Fluid Dynamics Models for Low Rank Discriminant Analysis

Yung-Kyun Noh[1,2] **Byoung-Tak Zhang**[2] **Daniel D. Lee**[1]
[1]GRASP Lab, University of Pennsylvania, Philadelphia, PA 19104, USA
[2]Biointelligence Lab, Seoul National University, Seoul 151-742, Korea
nohyung@seas.upenn.edu, btzhang@bi.snu.ac.kr, ddlee@seas.upenn.edu

Abstract

We consider the problem of reducing the dimensionality of labeled data for classification. Unfortunately, the optimal approach of finding the low-dimensional projection with minimal Bayes classification error is intractable, so most standard algorithms optimize a tractable heuristic function in the projected subspace. Here, we investigate a physics-based model where we consider the labeled data as interacting fluid distributions. We derive the forces arising in the fluids from information theoretic potential functions, and consider appropriate low rank constraints on the resulting acceleration and velocity flow fields. We show how to apply the Gauss principle of least constraint in fluids to obtain tractable solutions for low rank projections. Our fluid dynamic approach is demonstrated to better approximate the Bayes optimal solution on Gaussian systems, including infinite dimensional Gaussian processes.

1 Introduction

Algorithms for discovering interesting low-dimensional projections of data have been used by the statistics community for many decades (Friedman & Tukey, 1974; Huber, 1985; Duda et al., 2000). Projection pursuit is a canonical approach to find a low dimensional subspace where the projected data maximizes certain statistical properties. One example of such an approach is Fisher Discriminant Analysis (FDA), which has been applied in many domains due to its simplicity and ease of implementation. For the special case of separating two classes of homoscedastic Gaussian data, it can be shown that the simple criterion used by FDA results in an optimal projection, in that no other subspace can better separate the data using an optimal Bayes classifier. However, if the labeled data is heteroscedastic, non-Gaussian, or consists of more than two classes, standard approaches such as FDA can easily fail.

Some recent work on discriminant analysis have focused on finding better low-dimensional projection subspaces in these more difficult cases. These algorithms optimize a modified criterion function between the projected data, rather than using the simple heuristic employed by FDA. Examples of such algorithms have used various approximations to the Bayes error, including criterion motivated by information theory such as the Bhattacharyya coefficient and mutual information (Das & Nenadic, 2008; Hamsici & Martinez, 2008; Loog & Duin, 2004; Nenadic, 2007). The common theme among these algorithms is that they analyze *projected* data distributions, and try to optimize a criterion that is a function of the projected data statistics. However, with these more general nonlinear criterion, the optimization over low-dimensional projection matrices can be very non-convex. These algorithms typically rely upon gradient-based optimization techniques which can easily get caught in local minima in these applications.

Our approach, on the other hand, does not start by immediately considering low-dimensional projections of the high dimensional labeled data. Instead, we consider the data as interacting fluids in the high dimensional space. As shown in Fig. 1, our algorithm analyzes the structure of the resulting motions between the fluids in the high-dimensional space. In particular for this work, we consider the fluids as interacting via a potential function derived from the Bhattacharyya coefficient, which has been shown to be closely related to the Bayes classification error. This interaction potential induces forces within the fluids, which will cause

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9.

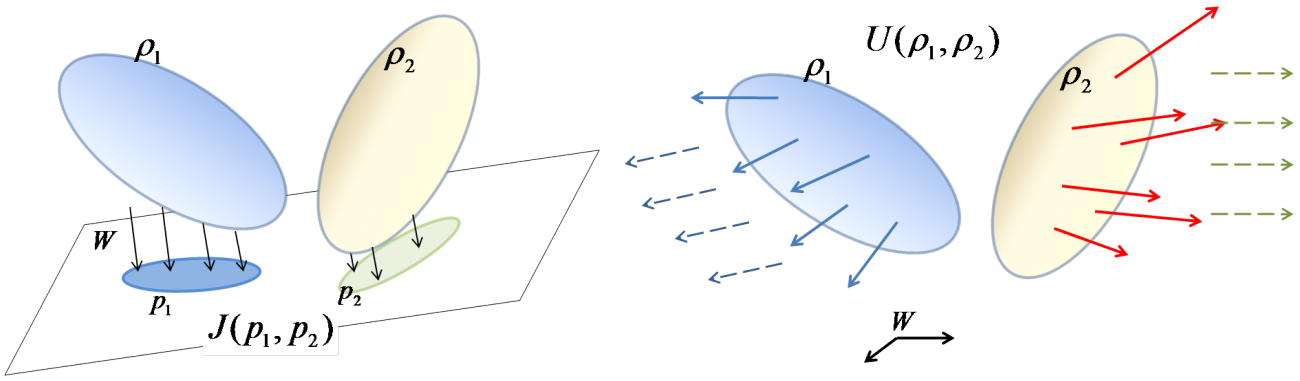


Figure 1: Conventional dimensionality reduction techniques measure the contrast between projected data distributions, while our approach analyzes low-rank constraints on the resulting flow fields of high dimensional interacting fluids. The projection matrix W will be found using the flow under the potential $U(\rho_1, \rho_2)$ instead of optimizing the non-convex function J of the projected distributions.

them to flow in response. We then introduce low-rank constraints on the induced flow fields, and consider constrained flows that best match the fluid forces.

