tilde{X}_i,$$

the pooled within-group sum-of-squares matrix, and

$$B = \sum_{i=1}^{k} n_i \, (\bar{x}_i - \bar{x}) \, (\bar{x}_i - \bar{x})^t \,,$$

the between-group sum-of-squares matrix.

LDA attempts to identify directions in which between-group variation is large relative to within-group variation. Assuming that W is invertible, we form $W^{-1}B$ and compute its spectral decomposition,

$$W^{-1}B = Q \Lambda Q^t.$$

For $p \leq \operatorname{rank}(W^{-1}B) \leq k-1$, let $q_1, \ldots, q_p$ denote the first p eigenvectors of $W^{-1}B$. These eigenvectors are the desired directions of maximal class separation. Some authors refer to them as *canonical variates*, but this terminology is not specific to discrimination. Following Gnanadesikan (1977) and Seber (1984), we prefer the more descriptive phrase *discriminant coordinates*.

Let $Q_p = [q_1 \cdots q_p]$. The representation of X in discriminant coordinates is

$$Z = X Q_p,$$

with inner products

$$z_i^t z_j = x_i^t Q_p Q_p^t x_j.$$

Because $Q_p^t W Q_p = I_p$, the $d \times d$ matrix $Q_p Q_p^t$ is a low-rank approximation of W^{-1}, so the Euclidean inner product in discriminant coordinates approximates the Mahalanobis inner product in the original coordinates. We would like to estimate these inner products directly from dissimilarity data, then extract the corresponding principal component representation to obtain Z.

To construct Z from Δ, we see no alternative to estimating X and Q_p separately. But estimating X from Δ *is* MDS. Furthermore, estimating Q_p from Δ necessarily entails CMDS. To understand why, consider that $\tau(\Delta_2) \approx \tilde{X}\tilde{X}^t$. We need covariance information to compute the Mahalanobis inner product, i.e., we need $\tilde{X}^t\tilde{X}$. The only way to pass from $\tilde{X}\tilde{X}^t$ to $\tilde{X}^t\tilde{X}$ is to factor the former matrix, explicitly identify X, then reverse the order of multiplication—and factoring an inner product matrix *is* CMDS.

The preceding arguments led Trosset (2004) to endorse a two-stage methodology: first embed (**M**) by CMDS, then label (**L**) by LDA. Previously, Anderson and Robinson (2003) studied the same two-stage methodology, describing multi-response permutation test statistics (for testing differences between groups) and deriving their asymptotic permutation distributions. The authors of both papers considered this methodology in the case of completely labeled

Table 1
Effect of embedding Δ in $\mathfrak{R}^d$ for $d = 1, \ldots, 17$

Principal component	λ_i	Percent variation	Cumulative percent	F_1	F_2
1	45.50	44.19	44.19	1.12	NA.
2	24.95	24.23	68.42	2.21	0.01
3	10.57	10.27	78.69	2.23	0.73
4	5.18	5.03	83.72	4.08	0.75
5	3.98	3.87	87.58	4.29	0.78
6	3.32	3.22	90.81	4.31	0.87
7	2.60	2.52	93.33	4.79	1.35
8	2.12	2.06	95.39	7.20	1.54
9	1.91	1.85	97.24	9.49	1.55
10	1.17	1.13	98.37	10.56	7.64
11	0.67	0.65	99.02	11.88	8.02
12	0.41	0.40	99.42	22.68	11.77
13	0.33	0.32	99.74	23.42	15.24
14	0.13	0.13	99.87	31.88	19.60
15	0.11	0.11	99.98	32.34	20.76
16	0.02	0.02	100.00	34.32	21.98
17	0.00	0.00	100.00	34.32	21.98

The λ_i are the positive eigenvalues of $\tau(\Delta_2)$; these quantities are proportional to the sample variances in the principal component directions. The quantities F_1 and F_2 are the F-ratios from a univariate analysis of variance with respect to the first and second discriminant coordinates. For comparison, note that the 0.95 quantile of an F distribution with 2 and 27 degrees of freedom is 3.35.

objects, i.e., as a fully supervised procedure. However, training $\mathbf{M}$ is unsupervised and training $\mathbf{L}$ is supervised; hence, training $\mathbf{L} \circ \mathbf{M}$ is in fact semisupervised. This insight creates additional possibilities: unlabeled objects that cannot be used when training $\mathbf{L}$ can be exploited when training $\mathbf{M}$.

3.2. Model selection

Decoupling the activities of embedding and classifying permits a semisupervised approach to learning from dissimilarity data, but it also poses a critical model selection problem. Given Δ, the embedding stage constructs $x_1, \ldots, x_n \in \mathfrak{R}^d$, which are then subjected to LDA. How should one choose the number of dimensions, d, in which to embed?

When MDS is used to display dissimilarity data in $\mathfrak{R}^d$, d is inevitably chosen to facilitate easy visualization, e.g., $d = 2$ or $d = 3$. More dimensions permit more faithful representation; however, in the case of what Torgerson (1952) called "fallible data," this may result in the faithful representation of noise. For example, if the dissimilarities are fallible measurements of a molecule's interatomic distances, then we would prefer $d = 3$ to $d > 3$ because the structure that we are attempting to reconstruct is 3-dimensional.

When embedding precedes classification, the demands of visualization might dictate $p = 2$ or $p = 3$ discriminant coordinates, but there is no *a priori* reason to restrict the intermediate step of embedding Δ in $\mathfrak{R}^d$ to $d = p$. Indeed, if the crucial directions for discriminating the classes are not directions of large variation, then it may be necessary to choose a large value of d in order to capture crucial information. But simply choosing d large is fraught with peril, as illustrated by the following example.

