Journal of Machine Learning Research 8 (2007) 1277-1305 Submitted 6/06; Revised 1/07; Published 6/07

Bayesian Quadratic Discriminant Analysis

Santosh Srivastava SANTOSH@AMATH.WASHINGTON.EDU
Department of Applied Mathematics
University of Washington
Seattle, WA 98195, USA

Maya R. Gupta GUPTA@EE.WASHINGTON.EDU
Department of Electrical Engineering
University of Washington
Seattle, WA 98195, USA

Béla A. Frigyik BFRIGYIK@MATH.PURDUE.EDU
Department of Mathematics
Purdue University
West Lafayette, IN 47907, USA

Editor: Saharon Rosset

Abstract

Quadratic discriminant analysis is a common tool for classification, but estimation of the Gaussian parameters can be ill-posed. This paper contains theoretical and algorithmic contributions to Bayesian estimation for quadratic discriminant analysis. A distribution-based Bayesian classifier is derived using information geometry. Using a calculus of variations approach to define a functional Bregman divergence for distributions, it is shown that the Bayesian distribution-based classifier that minimizes the expected Bregman divergence of each class conditional distribution also minimizes the expected misclassification cost. A series approximation is used to relate regularized discriminant analysis to Bayesian discriminant analysis. A new Bayesian quadratic discriminant analysis classifier is proposed where the prior is defined using a coarse estimate of the covariance based on the training data; this classifier is termed BDA7. Results on benchmark data sets and simulations show that BDA7 performance is competitive with, and in some cases significantly better than, regularized quadratic discriminant analysis and the cross-validated Bayesian quadratic discriminant analysis classifier Quadratic Bayes.

Keywords: quadratic discriminant analysis, regularized quadratic discriminant analysis, Bregman divergence, data-dependent prior, eigenvalue decomposition, Wishart, functional analysis

1. Introduction

A standard approach to supervised classification problems is quadratic discriminant analysis (QDA), which models the likelihood of each class as a Gaussian distribution, then uses the posterior distributions to estimate the class for a given test point (Hastie et al., 2001). The Gaussian parameters for each class can be estimated from training points with maximum likelihood (ML) estimation. The simple Gaussian model assumption is best suited to cases when one does not have much information to characterize a class, for example, if there are too few training samples to infer much about the class distributions. Unfortunately, when the number of training samples n is small compared to the number of dimensions of each training sample d, the ML covariance estimation can be ill-posed.

One approach to resolve the ill-posed estimation is to regularize the covariance estimation; another approach is to use Bayesian estimation.

Bayesian estimation for QDA was first proposed by Geisser (1964), but this approach has not become popular, even though it minimizes the expected misclassification cost. Ripley (2001) in his text on pattern recognition states that such predictive classifiers are mostly unmentioned in other texts and that "this may well be because it usually makes little difference with the tightly constrained parametric families." Geisser (1993) examines Bayesian QDA in detail, but does not show that in practice it can yield better performance than regularized QDA. In preliminary experiments we found that the performance of Bayesian QDA classifiers is very sensitive to the choice of prior (Srivastava and Gupta, 2006), and that priors suggested by Geisser (1964) and Keehn (1965) produce error rates similar to those yielded by ML.

In this paper we propose a Bayesian QDA classifier termed *BDA7*. BDA7 is competitive with regularized QDA, and in fact performs better than regularized QDA in many of the experiments with real data sets. BDA7 differs from previous Bayesian QDA methods in that the prior is selected by crossvalidation from a set of data-dependent priors. Each data-dependent prior captures some coarse information from the training data. Using ten benchmark data sets and ten simulations, the performance of BDA7 is compared to that of Friedman's regularized quadratic discriminant analysis (RDA) (Friedman, 1989), to a model-selection discriminant analysis (EDDA) (Bensmail and Celeux, 1996), to a modern cross-validated Bayesian QDA (QB) (Brown et al., 1999), and to ML-estimated QDA, LDA, and the nearest-means classifier. Our focus is on cases in which the number of dimensions d is large compared to the number of training samples n. The results show that BDA7 performs slightly better than the other approaches averaged over the real data sets. The simulations help analyze the methods under controlled conditions.