For optimizing over the various constrained motions, we use the Gauss principle of least constraint, which describes how constrained physical systems move in response to applied forces. In this work, the Gauss principle is used to derive a phyically-motivated criterion that describes the low-rank flow field in response to the Bhattacharyya potential energy. In particular, we show how this optimization with low-rank affine constraints on the acceleration and velocity fields reduces to an eigenvector problem even for heteroscedastic Gaussian data.

Earlier studies in discriminant analysis for Gaussian distributions presented two special cases where optimal solutions exist for two-class problems. One situation arises when the covariance matrices are equal (homoscedastic), whereas the other less well-known solution occurs when the two means are the same. These two extreme cases are situations where there are known analytic solutions that minimize Bayes classification error. Standard discriminant algorithms such as FDA do not agree with these analytic solutions in both cases. We show that our physical fluid discriminant model approximates the optimal Bayes error criterion in both of these special cases.

The remainder of the paper is organized as follows. Section 2 reviews discriminant analysis using the Bhattacharyya coefficient and other optimization approaches. Section 3 derives our fluid dynamical algorithm for optimizing constrained low-rank fluid flows under an interaction potential, and compares our resulting solutions with known analytic solutions in special cases. We compare results on machine learning datasets in Section 4, and an application to Gaussian processes in Section 5. Finally, we conclude with a discussion in Section 6.

2 Discriminant Analysis for Classification

The goal of dimensionality reduction with labeled data is to find a low dimensional subspace where classification is better than other subspaces, so that the classification error is minimal. We consider the Bayes classification error within the subspace to be the fundamental criterion for subspace search. It is

$$J_1 = \frac{1}{2}\int \min[p_1(\mathbf{x}), p_2(\mathbf{x})]d\mathbf{x} \tag{1}$$

for two classes having projected distributions $p_1(\mathbf{x})$ and $p_2(\mathbf{x})$, and equal priors. The projected distributions are considered as Gaussians with means $W^T\mu_1$ and $W^T\mu_2$, and covariance matrices $W^T\Sigma_1 W$ and $W^T\Sigma_2 W$ for projection matrix $W \in \mathbb{R}^{D\times d}$ where D and d are the dimensions of the original space and the projected space ($D > d$). Unfortunately, minimizing J_1 is intractable due to the non-analytic behavior of the $\min[\cdot,\cdot]$ function.

In this work, we focus on an alternative criterion which is the Bhattacharyya distance (negative log of the Bhattacharyya coefficient), $J_2 = -\ln \int \sqrt{p_1(\mathbf{x})p_2(\mathbf{x})}d\mathbf{x}$. For Gaussians, its integrated form is the sum of two simple terms:

$$J_2(W) = \frac{1}{8}tr[(W^T S_w W)^{-1} W^T S_b W] + \frac{1}{2}\ln\frac{|\frac{1}{2}(W^T\Sigma_1 W + W^T\Sigma_2 W)|}{|W^T\Sigma_1 W|^{1/2}|W^T\Sigma_2 W|^{1/2}} \tag{2}$$

where S_w and S_b are defined as $S_w = \frac{\Sigma_1+\Sigma_2}{2}$ and $S_b = \Delta\mu\Delta\mu^T$ for $\Delta\mu = \mu_1 - \mu_2$, and $|\cdot|$ is the matrix determinant. Maximizing this distance also reduces the J_1 error criterion due to the inequality $\min[p_1(\mathbf{x}), p_2(\mathbf{x})] \leq \sqrt{p_1(\mathbf{x})p_2(\mathbf{x})}$. However, analytic solutions are intractable for general cases, and gradient-based optimization techniques for this criterion (Choi & Lee, 2003) may suffer from many local optima.

Fukunaga previously showed two special cases where this criterion is exactly solved as a generalized eigenvector problem (Fukunaga, 1990). The first case is the homoscedastic case ($\Sigma_1 = \Sigma_2$) and the second case is the equal mean case ($\mu_1 = \mu_2$). When the two covariances are the same, the optimal solution reduces to finding W which maximizes the first term $tr[(W^T S_w W)^{-1} W^T S_b W]$. This is the criterion used in FDA, whose solution are the eigenvectors of the generalized eigenvector problem,

$$S_b W = S_w W \Lambda \tag{3}$$

for diagonal eigenvalue matrix Λ. Dimensionality reduction is then performed by projecting the data onto the principal eigenvectors of the FDA solution.

On the other hand, when the two means are the same, the first term disappears, and we can maximize solely the second term $\ln \frac{|\frac{1}{2}(W^T\Sigma_1 W + W^T\Sigma_2 W)|}{|W^T\Sigma_1 W|^{1/2}|W^T\Sigma_2 W|^{1/2}}$. This can again be expressed as an eigenvector problem:

$$(\Sigma_1^{-1}\Sigma_2 + \Sigma_2^{-1}\Sigma_1 + 2I)W = W\Lambda. \tag{4}$$

To solve this, Fukunaga noted that $\Sigma_1^{-1}\Sigma_2$ and $\Sigma_2^{-1}\Sigma_1$ share the same eigenvectors because they are inverses. The optimal basis W can easily be obtained by solving the eigenvector problem $\Sigma_2 W = \Sigma_1 W \Lambda'$ and choosing eigenvectors according to the order of the values $\lambda_i' + \frac{1}{\lambda_i'} + 2$, where λ_i' is the corresponding eigenvalue in Λ'.