Pseudorandom samples of $n_i = 10$ subject profiles were drawn from each of $k = 3$ distributions, resulting in an $n \times n$ dissimilarity matrix, Δ, of $n = 30$ labeled subjects. We defer explaining precisely how Δ was obtained, instead proceeding directly to the question of how to embed Δ in $p = 2$ discriminant coordinates.

The first step is a spectral analysis of the 30×30 matrix $\tau(\Delta_2)$. We found 17 positive eigenvalues, displayed in the second column of Table 1. Because negative eigenvalues correspond to non-Euclidean structure in Δ, the best (in the sense defined by CMDS) Euclidean representation of Δ is obtained in $\mathfrak{R}^{17}$. We must choose $\mathfrak{R}^d$ with $d \in \{1, \ldots, 17\}$. Of course, $d \geq 2$ is necessary if we are to compute $p = 2$ discriminant coordinates.

It is not clear how to proceed. If we only desired a faithful representation of Δ, then we might choose d large enough to account for a comfortable proportion of the total variation that can be represented in Euclidean space. For example, $d = 3$ allows us to represent 78.69% of the total Euclidean variation, while $d = 6$ allows us to represent

90.81% and $d = 9$ allows 97.24%. However, by not representing all of the variation, one risks suppressing crucial information needed to discriminate the classes.

To measure the degree of class separation, we computed univariate F ratios with respect to each discriminant coordinate, obtaining the quantities in the last two columns of Table 1. (To facilitate interpretation, note that the 0.95 quantile of an F distribution with 2 and 27 degrees of freedom is 3.354.) These results suggest that $d = 3$ is too small for effective discrimination.

Both F_1 and F_2 increase with d. What does this mean? On the one hand, it might mean that information that is essential for discrimination is not captured by the directions of greatest variation. Note the substantial increase in F_2 as one passes from $d = 9$ to $d = 10$, as well as the substantial increase in F_1 as one passes $d = 11$ to $d = 12$. On the other hand, one should consider that nominal discrimination becomes intrinsically easier as d increases. If $d \geq n - 1$, then LDA will discriminate perfectly, regardless of class structure, which is why LDA with large numbers of variables is discouraged.

Indeed, for the present example, the 30 subject profiles were drawn from one bivariate normal distribution with covariance matrix I_2. Euclidean interpoint distances d_{ij} were computed, then contaminated by drawing from the error model

$$\delta_{ij} = \exp\left(\log\left(d_{ij}\right) + \epsilon_{ij}\right),$$

where $\epsilon_{ij} \sim \text{Normal}(0, 0.1)$. The population structure that underlies these fallible dissimilarities is 2-dimensional; choosing $d > 2$ only succeeds in representing noise with greater fidelity. Separate classes do not exist. Our apparent ability to better separate the designated classes by embedding Δ in more dimensions is an illusion.

When constructing discriminant coordinates from dissimilarity data, the dimension, d, in which Δ is initially embedded should be regarded as a smoothing parameter for the subsequent LDA. As with any smoothing parameter, the choice of d involves a tradeoff between underfitting and overfitting. The antidote is the same: cross-validation, or some other procedure for balancing model fit and model complexity. Anderson and Robinson (2003, Section 2.4) proposed three ways to choose d, including leave-one-out cross-validation.

3.3. Out-of-sample classification

Inherent in cross-validation are the concepts of training data, from which a classifier is constructed, and test data, to which the classifier is applied. In traditional methods for supervised learning, the distinction between training and test data is unambiguous. Only labeled subjects can be used to construct the classifier. The labels of the test subjects are withheld, to be compared to the predictions made by the classifier. Because the test subjects thus assume a temporary identity as unlabeled subjects, it is not possible to use the test subjects to augment the training subjects in constructing the classifier that will then be used to classify the test subjects. In contrast, the semisupervised methods that we have described admit the possibility of using unlabeled subjects to facilitate construction of the Euclidean representation in which classification is then performed. In this setting, when a set of subjects (including possibly some unlabeled subjects) is augmented with additional unlabeled subjects that require classification, there are two ways to proceed:

(1) The exclusive approach to out-of-sample classification:

Use the additional dissimilarities to embed (e.g., by CMDS) the out-of-sample subjects in the Euclidean representation of the original subjects. Use the classifier that was trained on the original subjects (e.g., LDA) to label the out-of-sample subjects.

The exclusive approach maintains a fixed representation of the original subjects. Thus, the nominal classification of the original subjects is not affected by the introduction of additional subjects. However, this approach does not exploit whatever additional information the out-of-sample subjects might provide about how to construct a Euclidean representation of the data.

The exclusive approach also entails a technical difficulty, viz., how to embed out-of-sample subjects in a previously constructed configuration. For CMDS, this *out-of-sample embedding problem* is nontrivial. In an accompanying paper (Trosset and Priebe, 2008), we demonstrate how to solve the out-of-sample problem for CMDS by solving an unconstrained nonlinear least-squares problem. The objective function is a fourth-order polynomial, easily minimized by standard gradient-based methods for numerical optimization. In our experience to date, nonglobal minimizers have not been a problem.

(2) The inclusive approach to out-of-sample classification:

Use the entire set of dissimilarities to re-embed (e.g., by CMDS) all of the subjects in a new Euclidean representation. Using the new representation, train a new classifier (e.g., LDA) on the original subjects, then use it to label the out-of-sample subjects.

The inclusive approach exploits additional information provided by the out-of-sample subjects to construct a new Euclidean representation. As a result, the original subjects are re-configured and the classifier changes, as may the nominal classification of the original points.