This paper also contributes to the theory of Bayesian QDA in several aspects. First, the Bayesian classifier is solved for in terms of the Gaussian distributions themselves, as opposed to the standard approach of formulating the problem in terms of the Gaussian parameters. This distribution-based Bayesian discriminant analysis removes the parameter-based Bayesian analysis restriction of requiring more training samples than feature dimensions, and removes the question of invariance to transformations of the parameters, because the estimate is defined in terms of the Gaussian distribution itself. We show that the distribution-based Bayesian discriminant analysis classifier has roughly the same closed-form as the parameter-based Bayesian discriminant analysis classifier, but has a different degree of freedom.

The second theoretical result links Bayesian QDA and Friedman's regularized QDA by a using series approximation. Third, we show that the Bayesian distribution-based classifier that minimizes the expected misclassification cost is equivalent to the classifier that uses the Bayesian minimum expected Bregman divergence estimates of the class conditional distributions.

We begin with a review of approaches to Bayesian QDA and regularized QDA in Section 2. We formulate the distribution-based Bayesian classifier in Section 3, and review recent results showing that the distribution-based performance is superior to the parameter-based Bayesian classifier given the same prior if no cross-validation is allowed. An approximate relationship between Bayesian QDA and Friedman's regularized QDA is given in Section 4. In Section 5 we establish that the Bayesian minimum expected misclassification cost estimate is equivalent to a plug-in estimate using the Bayesian minimum expected Bregman divergence estimate for each class conditional. Then, we turn to the practical matter of classification: we propose the cross-validated Bayesian QDA classifier BDA7 in Section 6. In Section 7, benchmark data set results compare BDA7 to other QDA

Notation	Description	Notation	Description
I	identity matrix	$I_{(\cdot)}$	indicator function
$x_i \in \mathbb{R}^d$	i^{th} training sample	n	number of training samples
y_i	class label corresponding to x_i	n_h	number of training samples of class h
G	number of class labels	$\bar{x_h}$	sample mean for class h
$\mathcal{T}$	n pairs of training samples	$\mathcal{T}_h$	n_h pairs of class h training samples
$x \in \mathbb{R}^d$	test sample	S	$\sum_{i=1}^{n}(x_i-\bar{x})(x_i-\bar{x})^T$
Y	class label corresponding to x	S_h	$\sum_{i=1}^{n}(x_i-\bar{x_h})(x_i-\bar{x_h})^T I_{(y_i=h)}$
C	misclassification cost matrix	$\|B\|$	determinant of B
$\mathrm{diag}(B)$	diagonal of B	$\mathrm{tr}(B)$	trace of B
N	random Gaussian distribution	$\mathcal{N}$	realization of a Gaussian distribution

Table 1: Key Notation

classifiers, followed by further analysis using simulation results in Section 8. The paper concludes with a discussion of the results and some open questions.

Table 1 details some of the notation used in this paper.

2. Prior Research on Ill-Posed QDA

We review the prior literature on Bayesian approaches to QDA, regularization approaches to QDA, and other approaches to resolving ill-posed QDA.

2.1 Bayesian Approaches to QDA

Discriminant analysis using Bayesian estimation was first proposed by Geisser (1964) and Keehn (1965). Geisser's work used a noninformative prior distribution to calculate the posterior odds that a test sample belongs to a particular class. Keehn's work assumed that the prior distribution of the covariance matrix is an inverse Wishart distribution. Raudys and Jain (1991) stated that Bayesian QDA does not solve the problems that occur with ML QDA, particularly when class sample sizes differ. Recent work by the authors showed that Bayesian QDA using the priors suggested by Geisser and Keehn perform similarly to ML QDA on six standard QDA simulations (Srivastava and Gupta, 2006).

The inverse Wishart prior is a conjugate prior for the covariance matrix, and it requires the specification of a "seed" positive definite matrix and a scalar degree of freedom. Following Brown et al. (1999), the term *Quadratic Bayes* (QB) is used to refer to a modern form of Bayesian QDA where the inverse Wishart seed matrix is kI, where I is the d-dimensional identity matrix, k is a scalar, and the parameters k and the degree of freedom q of the inverse Wishart distribution are chosen by crossvalidation.

Previous Bayesian QDA work is *parameter-based* in that the mean μ and covariance Σ are treated as random variables, and the expectations of the Gaussian class-conditional distributions are

calculated with respect to Lebesgue measure over the domain of the parameters. In this paper we solve for the *distribution-based* Bayesian QDA, such that the uncertainty is considered to be over the set of Gaussian distributions and the Bayesian estimation is formulated over the domain of the Gaussian distributions.