The intuitive interpretation of Eq. (4) leads to the conclusion that they correspond to the optimal subspace in terms of Bayes error. When Σ_1 and Σ_2 are simultaneously diagonalized by matrix W, the columns of W can be scaled to satisfy $W^T(\Sigma_1 + \Sigma_2)W = I$, while keeping $W^T\Sigma_1 W = D_1$ and $W^T\Sigma_2 W = D_2$ to be diagonal. In this case, the ith diagonal elements of D_1 and D_2 satisfy $d_{1i} \geq 0$ and $d_{2i} \geq 0$, and the sum of two elements always satisfies $d_{1i} + d_{2i} = 1$. The solution of Fukunaga's equal mean solution are the leading columns of W having maximum values of $\frac{d_{1i}}{d_{2i}} + \frac{d_{2i}}{d_{1i}} + 2 = \frac{d_{1i}+d_{2i}}{d_{2i}} + \frac{d_{1i}+d_{2i}}{d_{1i}} = \frac{1}{d_{1i}} + \frac{1}{d_{2i}}$. The problem of maximizing $\frac{1}{d_{1i}} + \frac{1}{d_{2i}}$ with constraints $d_{1i} + d_{2i} = 1$, $d_{1i} \geq 0$, and $d_{2i} \geq 0$ is equivalent to finding d_{1i} and d_{2i} that differ the most. So Fukunaga's equal mean analysis is equivalent to finding a basis having maximal difference in variances, resulting in the optimal subspace in terms of minimizing Bayes error.

Thus, in these two special cases, we know exactly which projections are the optimal subspace for Bayes classification, and the objective J_2 reduces to a generalized eigenvector problem. In the following, we compare our dimensionality reduction model in these special cases to these analytic solutions.

3 Derivation of the Equation of Motion

In this section, we construct an analogous fluid model for dimensionality reduction. We derive equations for fluid flow in high dimensional data spaces, and see how the solution can be appropriately constrained so that it can be used for discriminant analysis.

3.1 Fluid densities and equation of continuity

Our basic idea is to consider the class data as high dimensional mass distributions and analyze their resulting motion under interaction forces. In particular, we consider high dimensional fluids so that every point $\mathbf{x}$ moves according to a velocity field $\mathbf{v}(\mathbf{x})$ that minimizes a potential energy corresponding to the overlap of the different class densities. The overall structure of the resulting motion is approximated by constrained velocity and acceleration fields, and these constraints correspond to dominant projections for discriminant analysis.

We first consider a Gaussian mass distribution for each class $c \in \{1, 2, \ldots, C\}$ having mean μ_c and covariance matrix Σ_c.

$$\rho_c(\mathbf{x}) = \frac{1}{\sqrt{2\pi}^D |\Sigma_c|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu_c)^T\Sigma_c^{-1}(\mathbf{x}-\mu_c)} \tag{5}$$

Mass conservation gives rise to the equation of continuity:

$$\frac{\partial \rho_c}{\partial t} + \nabla \cdot (\rho_c \mathbf{v}) = 0. \tag{6}$$

which restricts the divergence of velocities $\mathbf{v}$ around a point $\mathbf{x}$ at time t. We next use this equation to derive the resulting force field corresponding to a given potential function.

3.2 Force for class separation

The Bhattacharyya coefficient is an information theoretic measure that can be used as a potential function U to describe the overlap between two fluid distributions:

$$U(\rho_1, \rho_2) = \int \sqrt{\rho_1 \rho_2} d\mathbf{x}. \tag{7}$$

This potential induces a local force at every spatial position that acts to minimize the interaction energy. This force field can be derived using the equation of continuity (6), and the relation $dU = -\int (F \cdot d\mathbf{s})\, d\mathbf{x}$, where $d\mathbf{s}$ is an infinitesimal displacement of fluid and $d\mathbf{x}$ is the integration measure. By analyzing how infinitesimal changes of the distribution $\rho_2(\mathbf{x})$ affect the potential U with fixed $\rho_1(\mathbf{x})$, the force field F_2 exerted on the class 2 is obtained:

$$\begin{aligned} F_2(\mathbf{x}) &= -\frac{1}{2}\rho_2 \nabla \sqrt{\frac{\rho_1}{\rho_2}} \\ &= \frac{1}{4} C_1 e^{-\frac{1}{4}(\mathbf{x}-\mu'_+)^T(\Sigma_1^{-1}+\Sigma_2^{-1})(\mathbf{x}-\mu'_+)} \\ &\quad \cdot(\Sigma_1^{-1} - \Sigma_2^{-1})(\mathbf{x}-\mu'_-) \end{aligned} \tag{8}$$

with the constants:

$$C_1 = \frac{e^{-\frac{1}{4}\Delta\mu^T(\Sigma_1+\Sigma_2)^{-1}\Delta\mu}}{(2\pi)^{d/2}|\Sigma_1|^{1/4}|\Sigma_2|^{1/4}} \tag{9}$$

$$\mu'_+ = \left(\Sigma_1^{-1} + \Sigma_2^{-1}\right)^{-1}(\Sigma_1^{-1}\mu_1 + \Sigma_2^{-1}\mu_2) \tag{10}$$

$$\mu'_- = \left(\Sigma_1^{-1} - \Sigma_2^{-1}\right)^{-1}(\Sigma_1^{-1}\mu_1 - \Sigma_2^{-1}\mu_2). \tag{11}$$

The same derivation can be used to obtain the forces F_1 on class 1, and we see $F_1(\mathbf{x}) = -F_2(\mathbf{x})$, in accordance with Newton's third law. The resulting force fields for some example density distributions are shown in Fig. 2.