The inclusive approach repeats the entire semisupervised analysis each time that additional data is collected. Whether or not such repetition is more expensive than inserting additional subjects into an existing configuration will depend on the numbers of subjects involved.

3.4. Sphericity

We have worried about the possibility that the principal component directions of the entire data set may be very different from the directions that most effectively discriminate classes. There is one situation in which these two sets of directions are necessarily aligned, hence in which the semisupervised approach is guaranteed to be asymptotically superior to the fully supervised approach.

Theorem 1. *Suppose that $\Omega = \Re^q$ and that P is given by (1), with class means μ_i and spherical class covariance matrices $\sigma_i^2 I$. Then population discriminant coordinate i coincides with population principal component i.*

Proof. Let E denote expectation with respect to P and let E_i denote expectation with respect to P_i. Then

$$E(y) = \sum_{i=1}^{k} \alpha_i \mu_i = \mu$$

and

$$\begin{aligned}
\mathrm{Cov}(y) &= E\left[(y-\mu)(y-\mu)^t\right] \\
&= \sum_{i=1}^{k} \alpha_i E_i\left[(y-\mu_i+\mu_i-\mu)(y-\mu_i+\mu_i-\mu)^t\right] \\
&= \sum_{i=1}^{k} \alpha_i \left[\sigma_i^2 I + (\mu_i-\mu)(\mu_i-\mu)^t\right] \\
&= \left(\sum_{i=1}^{k} \alpha_i \sigma_i^2\right) I + \sum_{i=1}^{k} \alpha_i (\mu_i-\mu)(\mu_i-\mu)^t \\
&= W + B.
\end{aligned}$$

The population principal components are the eigenvectors of $W + B$, which are also the eigenvectors of B because $W = t^2 I$, and the population discriminant coordinates are the eigenvectors of $W^{-1}B = B/t^2$. □

Despite not observing $y_1, \ldots, y_n \sim P$, the assumption of spherical covariances can be tested using $\delta_{ij} = \|y_i - y_j\|$. Let $\Delta(i)$ denote the within-class dissimilarity matrices for the labeled subjects. The likelihood ratio statistic for testing the sphericity of a covariance matrix is a function of the eigenvalues of the sample covariance matrix Seber (1984, Section 3.5.4), and these eigenvalues are also the eigenvalues of $\tau(\Delta_2(i))$. Unfortunately, this procedure does extend transparently to the case of fallible dissimilarities. In the example discussed in Section 3.2, the population covariance matrix was spherical, but the noise that contaminated the dissimilarities introduced spurious dimensions with apparently smaller variation. To test sphericity in this case, one must first discard spurious dimensions.

For the data displayed in Table 1, it was not clear which dimensions were spurious. These data were generated by drawing $n = 30$ subject profiles from a single bivariate population with spherical covariance, then contaminating the Euclidean interpoint distances. If we instead draw $n = 300$ subject profiles, then it becomes considerably easier to

distinguish the spurious dimensions. For example, we performed this experiment and observed that $\tau(\Delta_2)$ had 151 positive eigenvalues, the largest 20 of which were as follows:

299.80	286.33	26.00	24.11	23.89	23.64	22.87	21.3	20.97	20.53
19.37	18.95	18.38	17.76	17.23	16.97	16.74	16.6	16.34	15.97

With these data, it does not require much imagination to guess that the underlying population is 2-dimensional with spherical covariance.

4. Example 1: Simulated dissimilarity data

To illustrate the concepts described in the preceding sections, we performed a simple simulation experiment. We began by drawing 200 objects from a bivariate normal distribution with population mean vector $\mu_0 = (0, 0)^t$ and another 200 objects from a bivariate normal distribution with population mean vector $\mu_0 = (1, 0)^t$. Both distributions had population covariance matrix

$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 25 \end{bmatrix}.$$

From these 400 objects, we computed pairwise Euclidean distances, then multiplied each distance by $\exp(\epsilon_{ij})$, where each $\epsilon_{ij} \sim \text{Normal}(0, 0.005)$, to obtain a 400×400 dissimilarity matrix, Δ. The challenge was to discriminate between the populations using Δ, together with labels for a small subset of the 400 objects. Notice that the populations were chosen so that the *second* principal component is the direction that best discriminates the populations.

For each of 100 replications of the experiment, we chose simple random samples of 10 objects from each sample of 200 objects. These 20 objects were labeled. We then performed four analyses:

(1) Supervised Learning in $\Re^2$. Using only the pairwise dissimilarities between the 20 labeled objects, we used CMDS to construct a configuration of 20 labeled points in $\Re^2$. LDA was applied to these points and the number of nominal (resubstitution) misclassification errors was recorded.
(2) Semisupervised Learning in $\Re^2$. Using the pairwise dissimilarities between all 400 objects, we used CMDS to construct a configuration of 400 points in $\Re^2$. LDA was applied to the 20 labelled points in this configuration and the number of nominal misclassification errors was recorded.
(3) Supervised Learning in $\Re^1$. Using only the pairwise dissimilarities between the 20 labeled objects, we used CMDS to construct a configuration of 20 labeled points in $\Re^1$. LDA was applied to these points and the number of nominal misclassification errors was recorded.
(4) Semisupervised Learning in $\Re^1$. Using the pairwise dissimilarities between all 400 objects, we used CMDS to construct a configuration of 400 points in $\Re^1$. LDA was applied to the 20 labelled points in this configuration and the number of nominal misclassification errors was recorded.

