

CPM: A Covariance-preserving Projection Method

Jieping Ye*

Tao Xiong[†]

Ravi Janardan[‡]

Abstract

Dimension reduction is critical in many areas of data mining and machine learning. In this paper, a Covariance-preserving Projection Method (CPM for short) is proposed for dimension reduction. CPM maximizes the class discrimination and also preserves approximately the class covariance. The optimization involved in CPM can be formulated as low rank approximations of a collection of matrices, which can be solved iteratively. Our theoretical and empirical analysis reveals the relationship between CPM and Linear Discriminant Analysis (LDA), Sliced Average Variance Estimator (SAVE), and Heteroscedastic Discriminant Analysis (HDA). This gives us new insights into the nature of these different algorithms. We use both synthetic and real-world datasets to evaluate the effectiveness of the proposed algorithm.

keywords: Dimension reduction, linear discriminant analysis, heteroscedastic discriminant analysis, covariance.

1 Introduction

Linear Discriminant Analysis (LDA) [4, 7, 8] is a well-known scheme for feature extraction and dimension reduction. LDA projects the data onto a lower-dimensional vector space such that the ratio of the between-class distance to the within-class distance is maximized, thus achieving maximum discrimination. LDA is equivalent to maximum likelihood classification assuming normal distribution for each class with a common covariance matrix. It has been applied successfully to many areas, such as computer vision, bioinformatics, etc. [1, 6, 11, 15] However, LDA has several limitations:

- (1) It fails to recover the features for classification, when all class centroids coincide;

- (2) it may not find the best projection, when class covariance matrices vary; and
- (3) the reduced dimension of LDA is no larger than $k - 1$ (k denotes the number of classes), which may not be sufficient for complex data.

Generalization of LDA by fitting Gaussian mixtures to each class has been studied by Hastie [10]. Cook *et al.* [2, 3] proposed the Sliced Average Variance Estimator (SAVE), which is shown to be capable of dealing with the limitations in LDA. Zhu and Hastie [16] developed a general method for finding important discriminant directions without assuming the class densities belong to any particular parametric family. Kumar and Andreou [13] proposed HDA, based on a different model, in which the classes are still Gaussian, yet are allowed to have different covariance matrices, under the condition that both centroids and covariance matrices coincide in a subspace of the observation space.

In this paper, we propose a new algorithm for dimension reduction, called CPM (which stands for Covariance-preserving Projection Method). CPM aims to maximize the class discrimination and at the same time preserve approximately the class covariance by applying a tuning parameter α between 0 and 1. One key feature of CPM is that the critical information on the class covariance is preserved under the projection. With a properly chosen tuning parameter α , which may be dependent on the data distribution, the CPM algorithm is able to deal with difficult situations encountered in LDA. In practice, the best value of α can be estimated by cross-validation.

To illustrate the difference between CPM and LDA, we generated a synthetic dataset with 20 dimensions and 3 classes as in [16]. Fig 1 (top) shows the first two coordinates. For the first two coordinates, class 1 is simulated from a standard multivariate Gaussian, while classes 2 and 3 are mixtures of two symmetrically shifted standard Gaussians. In the remaining 18 coordinates, Gaussian noise with zero mean and standard deviation 1 is used for all three classes. It is clear from Fig 1 (middle) that LDA fails to extract important features because the class centroids coincide. CPM separates the three classes completely as shown in Fig 1 (bottom), which shows the advantage of incorporating the class

*Department of Computer Science and Engineering, Arizona State University

[†]Department of Electrical and Computer Engineering, University of Minnesota

[‡]Department of Computer Science and Engineering, University of Minnesota