3.3 Gauss principle of least constraint

Given the force fields derived in the previous section, the equations of motion can be derived. In general, without constraints on the system, the resulting motions will be high-dimensional. For our purposes, we constrain the resulting motion so that the fluids can only flow along certain directions. The optimal fluid flow directions will then correspond to a low dimensional space for discriminant analysis.

When constraints are applied to a system of many interacting particles, the Gauss principle of least constraint is a useful method to derive the equation of motion for each particle. The principle states that the motion follows the trajectory having the least amount of constraint force. When the force and constraints are given, this results in an optimization problem minimizing the objective function $\frac{1}{2}\sum_i m_i(\ddot{\mathbf{x}}_i - \frac{F_i}{m_i})^2$, where $\ddot{\mathbf{x}}_i$ is the acceleration of mass m_i satisfying the imposed constraints.

The constraint force is the difference between the applied force and the force governing the actual movement, $F - m\ddot{\mathbf{x}}$, and the objective is a weighted linear combination of these constraint forces. For continuous fluids, the objective becomes an integral of constraint forces over space, where the mass density function $\rho(\mathbf{x})$ is used in place of point masses:

$$L = \frac{1}{2}\int \rho(\mathbf{x}) \left(\ddot{\mathbf{x}} - \frac{F(\mathbf{x})}{\rho(\mathbf{x})}\right)^2 d\mathbf{x} \tag{12}$$

We will see how various constraints on the fluid flow can be analyzed using (12). In particular, a very restrictive constraint that only allows uniform translational motion will result in global fluid flow along a single direction. On the other hand, less restrictive constraints that allow local differences in flow velocities can capture more of the fine structure of fluid flow motion. Next we show how to tractably obtain the optimal low dimensional subspace from these types of constraints for discriminant analysis.

3.4 Constraint on motion: uniform translational movement

We first assume very hard constraints so that the fluid is a rigid body having only translational motion. In this case, the flow acceleration field is constant over the entire space, $\ddot{\mathbf{x}} = \mathbf{w}$. Minimizing the objection function

$$L(\mathbf{w}) = \frac{1}{2}\int \rho_c(\mathbf{x}) \left(\mathbf{w} - \frac{F_c(\mathbf{x})}{\rho_c(\mathbf{x})}\right)^2 d\mathbf{x} \tag{13}$$

with respect to the direction of the uniform acceleration field $\mathbf{w}$ yields the FDA direction:

$$\mathbf{w} = \int F d\mathbf{x} = C_2 \left(\frac{\Sigma_1 + \Sigma_2}{2}\right)^{-1} \Delta\mu \tag{14}$$

for class 2, and the opposite direction $C_2(\frac{\Sigma_1+\Sigma_2}{2})^{-1}(-\Delta\mu)$ for class 1. Here, the constant C_2 is $C_2 = \frac{1}{4}\left(\frac{|\Sigma_1|^{\frac{1}{2}}|\Sigma_2|^{\frac{1}{2}}}{\left|\frac{\Sigma_1+\Sigma_2}{2}\right|}\right)^{\frac{1}{2}} e^{-\frac{1}{4}\Delta\mu^T(\Sigma_1+\Sigma_2)^{-1}\Delta\mu}$. From this example, we see that the optimal rigid body translational motion under the Bhattacharyya coefficient interaction is equivalent to motion along the direction of the FDA solution.

3.5 Constraint on motion: low rank affine acceleration

We relax the constraint so that the fluid acceleration is described by an affine function in a low dimensional space. First, we define a low rank acceleration field as $\ddot{\mathbf{x}} = Wa(\mathbf{x})$ where $W \in \mathbb{R}^{D\times d}$ for $d < D$ is a low rank tall rectangular matrix. The acceleration field $a_c(\mathbf{x})$ of class $c \in \{1, 2\}$ is an affine function that can be expressed by $a_c(\mathbf{x}) = U_c^T\mathbf{x}_e$ where $\mathbf{x}_e = [\mathbf{x}^T\ 1]^T$ and $U_c \in \mathbb{R}^{(D+1)\times d}$. This constrains the transformation matrix WU^T to be a low rank affine transformation.