Because $d = 2$ dimensions suffice to discriminate the populations, we would expect semisupervised learning to outperform supervised learning in the 2-dimensional case. Indeed, the semisupervised approach resulted in fewer nominal errors than the supervised approach in 75 of 100 replications. (The supervised approach had fewer errors in 17 replications; the approaches tied in the remaining 8 replications.) On average, the semisupervised approach resulted in 2.08 fewer nominal errors per replication than did the supervised approach.

Because the first principal component is orthogonal to the discriminant coordinate, we would not expect semisupervised learning to outperform supervised learning in the 1-dimensional case. In this example, the two approaches performed almost identically, the semisupervised approach resulting in an average of 0.10 more nominal errors per replication than the supervised approach.

5. Example 2: Hippocampal dissimilarity data

Finally, we apply our methods to the problem of distinguishing patients with Alzheimer's disease (AD) from normal elderly subjects on the basis of how the shapes of their hippocampi differ. Here, dissimilarities are obtained by (1) scanning individual whole brain structure using high-resolution T1-weighted structural MRI (magnetic resonance imaging), (2) segmenting the scans using FreeSurfer (see http://surfer.nmr.mgh.harvard.edu/fswiki for documentation

and citations), and (3) measuring asymmetric pairwise dissimilarity by large deformation diffeomorphic metric mapping (LDDMM), as described by Beg et al. (2005). After symmetrizing, the methods described herein apply. Because of the extensive preprocessing necessary to obtain the dissimilarities, learning strategies that operate directly on feature vectors are not appropriate.

To illustrate our methods, we analyze data obtained from the left and right hippocampi of 38 Alzheimer's patients and 57 normal elderly control subjects through the Biomedical Informatics Research Network (http://www.rbirn.net). This is a subset of the data analyzed by Miller et al. (submitted for publication), in which 6 patients with semantic dementia were also considered. Step (1) was performed at Washington University, step (2) at the Martinos Center at Massachusetts General Hospital, and step (3) at the Center for Imaging Science at Johns Hopkins University. Step (2) was performed at two different times, first for a training set of $n = 39$ subjects (18 AD patients and 21 normal elderly control subjects) and subsequently for 56 additional test subjects. Complicating the task of classifying the test subjects, the segmentation methodology used in step (2) had been modified in the interim. Such a change in methodology could result in test data for which the training data are not representative, thereby complicating the task of classifier construction. We do not pursue that possibility here, as our present concern is with illustrating the semisupervised methodology described in the preceding sections.

Each set (left and right) of asymmetric dissimilarities computed by LDDMM was symmetrized by averaging, resulting in two 95×95 dissimilarity matrices, L and R. It is not clear how to combine the information contained in L and R. Various approaches are possible. Here, we embed L and R separately, then apply LDA to the product of these representations. Because there are only two classes, LDA reduces to Fisher's best linear discriminator.

First, suppose that we had decided *a priori* to construct a 2-dimensional representation by forming the product of the first principal components of the left and right embeddings. We treat the training subjects as labeled, the test subjects as unlabeled, and compare three procedures for classifying the test subjects:

(1) Fully supervised classification with individual out-of-sample embedding.

Let $L(\text{train})$ and $R(\text{train})$ denote the 39×39 dissimilarity matrices for the training subjects. Let $X_l(\text{train})$ denote the 39×1 configuration matrix that results from applying CMDS to $L(\text{train})$ and let $X_r(\text{train})$ denote the 39×1 configuration matrix that results from applying CMDS to $R(\text{train})$. Train a classifier by applying LDA to the 39×2 configuration matrix $X(\text{train}) = [X_l(\text{train})|X_r(\text{train})]$.

To apply the classifier so constructed, individually embed each of the 56 test subjects in $X(\text{train})$. This is accomplished by applying the technique described in Section 3 (the case of $k = 1$ out-of-sample point) of Trosset and Priebe (2008). After embedding, use the classifier constructed from the training subjects to classify the test subjects.

(2) Fully supervised classification with simultaneous out-of-sample embedding.

Construct the same configuration and classifier as in the preceding procedure, then simultaneously embed all 56 test subjects in $X(\text{train})$. This is accomplished by applying the technique described in Section 5 (the case of $k > 1$ out-of-sample points) of Trosset and Priebe (2008). After embedding, use the classifier constructed from the training subjects to classify the test subjects.

Note that this procedure exploits information about the relations between unlabeled subjects, not in its fully supervised construction of a classifier, but in its determination of the points to which that classifier will be applied. Depending on one's perspective, one might reasonably argue that this procedure is also a form of semisupervised learning.

(3) Semisupervised classification.

Let X_l denote the 95×1 configuration matrix that results from applying CMDS to L and let X_r denote the 95×1 configuration matrix that results from applying CMDS to R. Train a classifier by applying LDA to the 39 training subjects in the 95×2 configuration matrix $X = [X_l|X_r]$, then apply that classifier to the 56 test subjects in X.

The 2-dimensional configurations of 95 points constructed by each of the three preceding procedures are displayed in Figs. 1 and 3. The first two procedures begin by using CMDS to embed the 39 training subjects; hence, the coordinates of these subjects in Figs. 1 and 2 are identical and the same decision boundary is obtained by LDA. The nominal misclassification error rate of this classifier is $13/39 \doteq 0.33$ (6 AD patients in the training sample lie below the decision boundary, in the region identified with normal elderly subjects; 7 normal elderly subjects in the training lie above the decision boundary, in the region identified with AD patients), but this resubstitution estimate of