2.2 Regularizing QDA

Friedman (1989) proposed regularizing ML covariance estimation by linearly combining a ML estimate of each class covariance matrix with the ML pooled covariance estimate and with a scaled version of the identity matrix to form an estimate $\hat{\Sigma}_h(\lambda,\gamma)$ for the hth class:

$$\hat{\Sigma}_h(\lambda) = \frac{(1-\lambda)S_h + \lambda S}{(1-\lambda)n_h + \lambda n}, \quad (1)$$

$$\hat{\Sigma}_h(\lambda,\gamma) = (1-\gamma)\hat{\Sigma}_h(\lambda) + \frac{\gamma}{d}\operatorname{tr}\left(\hat{\Sigma}_h(\lambda)\right) I. \quad (2)$$

The above notation and other key notation is defined in Table 1.

In Friedman's regularized QDA (RDA), the parameters λ,γ are trained by crossvalidation to be those parameters that minimize the number of classification errors. Friedman's comparisons to ML QDA and ML linear discriminant analysis on six simulations showed that RDA could deal effectively with ill-posed covariance estimation when the true covariance matrix is diagonal. RDA is perhaps the most popular approach to discriminant analysis when the covariance estimation is expected to be ill-posed (Hastie et al., 2001).

Hoffbeck and Landgrebe (1996) proposed a similar regularized covariance estimate for classification of the form

$$\hat{\Sigma} = \alpha_1 \operatorname{diag}(\hat{\Sigma}_{ML}) + \alpha_2 \hat{\Sigma}_{ML} + \alpha_3 \hat{\Sigma}_{\text{pooled}\,ML} + \alpha_4 \operatorname{diag}(\hat{\Sigma}_{\text{pooled}\,ML}),$$

where $\hat{\Sigma}_{\text{ML}}$ and $\hat{\Sigma}_{\text{pooled}\,ML}$ are maximum likelihood estimates of class and pooled covariance matrices, respectively, and the parameters $\alpha_1,\alpha_2,\alpha_3,\alpha_4$ are trained by crossvalidation to maximize the likelihood (whereas Friedman's RDA crossvalidates to maximize classification accuracy). Results on Friedman's simulation suite and experimental results on a hyperspectral classification problem showed that the two classifiers achieved similar accuracy (Hoffbeck and Landgrebe, 1996). Another restricted model used to regularize covariance matrix estimation is a banded covariance matrix (Bickel and Li, 2006).

RDA and the Hoffbeck-Landgrebe classifiers linearly combine different covariance estimates. A related approach is to select the best covariance model out of a set of models using crossvalidation. Bensmail and Celeux (1996) propose eigenvalue decomposition discriminant analysis (EDDA), in which fourteen different models for the class covariances are considered, ranging from a scalar times the identity matrix to a full class covariance estimate for each class. The model that minimizes the crossvalidation error is selected for use with the test data. Each individual model's parameters are estimated by ML; some of these estimates require iterative procedures that are computationally intensive.

The above techniques consider a discrete set of possible models for Σ, and either linearly combine models or select one model. In contrast, Bayesian QDA combines a continuous set of possible models for Σ, weighting each model by its posterior probability.

2.3 Other Approaches to Quadratic Discriminant Analysis

Other approaches have been developed for ill-posed quadratic discriminant analysis. Friedman (1989) notes that, beginning with work by James and Stein in 1961, researchers have attempted to improve the eigenvalues of the sample covariance matrix. Another approach is to reduce the data dimensionality before estimating the Gaussian distributions, for example by principal components analysis (Swets and Weng, 1996). One of the most recent algorithms of this type is orthogonal linear discriminant analysis (Ye, 2005), which was shown by the author of that work to perform similarly to Friedman's regularized linear discriminant analysis on six real data sets.

3. Distribution-based Bayesian Discriminant Analysis

Parameter estimation depends on the form of the parameter, for example, Bayesian estimation can yield one result if the expected standard deviation is solved for, or another result if the expected variance is solved for. To avoid this issue we derive Bayesian QDA by formulating the problem in terms of the Gaussian distributions explicitly. This section extends work presented in a recent conference paper (Srivastava and Gupta, 2006).