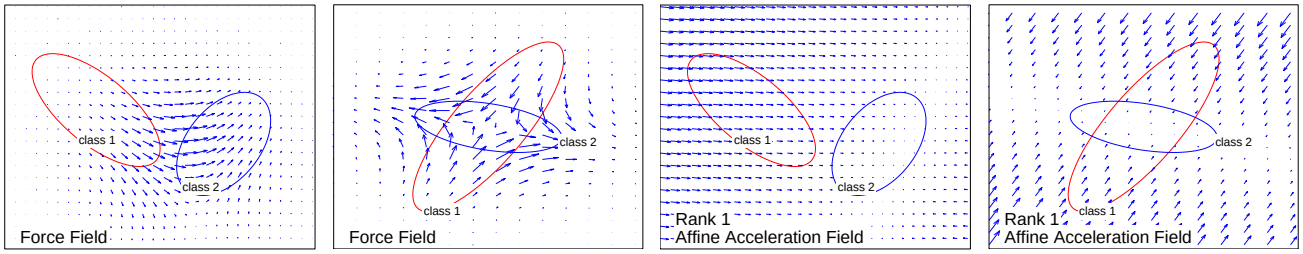


Figure 2: Two Gaussian distributions of different classes are represented as ellipses (standard deviation from the mean). The left two panels show the force field on class 2 from class 1. The force is a repulsive force (first panel) to minimize the overlap between classes; but if the distributions are overlapped (second panel), it tries to minimize $U(\rho_1, \rho_2)$ by squeezing one class rather than pushing it. The right two panels show the resulting acceleration field when the low rank affine constraint is applied.

In our setting, W is shared by different classes. The Gauss principle then yields the objective function

$$L = \frac{1}{2}\sum_{c=1}^{2}\int \rho_c \left(W U_c^T \mathbf{x}_e - \frac{F_c}{\rho_c} \right)^2 d\mathbf{x}. \tag{15}$$

With the force field defined in (8), the low rank matrix W that optimizes the objective function L is given by the span of the leading eigenvectors of the following symmetric matrix:

$$\sum_{c=1}^{2} \langle F_c \mathbf{x}_e^T \rangle \langle \mathbf{x}_e \mathbf{x}_e^T \rangle_{\rho_c}^{-1} \langle \mathbf{x}_e F_c^T \rangle. \tag{16}$$

The sufficient statistics for this computation are $\langle \mathbf{x}_e \mathbf{x}_e^T \rangle_{\rho_c} = \int \rho_c \mathbf{x}_e \mathbf{x}_e^T d\mathbf{x}$ and $\langle \mathbf{x}_e F_c^T \rangle = \int \mathbf{x}_e F_c^T d\mathbf{x}$. For multiway classification with $c = \{1, 2, \ldots, C\}$, we simply extend the previous analysis to every pair of classes to see that the optimal W is given by the principal eigenvectors of $\sum_{c=1}^{C} \langle F_c \mathbf{x}_e^T \rangle \langle \mathbf{x}_e \mathbf{x}_e^T \rangle_{\rho_c}^{-1} \langle \mathbf{x}_e F_c^T \rangle$ where F_c is the sum of forces from all other classes to class c. Fig. 2 shows an example force field and the acceleration field constrained by a rank one affine transformation for two class Gaussians.

Previously, we noted two scenarios where we know analytic solutions minimizing the Bayes classification error. These cases are when the data are two homoschedastic Gaussians, and when the Gaussians have the same mean. Now, we check how our fluid dynamic solution Eq. (16) approximates the known analytic solutions in these two special cases.

3.6 Validity analysis of the low rank affine approximation

The matrix (16) is given from μ_1, μ_2, Σ_1, and Σ_2 using the sufficient statistics:

$$\langle \mathbf{x}_e \mathbf{x}_e^T \rangle_{\rho_c}^{-1} = \begin{pmatrix} \Sigma_c + \mu_c \mu_c^T & \mu_c \\ \mu_c^T & 1 \end{pmatrix}^{-1}, \quad c \in \{1, 2\} \tag{17}$$

$$\langle \mathbf{x}_e F_2^T \rangle = -\langle \mathbf{x}_e F_1^T \rangle = \tag{18}$$

$$2C_2 \begin{bmatrix} \Sigma_2 - \Sigma_1 + \{\Sigma_2(\Sigma_1+\Sigma_2)^{-1}\mu_1 + \\ \Sigma_1(\Sigma_1+\Sigma_2)^{-1}\mu_2\}(-\Delta\mu)^T \\ \\ (-\Delta\mu)^T \end{bmatrix} (\Sigma_1+\Sigma_2)^{-1}$$

where the constant C_2 is the same as in Eq. (14).

When the two covariance matrices Σ_1 and Σ_2 are the same, the matrix (16) becomes the rank one symmetric matrix

$$4C_2^2(\Sigma_1+\Sigma_2)^{-1}\Delta\mu\Delta\mu^T(\Sigma_1+\Sigma_2)^{-1}\Big[2I + \Delta\mu\Delta\mu^T(\Sigma_1+\Sigma_2)^{-1}\Big], \tag{19}$$

and the only eigenvector with nonzero eigenvalue of this matrix is $\mathbf{w} = (\Sigma_1+\Sigma_2)^{-1}\Delta\mu$, which is equivalent to the FDA solution.