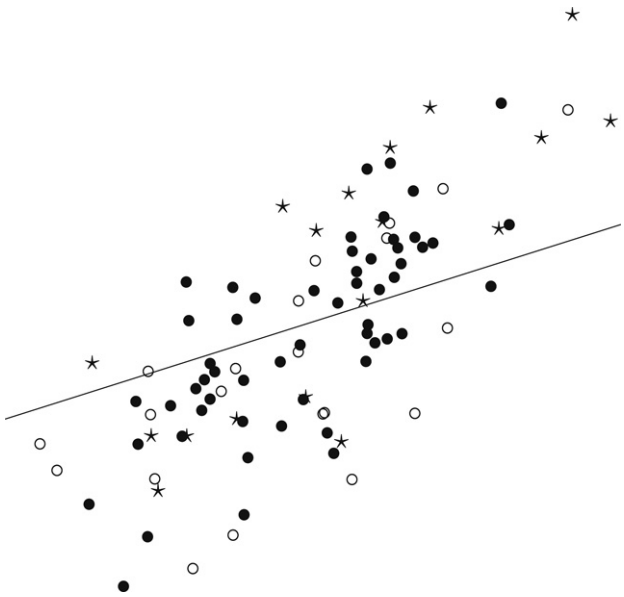

Fig. 1. A 2-dimensional Euclidean representation of 18 AD patients (⋆), 21 normal elderly control subjects (○), and 56 test subjects (●), constructed from dissimilarities in their hippocampal shapes. The horizontal coordinates were obtained from the 1-dimensional CMDS embedding of the left-side dissimilarities of the 39 training subjects; the vertical coordinates were obtained from the 1-dimensional CMDS embedding of the corresponding right-side dissimilarities. The 56 test subjects were then embedded individually, without trying to approximate dissimilarities between test subjects. Compare the individual out-of-sample embedding displayed here to the simultaneous out-of-sample embedding displayed in Fig. 2. The line is the decision boundary that corresponds to Fisher's best linear discriminator, inferred from the 39 training subjects.

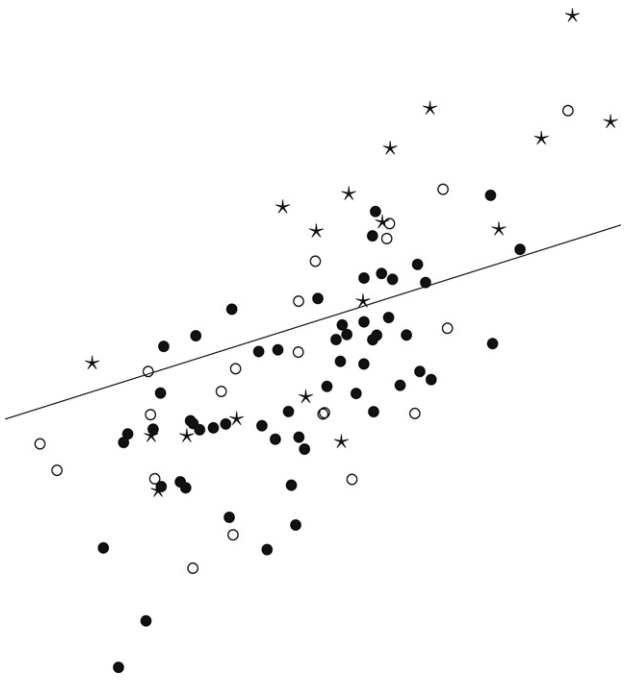

Fig. 2. A 2-dimensional Euclidean representation of 18 AD patients (⋆), 21 normal elderly control subjects (○), and 56 test subjects (●), constructed from dissimilarities in their hippocampal shapes. The horizontal coordinates were obtained from the 1-dimensional CMDS embedding of the left-side dissimilarities of the 39 training subjects; the vertical coordinates were obtained from the 1-dimensional CMDS embedding of the corresponding right-side dissimilarities. The 56 test subjects were then embedded simultaneously, in an attempt to approximate dissimilarities between pairs of test subjects. Compare the simultaneous out-of-sample embedding displayed here to the individual out-of-sample embedding displayed in Fig. 1. The line is the decision boundary that corresponds to Fisher's best linear discriminator, inferred from the 39 training subjects.

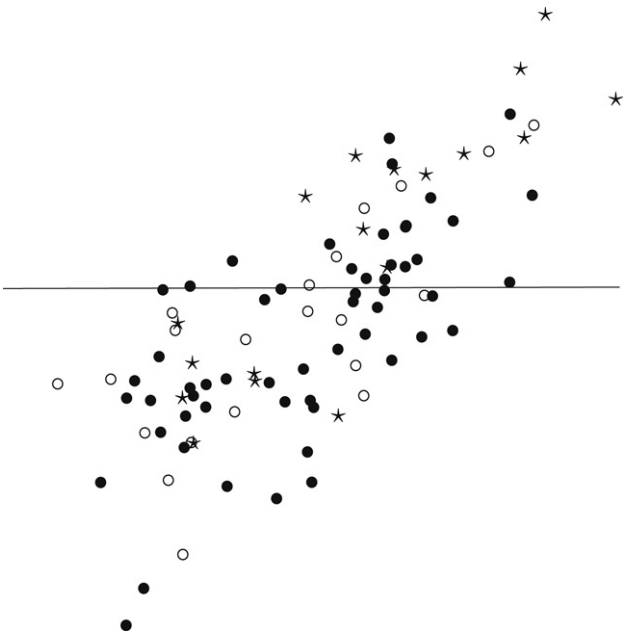

Fig. 3. A 2-dimensional Euclidean representation of 18 AD patients (⋆), 21 normal elderly control subjects (○), and 56 test subjects (●), constructed from dissimilarities in their hippocampal shapes. The horizontal coordinates were obtained from the 1-dimensional CMDS embedding of the left-side dissimilarities of all 95 subjects; the vertical coordinates were obtained from the 1-dimensional CMDS embedding of the corresponding right-side dissimilarities. The line is the decision boundary that corresponds to Fisher's best linear discriminator, inferred from the 39 training subjects.

the true probability of misclassification is too optimistic because it is computed from the same subjects whose labels were used to construct the classifier. Better estimates can be obtained by examining the 56 test subjects.