Suppose one is given an iid training set $\mathcal{T} = \{(x_i, y_i), i = 1, 2, \ldots n\}$ and a test sample x, where $x_i, x \in \mathbb{R}^d$, and y_i takes values from a finite set of class labels $y_i \in \{1, 2, \ldots, G\}$. Let C be the misclassification cost matrix such that $C(g,h)$ is the cost of classifying x as class g when the truth is class h. Let $P(Y = h)$ be the prior probability of class h. Suppose the true class conditional distributions $p(x|Y = h)$ exist and are known for all h, then the estimated class label for x that minimizes the expected misclassification cost is

$$Y^* \stackrel{\triangle}{=} \underset{g=1,\ldots,G}{\operatorname{argmin}} \sum_{h=1}^{G} C(g,h)\, p(x|Y=h) P(Y=h). \tag{3}$$

In practice the class conditional distributions and the class priors are usually unknown. We model each unknown distribution $p(x|h)$ by a random Gaussian distribution N_h, and we model the unknown class priors by the random vector Θ, which has components $\Theta_h = P(Y = h)$ for $h = 1, \ldots, G$. Then, we estimate the class label that minimizes the expected misclassification cost, where the expectation is with respect to the random distributions Θ and $\{N_h\}$ for $h = 1, \ldots, G$. That is, define the *distribution-based Bayesian QDA* class estimate by replacing the unknown distributions in (3) with their random counterparts and taking the expectation:

$$\hat{Y} \stackrel{\triangle}{=} \underset{g=1,\ldots,G}{\operatorname{argmin}} E\left[\sum_{h=1}^{G} C(g,h) N_h(x) \Theta_h\right]. \tag{4}$$

In (4) the expectation is with respect to the joint distribution over Θ and $\{N_h\}$ for $h = 1, \ldots, G$, and these distributions are assumed independent. Therefore (4) can be rewritten as

$$\hat{Y} = \underset{g=1,\ldots,G}{\operatorname{argmin}} \sum_{h=1}^{G} C(g,h) E_{N_h}[N_h(x)] E_{\Theta}[\Theta_h]. \tag{5}$$

Straightforward integration yields an estimate of the class prior, $E_{\Theta}[\Theta_h] = \frac{n_h+1}{n+G}$; this Bayesian estimate for the multinomial is also known as Laplace correction (Jaynes and Bretthorst, 2003).

In this next section we discuss the evaluation of $E_{N_h}[N_h(x)]$.

3.1 Statistical Models and Measure

Consider the family M of multivariate Gaussian probability distributions on $\mathbb{R}^d$. Let each element of M be a probability distribution $\mathcal{N} : \mathbb{R}^d \to [0,1]$, parameterized by the real-valued variables (μ, Σ) in some open set in $\mathbb{R}^d \otimes \mathbb{S}$, where $\mathbb{S} \subset \mathbb{R}^{d(d+1)/2}$ is the cone of positive semi-definite symmetric matrices. That is $M = \{\mathcal{N}(\cdot\,; \mu, \Sigma)\}$ defines a $\frac{d^2+3d}{2}$-dimensional statistical model, (Amari and Nagaoka, 2000, pp. 25–28).

Let the differential element over the set M be defined by the Riemannian metric (Kass, 1989; Amari and Nagaoka, 2000),

$$\begin{aligned} dM &= |I_F(\mu,\Sigma)|^{\frac{1}{2}} d\mu d\Sigma, \quad \text{where} \\ I_F(\mu,\Sigma) &= -E_X[\nabla^2 \log \mathcal{N}(X;(\mu,\Sigma))], \end{aligned}$$

where ∇^2 is the Hessian operator with respect to the parameters μ and Σ, and this I_F is also known as the Fisher information matrix. Straightforward calculation shows that

$$dM = \frac{d\mu}{|\Sigma|^{\frac{1}{2}}} \frac{d\Sigma}{|\Sigma|^{\frac{d+1}{2}}} = \frac{d\mu d\Sigma}{|\Sigma|^{\frac{d+2}{2}}}. \tag{6}$$

Let $\mathcal{N}_h(\mu_h, \Sigma_h)$ be a possible realization of the Gaussian pdf N_h. Using the measure defined in (6),

$$E_{N_h}[N_h(x)] = \int_M \mathcal{N}_h(x) r(\mathcal{N}_h) dM, \tag{7}$$

where $r(\mathcal{N}_h)$ is the posterior probability of $\mathcal{N}_h$ given the set of class h training samples $\mathcal{T}_h$; that is,

$$r(\mathcal{N}_h) = \frac{\ell(\mathcal{N}_h, \mathcal{T}_h) p(\mathcal{N}_h)}{\alpha_h}, \tag{8}$$