Otherwise, when $\mu_1 = \mu_2$, all terms including $\Delta\mu$ disappear, and the matrix becomes

$$4C_2^2(\Sigma_1+\Sigma_2)^{-1}\Big[(\Sigma_1-\Sigma_2)\Sigma_2^{-1}(\Sigma_1-\Sigma_2) + (\Sigma_2-\Sigma_1)\Sigma_1^{-1}(\Sigma_2-\Sigma_1)\Big](\Sigma_1+\Sigma_2)^{-1} \tag{20}$$

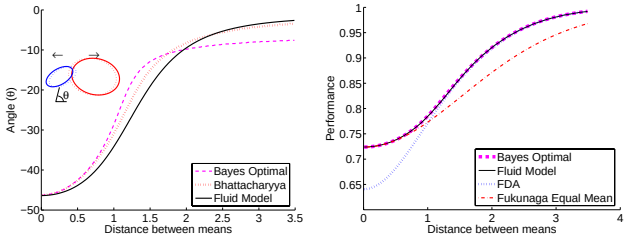
The eigenvector equation using this symmetric matrix can be rearranged as

$$(\Sigma_1\Sigma_2^{-1} + \Sigma_2\Sigma_1^{-1} - 2I)W = (\Sigma_1+\Sigma_2)W\Lambda, \tag{21}$$

which has a similar form to the eigenvector equation for Fukunaga's equal mean case. If we compare this to (4), we can see the solutions are equivalent if Σ_1 and Σ_2 commute and the sum of Σ_1 and Σ_2 is isotropic: $\Sigma_1 + \Sigma_2 = \alpha I$. If symmetric matrices Σ_1 and Σ_2 commute, there is an orthonormal matrix W (satisfying $W^TW = I$) that can diagonalize both Σ_1 and Σ_2. Therefore, we can write $W^T\Sigma_1 W = D_1$ and $W^T\Sigma_2 W = D_2$ for two diagonal matrices D_1 and

D_2. Solving Eq. (21) becomes the problem of finding $\mathbf{w}_i$s that are the columns of W with eigenvalues $\frac{1}{d_{1i}+d_{2i}}\left(\frac{d_{1i}}{d_{2i}}+\frac{d_{2i}}{d_{1i}}-2\right)$, where d_{1i} and d_{2i} are the ith elements of D_1 and D_2. It will be $\frac{1}{d_{2i}}+\frac{1}{d_{2i}}$ when $d_{1i}+d_{2i}=\alpha$, or $\mathbf{w}_i^T(\Sigma_1+\Sigma_2)\mathbf{w}_i=\alpha$, giving the same solution as Fukunaga's equal mean solution in Section 2.

To show the validity of our model on a simple 2-dimensional example, Fig. 3 shows the results obtained by directly optimizing the Bayes error J_1, the Bhattacharrya criterion J_2, along with our fluid model approximation. In this simple example, finding the optimal projection angle for J_1 and J_2 is performed by numerically scanning over all possible angles. On the other hand, the fluid model yields a tractable 2-dimensional eigenvector problem that performs quite well compared with the optimal solutions as in Fig. 3(a). In terms of the expected performance in Fig. 3(b), the fluid model achieves the optimal Bayes performance over the entire distance between means.



(a) (b)

Figure 3: (a) This panel shows the optimal discriminating direction for two 2-D Gaussians as we increase the distance between means. The Bayes optimal direction, the Bhattacharyya optimal direction, and the fluid approximation are presented. (b) The expected classification performances for the same distributions as (a) are presented using the Bayes optimal direction, fluid solution, FDA solution, and Fukunaga's equal mean solution.

4 Experiments on Standard Datasets

The proposed method is evaluated on several benchmark datasets. The data are projected onto the subspaces found by FDA, Fukunaga's equal mean analysis, and the fluid model, then Bayes classification is performed in the subspace assuming Gaussian distributions. The results are presented in Table 1, and they show our fluid approach generally outperforms or is similar to the other two analyses. We also note that Fukunaga's equal mean analysis often beats FDA, even though it is less well-known.

The experimental datasets are the USPS handwritten digit dataset and six UC Irvine datasets. For the USPS set, the original images of 16×16 pixels are first resized to 10×10 to reduce the sparsity of pixels. Non-overlapping training and testing data are randomly sampled from the data, and the sampling numbers are balanced over all classes. The experimental results are the averages over 100 realizations for UC Irvine datasets, and 20 realizations for USPS dataset. The number of extracted dimensions is $C-1$ which is the maximum number of dimensions standard FDA can extract with C number of classes. The empirical covariance matrices are regularized by adding a small multiple of the identity matrix, which is determined by cross validation for each algorithm.

Table 1: Performance on benchmark datasets (%). For each dataset, the number of classes (C), and performance of Fukunaga's equal mean solution, FDA, and our fluid model are presented.

Dataset	C	Fukunaga	FDA	Fluid
SpectF	2	78.50	80.20	**81.70**
Ionosphere	2	86.83	85.96	**87.54**
Parkinsons	2	86.83	82.33	**89.33**
Ozone	2	71.27	**84.54**	84.20
Breast Cancer	2	93.83	97.87	**97.92**
Glass	6	·	52.53	**55.67**
USPS	10	·	90.38	**91.48**

5 Application to Gaussian Processes

We also extend our fluid model to the problem of treating data in an infinite dimensional space using Gaussian processes. The problem of optimally discriminating two Gaussian processes is intractable, similar to the problems of analyzing finite dimensional problems mentioned in Section 2. In this section, we introduce our method for low rank discrimination of two different stochastic dynamical systems that can be modeled by Gaussian processes.