To classify the test subjects, one must first embed them in the configuration of training subjects. Figs. 1 and 2 display two different out-of-sample embeddings. In Fig. 1, each training subject was embedded individually, without trying to approximate dissimilarities between pairs of test subjects. The misclassification error rate for the test sample is $18/56 \doteq 0.32$.

In Fig. 2, the $k = 56$ test subjects were embedded simultaneously. In simultaneous out-of-sample embedding, the error criterion includes test–test dissimilarities as well as test–train dissimilarities. As a result, the positioning of the test subjects in Fig. 2 is different than in Fig. 1. The estimated misclassification error rate is $17/56 \doteq 0.30$.

The third procedure begins by using CMDS to embed all 95 subjects. As a result, the configuration of 39 training subjects in Fig. 3 differs from the configuration of training subjects in Figs. 1 and 2. This difference, in turn, results in a different decision boundary when LDA is applied to the same test subjects. This classifier also has a misclassification error rate of $13/39 \doteq 0.33$ for the training sample and $17/56 \doteq 0.30$ for the test sample.

What cannot be discerned from Figs. 2 and 3 is the extent of agreement between the second and third procedures. This information is summarized in Table 2. There were 6 test subjects (3 normal, 3 AD) that were classified as normal by the second procedure and as AD by the third procedure; the two procedures agreed on the other 50 test subjects. One also discerns that the normal test subjects were easier to classify than the AD test subjects. Only 5 of the 33 normal test subjects on which the second and third procedures agreed were misclassified, an error rate of 0.15. In contrast, 9 of the 17 AD test subjects on which the second and third procedures agreed were misclassified, an error rate of 0.53.

The illustrative 2-dimensional Euclidean representations displayed in Figs. 1 and 3 suggest considerable overlap in the hippocampal shapes of AD patients and normal elderly. Perhaps hippocampal shape is not systematically affected by AD, or perhaps these representations fail to capture critical information needed for better discrimination. We proceed to investigate the latter possibility.

Each of the representations displayed in Figs. 1 and 3 was constructed by forming the product of the first principal components from left and right embeddings. We now consider a variety of Euclidean representations, constructed by

Table 2
Performance of the classifiers in Figs. 2 and 3 on 56 test subjects, comprising 36 normal elderly control (NC) subjects and 20 AD patients

		NC Subjects		AD Patients	
		Procedure 3 prediction		Procedure 3 prediction	
		NC	AD	NC	AD
Procedure 2 prediction	NC	28	3	9	3
	AD	0	5	0	8

Procedure 2 is fully supervised classification with simultaneous out-of-sample embedding; Procedure 3 is semisupervised classification.

Table 3
Numbers of misclassification errors (of 56) for three classifiers constructed from 18 AD patients and 21 normal elderly control subjects

	0	1	2	3	4	5	6
0	−|−|−	20|36|17	20|36|22	20|36|13	20|36|17	19|36|13	20|36|16
1	23|21|21	18|17|17	18|17|20	18|17|12	18|17|19	17|16|14	16|22|17
2	21|22|30	$18\left\|\binom{20}{18}\right\|28$	19|23|28	17|17|17	$17\left\|\binom{23}{16}\right\|26$	$19\left\|\binom{16}{18}\right\|23$	18|21|25
3	22|23|32	21|20|26	21|21|27	19|22|17	19|26|27	19|17|25	18|24|26
4	22|23|25	22|21|22	23|22|25	19|23|16	19|26|22	20|17|21	18|25|22
5	23|23|24	19|16|22	18|15|24	21|17|15	21|18|21	21|15|19	19|22|22
6	23|23|25	19|17|19	18|18|22	21|17|16	20|20|22	20|15|21	20|21|20

Rows are labeled by d_l, columns by d_r. The classifiers were constructed by LDA in Euclidean representations with $d_l + d_r$ dimensions. The representations were constructed from dissimilarities in hippocampal shapes by forming the product of the d_l-dimensional CMDS embedding of left-side dissimilarities and the d_r-dimensional CMDS embedding of right-side dissimilarities. For procedure (1) and (2), CMDS was applied to the 39 training subjects, then the 56 out-of-sample test subjects were positioned in relation to the training subjects. The out-of-sample embedding was performed individually for (1), simultaneously for (2). For procedure 3, CMDS was applied to all 95 subjects. For each (d_l, d_r), the three entries are the respective numbers of errors for the three procedures. The test sample comprised 36 normal subjects and 20 AD patients; hence, classifying all test subjects as normal results in 20 misclassification errors. For $d_l = 2$ and $d_r = 1, 4, 5$, there appear to be two different globally optimal embeddings for procedure (2), resulting in two different numbers of errors.

forming products of the first d_l principal components of the left embedding and the first d_r principal components of the right embedding. Thus, in the case of semisupervised classification, if X_l is the $95 \times d_l$ configuration matrix that results from applying CMDS to L and X_r is the $95 \times d_r$ configuration matrix that results from applying CMDS to R, then $X = [X_l|X_r]$ is a $95 \times (d_l + d_r)$ configuration matrix and LDA is applied to the 39 training subjects in X.