where α_h is a normalization constant, $p(\mathcal{N}_h)$ is the prior probability of $\mathcal{N}_h$ (treated further in Section 3.2), and $\ell(\mathcal{N}_h, \mathcal{T}_h)$ is the likelihood of the data $\mathcal{T}_h$ given $\mathcal{N}_h$, that is,

$$\ell(\mathcal{N}_h(\mu_h,\Sigma_h), \mathcal{T}_h) = \frac{\exp[-\frac{1}{2}\operatorname{tr}\left(\Sigma_h^{-1} S_h\right) - \frac{n_h}{2}\operatorname{tr}\left(\Sigma_h^{-1}(\mu_h - \bar{X_h})(\mu_h - \bar{X_h})^T\right)]}{(2\pi)^{\frac{dn_h}{2}} |\Sigma_h|^{\frac{n_h}{2}}}. \tag{9}$$

3.2 Priors

A prior probability distribution of the Gaussians, $p(\mathcal{N}_h)$, is needed to solve the classification problem given in (4). A common interpretation of Bayesian analysis is that the prior represents information that one has prior to seeing the data (Jaynes and Bretthorst, 2003). In the practice of statistical learning, one often has very little quantifiable information apart from the data. Instead of thinking of the prior as representing prior information, we considered the following design goals: the prior should

- regularize the classification to reduce estimation variance, particularly when the number of training samples n is small compared to the number of feature dimensions;
- add minimal bias;

- allow the estimation to converge to the true generating class conditional normals as $n \to \infty$ if in fact the data was generated by class conditional normals;
- lead to a closed-form result.

To meet these goals, we use as a prior

$$p(\mathcal{N}_h) = p(\mu_h)p(\Sigma_h) = \gamma_0 \frac{\exp[-\frac{1}{2}\operatorname{tr}\left(\Sigma_h^{-1}B_h\right)]}{|\Sigma_h|^{\frac{q}{2}}}, \tag{10}$$

where B_h is a positive definite matrix (further specified in (25)) and γ_0 is a normalization constant. The prior (10) is equivalent to a noninformative prior for the mean μ, and an inverted Wishart prior with q degrees of freedom over Σ.

This prior is unimodal and leads to a closed-form result. Depending on the choice of B_h and q, the prior probability mass can be focused over a small region of the set of Gaussian distributions M in order to regularize the estimation. Regularization is important for cases where the number of training samples n is small compared to the dimensionality d. However, the tails of this prior are sufficiently heavy that the prior does not hinder convergence to the true generating distribution as the number of training samples increases if the true generating distribution is normal.

The positive definite matrix B_h specifies the location of the maximum of the prior probability distribution. Using the matrix derivative (Dwyer, 1967) and the knowledge that the inverse Wishart distribution has only one maximum, one can calculate the location of the maximum of the prior. Take the log of the prior,

$$\log p(\mathcal{N}_h) = -\frac{1}{2}\operatorname{tr}\left(\Sigma_h^{-1}B_h\right) - \frac{q}{2}\log|\Sigma_h| + \log\gamma_0.$$

Differentiate with respect to Σ_h to solve for $\Sigma_{h,max}$,

$$\begin{aligned} -\frac{1}{2}\frac{\partial}{\partial\Sigma_h}\operatorname{tr}\left(\Sigma_{h,max}^{-1}B_h\right) - \frac{q}{2}\frac{\partial}{\partial\Sigma_h}\log|\Sigma_{h,max}| &= 0, \\ \Sigma_{h,max}^{-1}B_h\Sigma_{h,max}^{-1} - q\Sigma_{h,max}^{-1} &= 0, \\ \Sigma_{h,max} &= \frac{B_h}{q}. \end{aligned} \tag{11}$$

Because this prior is unimodal, a rough interpretation of its action is that it regularizes the likelihood covariance estimate towards the maximum of the prior, given in (11). To meet the goal of minimizing bias, we encode some coarse information about the data into B_h. In QB (Brown et al., 1999), the prior seed matrix $B_h = kI$, where k is a scalar determined by crossvalidation. Setting $B_h = kI$ is reminiscent of Friedman's RDA (Friedman, 1989), where the covariance estimate is regularized by the trace: $\frac{\operatorname{tr}(\hat{\Sigma}_{ML})}{d}I$.