The dynamical systems are expressed by the following linear state equation, resulting in a Gaussian process. We consider an infinite length sequence $X=[\ldots,x_{m-1},x_m,x_{m+1},\ldots]$ along with the following recursive update rule,

$$x_m=\alpha x_{m-1}+b_m+\epsilon_m,\quad 0<\alpha<1 \tag{22}$$

where each $\epsilon_m\in\mathbb{R}$ follows a Gaussian distribution $N(0,\sigma^2)$, and b_m, a input value at time m. Then the sequence X is a Gaussian process which has a mean function $\mu(m)=\langle x_m\rangle$ and a covariance function $f(m,n)=\langle x_m x_n\rangle-\langle x_m\rangle\langle x_n\rangle$ which is translationally invariant.

To obtain $\mu(m)$ and $f(m,n)$, we use the recursive equation (22) to obtain a general expression for x_m,

$$x_m = \sum_{i=1}^{\infty} \alpha^{i-1}(b_{m-i} + \epsilon_{m-i}). \tag{23}$$

This equation leads to the mean and the covariance function below:

$$\mu(m) = \langle x_m \rangle = \sum_{i=1}^{\infty} \alpha^{i-1} b_{m-i} \tag{24}$$

$$f(m,n) = \langle x_m x_n^T \rangle - \langle x_m \rangle \langle x_n \rangle^T = \frac{\sigma^2 \alpha^{|m-n|}}{1-\alpha^2} \tag{25}$$

We now consider two different systems having different α_c and b_{cm}, for $c \in \{1,2\}$, and see how our methods can be applied to finding a low dimensional projection for discriminating different dynamical systems.

5.1 Finding low dimensional filters

The one dimensional projection vector for infinite sequence data can be considered as a filter whose inner product with the sequence yields a scalar output. Thus, the projection vector can be expressed as a filter function $w(m)$, and we seek the best filter that can be used to discriminate two dynamical systems. If we use the mean functions and the covariance functions of Gaussian processes, Fukunaga's two special cases can be equivalently analyzed. We show how our tractable fluid model can well approximate this optimal filter $w(m)$.

We first look at a simple example where the FDA solution is optimal to get the intuition about how Gaussian process covariances are applied, then we extend the analysis for more general cases using our fluid model.

5.1.1 FDA analysis

Consider the special case where the two covariance functions are the same, and the two mean functions are different. In this case, the optimal discriminating filter between the dynamical systems is the FDA solution, which is , $w(m) = \int (f_1(m,n) + f_2(m,n))^{-1} (\mu_1(n) - \mu_2(n)) dn$ analogous to the finite dimensional form $(\Sigma_1 + \Sigma_2)^{-1}\Delta\mu$. Here, the subscripts indicate classes, and the inverse function satisfies $\int f^{-1}(l,m) f(m,n) dm = \delta(l,n)$ for Kronecker delta δ. The inverse function can be obtained through the Fourier and the inverse Fourier transforms.

An example is presented in Fig. 4 containing two dynamical systems satisfying Eq. (22). The two systems have the same α and different b_{cm} for different classes c, so they are processes having the same covariance and different means. The samples of different classes in Fig. 4(a) show variations in the region where there is an obvious difference in the mean functions. We consider two filters given by FDA (Fig. 4(b)) and the mean difference (Fig. 4(c)). The FDA solution mainly uses the difference between the starting point and the maximum point of mean differences, whereas the mean difference averages over this region. The difference is given by the influence of the inverse covariance function in the FDA solution that deconvolves the smearing of the Gaussian process. If we look at the projected distributions in Fig. 4(d), we see the role of this deconvolution. The mean difference direction just tries to maximize the distance between the means in the projection, while the deconvolved direction considers both the maximization of mean difference and the minimization of within class variance.

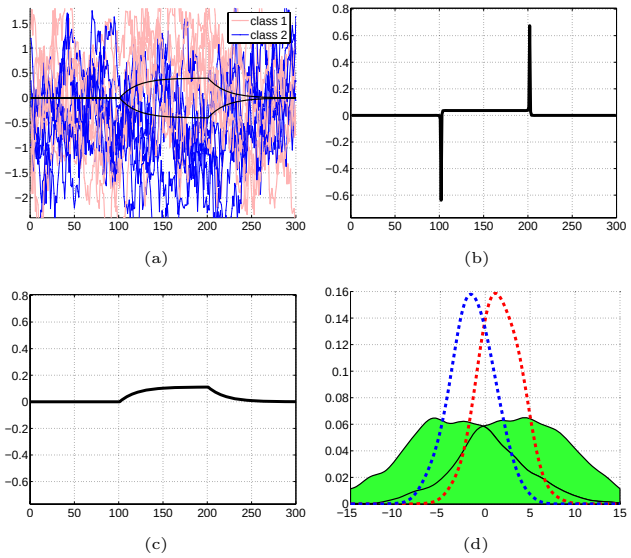


Figure 4: (a) The sample signals and the mean functions of two Gaussian processes, (b) FDA filter solution, (c) mean difference filter, and (d) the projected distributions. In (d), the two colored distribution in green are the projected distributions onto the mean difference direction. The dotted distributions are the projected distributions onto the FDA direction.