We considered $d_l, d_r \in \{0, 1, \ldots, 6\}$, excepting the pair $(0, 0)$. As in the previous example, $(d_l, d_r) = (1, 1)$, we used each of the three procedures described above to classify each of the 56 test subjects. The resulting numbers of misclassification errors are reported in Table 3. Interpretation of these results is facilitated by considering that the test sample comprised 36 normal subjects and 20 AD patients. The expected number of misclassification errors that would result from randomly classifying each test subject as normal with probability $1/2$ and AD with probability $1/2$ is 28, and the probability of observing ≤ 20 errors is just 0.022. However, classifying all test subjects as normal would result in 20 misclassification errors and classifying all test subjects as AD would result in 36 errors. Hence, the expected number of misclassification errors that would result from randomly choosing to classify all test subjects as normal with probability $1/2$ and AD with probability $1/2$ is also 28, but the probability of observing 20 errors is 0.500. How one assesses the performance of a classifier that produces 20 misclassification errors depends on which of these models one uses for comparison.

The numbers of test subjects classified as AD are reported in Table 4. For most (d_l, d_r), each procedure classified some test subjects as normal and others as AD; hence, it seems reasonable to assess the performances of these

Table 4
Numbers of test subjects classified as AD by the procedures described in Table 3

	0	1	2	3	4	5	6
0	−|−|−	26|56|19	26|56|30	22|56|9	22|56|15	21|56|13	24|56|12
1	29|27|27	26|13|19	26|25|28	22|21|10	22|21|17	19|18|14	20|28|15
2	29|34|48	$28\left\|\binom{26}{4}\right\|38$	27|31|42	21|9|25	$21\left\|\binom{29}{12}\right\|36$	$21\left\|\binom{22}{6}\right\|33$	20|31|33
3	32|37|50	29|28|36	29|31|41	27|34|25	27|42|37	25|29|35	26|42|38
4	32|39|35	28|29|32	27|32|41	27|35|20	27|40|34	26|29|31	26|43|26
5	29|29|36	31|14|30	30|17|40	27|21|23	27|24|33	27|17|29	27|34|28
6	29|29|35	31|17|27	30|24|38	27|21|24	26|26|36	26|17|33	30|27|26

The test sample comprised 56 subjects, of whom 20 were AD patients. Notice that the two different simultaneous embeddings obtained for $d_l = 2$ and $d_r = 1, 4, 5$ resulted in very different classifiers.

Table 5
Performance of two classifiers on 56 test subjects, comprising 36 normal elderly control (NC) subjects and 20 AD patients

		NC Subjects		AD Patients	
		Procedure 3 prediction		Procedure 3 prediction	
		NC	AD	NC	AD
Procedure 2 prediction	NC	27	0	7	1
	AD	8	1	4	8

In contrast to Table 2, in which the subjects were represented in $d_l = 1$ plus $d_r = 1$ dimensions, here the subjects were represented in $d_l = 1$ plus $d_r = 3$ dimensions. Procedure 2 is fully supervised classification with simultaneous out-of-sample embedding; Procedure 3 is semisupervised classification.

procedures by comparing them to the null procedure that randomly classifies each test subject by a fair coin toss. The probability that coin-tossing would produce ≤ 12 misclassification errors (the best result reported in Table 3, by semisupervised classification for $d_l = 1$ and $d_r = 3$) is just 1.04×10^{-5}. Although it is not clear how to adjust this probability in light of 48 highly dependent analyses, it does provide some evidence that hippocampal shape is predictive of AD.

Table 4 also reveals that fully supervised classification with simultaneous out-of-sample embedding can behave quite differently than semisupervised classification. A dramatic example occurs when $d_l = 2$ and $d_r = 1$. In this case, the former procedure classified 4 test subjects as AD whereas the latter classified 38 test subjects as AD. Table 5 compares these procedures in the more interesting case of $d_l = 1$ and $d_r = 3$. In this case, the procedures disagreed on their labeling of 13 test subjects. For 12 of these 13 subjects, the fully supervised procedure diagnosed AD when the semisupervised procedure did not.

The general picture that emerges from our results is that, properly represented, some AD patients can be distinguished from normal elderly on the basis of hippocampal shape while others overlap normal elderly. This finding is entirely plausible, as AD is progressive. It would be interesting to investigate if the AD patients that could be distinguished from normal were suffering from a more advanced stage of the disease. See Miller et al. (submitted for publication) for further discussion of the effect of AD on hippocampal shape.

6. Discussion

We have described a two-stage approach to learning from dissimilarity data:

(1) Embed all objects (labeled and unlabeled) in a Euclidean space.
(2) Train a classifier on the labeled objects in the Euclidean representation.

The embedding stage is not supervised: information from unlabeled objects can be used to construct the representation in which classification will proceed. Thus, the entire procedure is properly regarded as semisupervised.

We have analyzed the case in which linear discriminant analysis is used in the classification stage. We argued in Section 3.1 that the use of LDA in the classification stage naturally invites the use of classical multidimensional scaling in the embedding stage; hence, our emphasis on CMDS followed by LDA. A significant challenge in implementing our methods is the choice of the Euclidean space in which to embed the objects. This is a model selection problem that involves the usual trade-off between underfitting (too few dimensions) and overfitting (too many dimensions).

Fully supervised approaches must construct a Euclidean representation using only labeled objects. The potential advantage of the semisupervised approach is its ability to exploit additional information in the embedding stage. Whether or not this additional information results in superior classifier performance depends on the extent to which the population principal component directions estimated in the embedding stage contain the population discriminant directions estimated in the classification stage. In the case of spherical covariances, these directions are perfectly aligned and the semisupervised approach is necessarily superior. The simulation study reported in Section 4 demonstrates that this superiority is not restricted to the case of sphericity, but the choice of dimension is critical.