We have shown in earlier work that setting $B_h = \frac{\operatorname{tr}(\hat{\Sigma}_{ML})}{d}I$ for a distribution-based discriminant analysis outperforms Geisser's or Keehn's parameter-based Bayesian discriminant methods, and does not require crossvalidation (Srivastava and Gupta, 2006). The trace of the ML covariance estimate is stable, and provides coarse information about the scale of the data samples. Thus, this proposed data-dependent prior can be interpreted as capturing the knowledge an application expert would have before seeing the actual data. There are many other approaches to data-dependent

priors, including hyperparameters and empirical Bayes methods (Efron and Morris, 1976; Haff, 1980; Lehmann and Casella, 1998). Data-dependent priors are also used to form proper priors that act similarly to improper priors, or to match frequentist goals (Wasserman, 2000). Next, we describe the closed form result with a prior of the form given in (10), then we return to the question of data-dependent definitions for B_h when we propose the BDA7 classifier in Section 6.

3.3 Closed-Form Result

In Theorem 1 we establish the closed-form result for the distribution-based Bayesian QDA classifier. The closed-form result for the parameter-based classifier with the same prior is presented after the theorem for comparison.

Theorem 1: The classifier (5) using the prior (10) can be written as

$$\hat{Y} = \underset{g=1,\ldots,G}{\operatorname{argmin}} \sum_{h=1}^{G} C(g,h) \frac{(n_h)^{\frac{d}{2}} \Gamma\left(\frac{n_h+q+1}{2}\right) \left|\frac{S_h+B_h}{2}\right|^{\frac{n_h+q}{2}}}{(n_h+1)^{\frac{d}{2}} \Gamma\left(\frac{n_h+q-d+1}{2}\right) |A_h|^{\frac{n_h+q+1}{2}}} \hat{P}(Y=h), \tag{12}$$

where $\hat{P}(Y=h)$ is an estimate of the class prior probability for class h, $\Gamma(\cdot)$ is the standard gamma function, and

$$A_h = \frac{1}{2}\left(S_h + \frac{n_h(x-\bar{x_h})(x-\bar{x_h})^T}{(n_h+1)} + B_h\right). \tag{13}$$

The proof of the theorem is given in Appendix A. The parameter-based Bayesian QDA class label using the prior given in (10) is

$$\hat{Y} = \underset{g=1,\ldots,G}{\operatorname{argmin}} \sum_{h=1}^{G} C(g,h) \frac{(n_h)^{\frac{d}{2}} \Gamma\left(\frac{n_h+q-d-1}{2}\right)}{(n_h+1)^{\frac{d}{2}} \Gamma\left(\frac{n_h+q-2d-1}{2}\right)} \frac{\left|\frac{S_h+B_h}{2}\right|^{\frac{n_h+q-d-2}{2}}}{|A_h|^{\frac{n_h+q-d-1}{2}}} \hat{P}(Y=h). \tag{14}$$

Equation (14) can be proved by following the same steps as the proof of Theorem 1. The parameter-based Bayesian discriminant result (14) will not hold if $n_h \leq 2d-q+1$, while the distribution-based result (12) holds for any $n_h > 0$ and any d.

In a previous publication (Srivastava and Gupta, 2006), we compared the distribution-based Bayesian QDA to the parameter-based Bayesian QDA classifier, maximum likelihood QDA, and nearest means. We compared the noninformative prior originally proposed by Geisser (1964), and the modified inverse Wishart prior given in Equation (10) with $B_h = \operatorname{tr}\left(\hat{\Sigma}_{ML,h}/d\right) I$. We fixed the degrees of freedom for the modified inverse Wishart to be $d+3$, so that in one-dimension it reduces to the common inverted gamma distribution. Results on six simulations (two-class versions of Friedman's original simulations) showed that given the same prior, the distribution-based performed better than the parameter-based when the ratio of feature dimensions to number of training samples is high. For the reader's convenience, we include a representative set of the simulation results from Srivastava and Gupta (2006) in Figure 1.

It is interesting to see that the only difference between the parameter-based Bayesian QDA (14) and the distribution-based Bayesian QDA (12) is a shift of the degree of freedom q; this is true whether the distribution-based formula (12) is solved for using either the Fisher or Lebesgue measure. QB, which is a modern parameter-based Bayesian QDA classifier (discussed further in Section