5.1.2 Fluid model extension

Next, we consider the general case where both the means and covariances of the dynamical processes show differences. As in the finite dimensional heteroscedastic problem, finding the optimal filter for discrimination is generally not tractable. However, applying our fluid model, the elements of $\langle F \mathbf{x}_e^T \rangle \langle \mathbf{x}\mathbf{x}^T \rangle_\rho \langle \mathbf{x}_e F^T \rangle$ can be expressed as $\iint e^{i\omega_1 m} e^{-i\omega_2 n} H(\omega_1, \omega_2) d\omega_1 d\omega_2$ where $H(\omega_1, \omega_2)$ is composed of two matrices which are diagonal ma-

trices whose components are $F_1(\omega)$ and $F_2(\omega)$, and the Hermitian matrix $M(\omega_1)M(\omega_2)^\dagger$ where $M(\omega) = \int \Delta\mu(t)e^{-i\omega t}dt$, through the Fourier transform. This problem can be solved by the eigendecomposition of $H(\omega_1, \omega_2)$.

In the example in Fig. 5(a), the noise is quite large compared to the mean difference. The FDA filter uses just two end points in the system responses, while our fluid model incorporates information about differences in the process covariances. The resulting filter projections shown in Fig. 5(d) show the advantages of using the fluid dynamics filter in this case. Quantitatively, the average of the resulting Bhattacharyya coefficients is 0.324 for the FDA filter, compared to 0.235 for the fluid dynamics filter. These results are averaged over 600 multiple realization of 1000 samples of Gaussian processes.

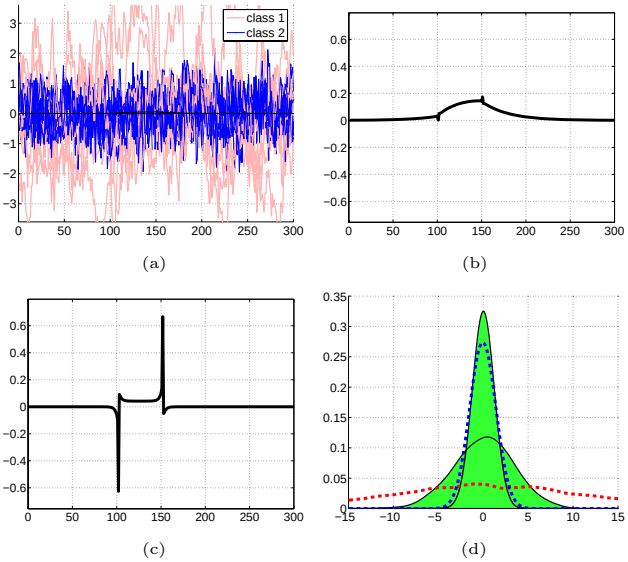


Figure 5: (a) Sample signals and the mean responses of two different dynamical Gaussian processes, (b) fluid model solution, (c) FDA solution, and (d) the projected filter distributions. In (d), the two colored distribution in green are the projected distributions by the FDA filter. The dotted distributions are the projected distributions using the fluid model filter.

6 Conclusions

We have presented a novel method for discriminant analysis by making an analogy to high-dimensional fluid dynamics. The model yields an optimal low-dimensional subspace for separating classes by analyzing the constrained motion of fluid densities under an information theoretic potential function. We have shown how this model relates well to the optimal subspace with Gaussian data as well as on standard machine learning databases. We also showed how this model can be applied to infinite dimensional systems, and can be used to discriminate Gaussian processes.

Future research will include dealing with nonlinear projections using kernel techniques, and extending the analysis to more general exponential families and non-parametric distributions.

Finally, we acknowledge the support provided by the IT R&D Program of MKE/KEIT (KI002138, MARS) and the NRF Grant of MEST (314-2008-1-D00377, Xtran).

References

Alvarez, M., Luengo, D., & Lawrence, N. (2009). Latent force models. *Twelfth International Conference on Artificial Intelligence and Statistics*.

Choi, E., & Lee, C. (2003). Feature extraction based on the Bhattacharyya distance. *Pattern Recognition*, *36*, 1703–1709.

Das, K., & Nenadic, Z. (2008). Approximate information discriminant analysis: A computationally simple heteroscedastic feature extraction technique. *Pattern Recognition*, *41*, 1548–1557.

Duda, R., Hart, P., & Stork, D. (2000). *Pattern classification (2nd edition)*. Wiley-Interscience.

Evans, D., & Morriss, G. (2008). *Statistical mechanics of nonequilibrium liquids*. Cambridge University Press, Cambridge. 2nd edition.

Friedman, J. H., & Tukey, J. W. (1974). A projection pursuit algorithm for exploratory data analysis. *IEEE Transactions on Computers*, *23*, 881–890.

Fukunaga, K. (1990). *Introduction to statistical pattern recognition*. San Diego, California: Academic Press.

Hamsici, O., & Martinez, A. (2008). Bayes optimality in linear discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *30*, 647–657.

Huber, P. J. (1985). Projection pursuit. *The Annals of Statistics*, *13*, 435–475.

Lifshitz, E., & Landau, L. (1987). *Fluid mechanics, second edition: Volume 6 (course of theoretical physics)*. Butterworth-Heinemann.

Loog, M., & Duin, R. (2004). Linear dimensionality reduction via a heteroscedastic extension of LDA: The Chernoff criterion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *26*, 732–739.

Nenadic, Z. (2007). Information discriminant analysis: Feature extraction with an information-theoretic objective. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*, 1394–1407.

Urtasun, R., & Darrell, T. (2007). Discriminative Gaussian process latent variable models for classification. *Proceedings of the 24th International Conference on Machine Learning*.