To classify unlabeled objects using the fully supervised approach, it is necessary to insert these out-of-sample objects into the original Euclidean representation, then apply the original classifier to the newly embedded out-of-sample objects. Thus, the fully supervised approach necessitates solving an out-of-sample embedding problem. This can be accomplished by the methods described in Trosset and Priebe (2008), either by embedding each out-of-sample object individually or by embedding all out-of-sample objects simultaneously. In contrast to individual embedding, simultaneous embedding exploits information about the relations between the out-of-sample objects.

Our analysis of the pairwise dissimilarities of left and right hippocampal shapes for 95 human subjects demonstrates that the above procedures can produce dramatically different results. In particular, there may be more than one globally optimal out-of-sample embedding, so that the same fully supervised procedure may produce dramatically different classifiers.

Acknowledgments

This research was supported in part by grants from the Office of Naval Research. We are grateful to Wendy L. Martinez. Support for the JHU group was provided in part by the National Center for Research Resources P41-RR015241 and the NCRR BIRN Morphometry Project U24 RR021382.

References

Anderson, M.-J., Robinson, J., 2003. Generalized discriminant analysis based on distances. Australian & New Zealand Journal of Statistics 45, 301–318.

Beg, M.F., Miller, M.I., Trouvé, A., Younes, L., 2005. Computing large deformation metric mapping via geodesic flows of diffeomorphisms. International Journal of Computer Vision 61, 139–157.

Blum, A., Mitchell, T., 1998. Combining labeled and unlabeled data with co-training. In COLT: Proceedings of the Workshop on Computational Learning Theory, pp. 92–100. Morgan Kaufman.

Borg, I., Groenen, P., 1997. Modern Multidimensional Scaling: Theory and Applications. Springer-Verlag, New York.

Carroll, J.D., Arabie, P., 1980. Multidimensional scaling. Annual Review of Psychology 31, 607–649.

Carroll, J.D., Arabie, P., 1998. Multidimensional scaling. In: Measurement, Judgment, and Decision Making. Academic Press, (Chapter 3).

Chapelle, O., Schölkopf, B., Zien, A. (Eds.), 2006. Semi-Supervised Learning. MIT Press, Cambridge, MA.

Cox, T.F., Cox, M.A.A., 1994. Multidimensional Scaling. Chapman & Hall, London.

de Leeuw, J., Heiser, W., 1982. Theory of multidimensional scaling. In: Krishnaiah, P.R., Kanal, I.N. (Eds.), Handbook of Statistics, vol. 2. North-Holland Publishing Company, Amsterdam, pp. 285–316 (Chapter 13).

Dempster, A.P., Laird, N.M., Rubin, D.B., 1977. Maximum likelihood from incomplete data via the EM algorithm. Journal of the Royal Statistical Society, Series B 39, 1–38.

Everitt, B.S., Dunn, G., 1991. Applied Multivariate Data Analysis. Edward Arnold, London.

Everitt, B.S., Rabe-Hesketh, S., 1997. The Analysis of Proximity Data, vol. 4. Arnold, London, in Kendall's Library of Statistics.

Gnanadesikan, R., 1977. Methods for Statistical Data Analysis of Multivariate Observations. John Wiley & Sons, New York.

Gower, J.C., 1966. Some distance properties of latent root and vector methods in multivariate analysis. Biometrika 53, 325–338.

Kruskal, J.B., 1977. Multidimensional scaling and other methods for discovering structure. In: Enslein, K., Ralston, A., Wilf, H.S. (Eds.), Statistical Methods for Digital Computers, vol. 3. John Wiley & Sons, New York, pp. 296–339. Mathematical Methods for Digital Computers.

Krzanowski, W.J., Marriott, F.H.C., 1994. Multivariate Analysis Part 1 Distributions, Ordination and Inference. Edward Arnold, London, Kendall's Library of Statistics 1.

Mardia, K.V., Kent, J.T., Bibby, J.M., 1979. Multivariate Analysis. Academic Press, Orlando.

McCallum, A.K., Thrun, S., Mitchell, T., Nigam, K., 2000. Text classification from labeled and unlabeled documents using EM. Machine Learning 39, 103–134.

Miller, M.I., Priebe, C., Qiu, A., Kolasny, A., Brown, T., Park, Y., Ratnanather, J.T., Busa, E., Jovicich, J., Yu, P., Dickerson, B., Buckner, R.L., Morphometry BIRN, 2008. Collaborative computational anatomy: an MRI morphometry study of the human brain via diffeomorphic metric mapping. Human Brain Mapping (submitted for publication).

Seber, G.A.F., 1984. Multivariate Observations. John Wiley & Sons, New York.

Seeger, M., 2000. Learning with labeled and unlabeled data. Technical Report, Institute for Adaptive and Neural Computation, University of Edinburgh.

Torgerson, W.S., 1952. Multidimensional scaling: I. Theory and method. Psychometrika 17, 401–419.

Trosset, M.W., 1997. Numerical algorithms for multidimensional scaling. In: Klar, R., Opitz, P. (Eds.), Classification and Knowledge Organization. Springer, Berlin, pp. 80–92. Proceedings of the 20th Annual Conference of the Gesellschaft für Klassifikation e.V., held March 6–8, 1996, in Freiburg, Germany.

Trosset, M.W., 2004. On the construction of discriminant coordinates from dissimilarity data. Computing Science and Statistics 35, Distributed on CD-ROM.

Trosset, M.W., Priebe, C.E., 2008. The out-of-sample problem for classical multidimensional scaling. Computational Statistics and Data Analysis.

Zhu, X., 2006. Semi-supervised learning literature survey. Technical Report 1530. Department of Computer Science, University of Wisconsin—Madison.