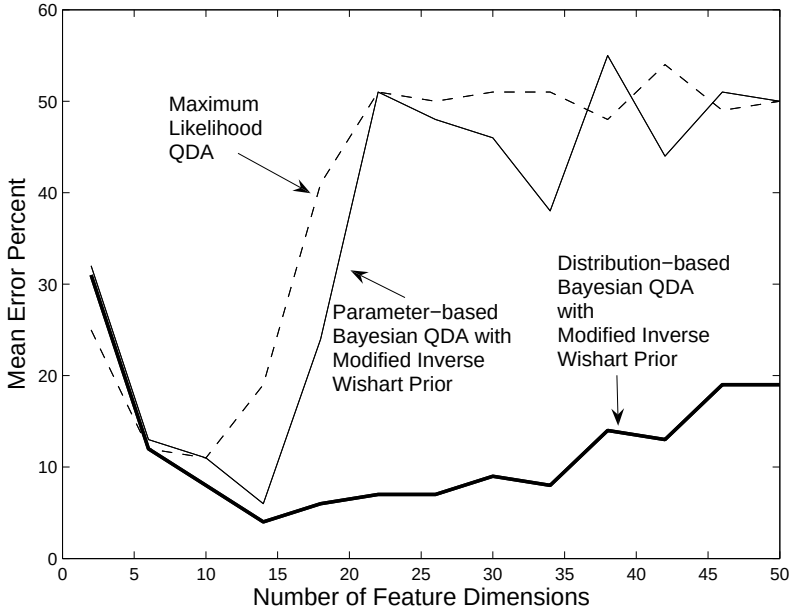

Figure 1: Mean error rates are shown for a two-class simulation where the samples from each class are drawn from a Gaussian distribution with the same mean but different, highly-ellipsoidal covariance matrices. The error rates were averaged over 1000 random trials, where on each trial 40 training samples and 100 test samples were drawn iid.

2.1), chooses the degree of freedom for the modified inverse Wishart prior by cross-validation. If one cross-validates the degree of freedom, then it does not matter if one starts from the parameter-based formula or the distribution-based formula. In Section 6, we propose a new Bayesian QDA classifier called *BDA7* that cross-validates the degree of freedom and a data-dependent seed matrix B_h for the prior. But first we consider further the analytic properties of Bayesian QDA; in the next section we develop a relationship between regularized QDA and Bayesian QDA, and then in Section 5 we show how Bayesian QDA is the solution to minimizing any expected Bregman divergence.

4. Relationship Between Regularized QDA and Bayesian QDA

In this section we show that Friedman's regularized form for the covariance matrix (Friedman, 1989) emerges from the Bayesian QDA formula.

Let $D_h = S_h + B_h, Z_h = x - \bar{x_h}$. The distribution-based Bayesian discriminant formula for the class conditional pdf (12) can be simplified to

$$\begin{aligned}
E_{N_h}[N_h] &= \frac{n_h^{\frac{d}{2}}\Gamma\left(\frac{n_h+q+1}{2}\right)}{(2\pi)^{\frac{d}{2}}(n_h+1)^{\frac{d}{2}}\Gamma\left(\frac{n_h+q-d+1}{2}\right)} \frac{\left|\frac{D_h}{2}\right|^{\frac{n_h+q}{2}}}{\left|\frac{D_h}{2}+\frac{n_h}{2(n_h+1)}Z_hZ_h^T\right|^{\frac{n_h+q+1}{2}}} \\
&= \frac{n_h^{\frac{d}{2}}\Gamma\left(\frac{n_h+q+1}{2}\right)\left|\frac{D_h}{2}\right|^{\frac{n_h+q}{2}}}{(2\pi)^{\frac{d}{2}}(n_h+1)^{\frac{d}{2}}\Gamma\left(\frac{n_h+q-d+1}{2}\right)\left|\frac{D_h}{2}\right|^{\frac{n_h+q+1}{2}}} \left|I+\frac{n_h}{n_h+1}Z_hZ_h^TD_h^{-1}\right|^{-\frac{n_h+q+1}{2}} \\
&= \frac{\Gamma\left(\frac{n_h+q+1}{2}\right)}{(\pi)^{\frac{d}{2}}\left|\left(\frac{n_h+1}{n_h}\right)D_h\right|^{\frac{1}{2}}\Gamma\left(\frac{n_h+q-d+1}{2}\right)}\left(1+\frac{n_h}{n_h+1}Z_h^TD_h^{-1}Z_h\right)^{-\frac{n_h+q+1}{2}}, \qquad (15)
\end{aligned}$$

where (15) follows by rearranging terms and applying the identity $|I + Z_hZ_h^TD_h^{-1}| = 1 + Z_h^TD_h^{-1}Z_h$ (Anderson, 2003).

Approximate $n_h/(n_h+1) \approx 1$ in (15). Recall the series expansion $e^r = 1 + r + r^2/2\ldots$, so if r is small, $1 + r \approx e^r$. We apply this approximation to the term $1 + Z_h^TD_h^{-1}Z_h$ in (15), and note that the approximation is better the closer the test point x is to the sample mean $\bar{x_h}$, such that Z_h is small. The approximation is also better the larger the minimum eigenvalue λ_{min} of D_h is, because of the bound $|Z_h^TD_h^{-1}Z_h| \leq \|Z_h\|^2/\lambda_{min}$. Then (15) becomes

$$\begin{aligned}
E_{N_h}[N_h] &\approx \frac{\Gamma\left(\frac{n_h+q+1}{2}\right)\left(\exp\left[\frac{n_h}{n_h+1}Z_h^TD_h^{-1}Z_h\right]\right)^{-\frac{n_h+q+1}{2}}}{(\pi)^{\frac{d}{2}}\left|\left(\frac{n_h+1}{n_h}\right)D_h\right|^{\frac{1}{2}}\Gamma\left(\frac{n_h+q-d+1}{2}\right)}, \\
&= \frac{\Gamma\left(\frac{n_h+q+1}{2}\right)\exp\left[-\frac{1}{2}Z_h^T\left[\frac{n_h+1}{n_h+q+1}\left(\frac{D_h}{n_h}\right)\right]^{-1}Z_h\right]}{(\pi)^{\frac{d}{2}}\left|\left(\frac{n_h+1}{n_h}\right)D_h\right|^{\frac{1}{2}}\Gamma\left(\frac{n_h+q-d+1}{2}\right)}. \qquad (16)
\end{aligned}$$

Let

$$\tilde{\Sigma}_h \triangleq \frac{n_h+1}{n_h+q+1}\frac{D_h}{n_h}. \qquad (17)$$

The approximation (16) resembles a Gaussian distribution, where $\tilde{\Sigma}_h$ plays the role of the covariance matrix. Rewrite (17),

$$\begin{aligned}
\tilde{\Sigma}_h &= \frac{n_h+1}{n_h+q+1}\left(\frac{S_h+B_h}{n_h}\right) \\
&= \frac{n_h+1}{n_h+q+1}\left(\frac{S_h}{n_h}\right)+\frac{n_h+1}{n_h+q+1}\left(\frac{B_h}{n_h}\right) \\
&= \left(1-\frac{q}{n_h+q+1}\right)\frac{S_h}{n_h}+\frac{1}{n_h+q+1}\left(\frac{n_h+1}{n_h}\right)B_h. \qquad (18)
\end{aligned}$$

In (18), make the approximation $\frac{n_h+1}{n_h} \approx 1$, then multiply and divide the second term of (18) by q,

$$\tilde{\Sigma}_h \approx \left(1-\frac{q}{n_h+q+1}\right)\frac{S_h}{n_h}+\left(\frac{q}{n_h+q+1}\right)\frac{B_h}{q}. \qquad (19)$$

The right-hand side of (19) is a convex combination of the sample covariance and the positive definite matrix $\frac{B_h}{q}$. This is the same general formulation as Friedman's RDA regularization (Friedman, 1989), re-stated in this paper in Equations (1) and (2). Here, the fraction $\frac{q}{n_h+q+1}$ controls the shrinkage of the sample covariance matrix toward the positive definite matrix $\frac{B_h}{q}$; recall from (11) that $\frac{B_h}{q}$ is the maximum of the prior probability distribution. Equation (19) also gives information about how the Bayesian shrinkage depends on the number of sample points from each class: fewer training samples n_h results in greater shrinkage towards the positive definite matrix $\frac{B_h}{q}$. Also, as the degree of freedom q increases, the shrinkage towards $\frac{B_h}{q}$ increases. However, as q increases, the shrinkage target $\frac{B_h}{q}$ moves towards the zero-matrix.

5. Bregman Divergences and Bayesian Quadratic Discriminant Analysis

In (5) we defined the Bayesian QDA class estimate that minimizes the expected misclassification cost. Then assuming a Gaussian class-conditional distribution, the expected class-conditional distribution is given in (12). A different approach to Bayesian estimation would be to estimate the hth class-conditional distribution to minimize some expected risk. That is, the estimated class-conditional distribution would be

$$\hat{f}_h = \operatorname*{argmin}_{f \in \mathcal{A}} \int_M R(\mathcal{N}_h, f) dM, \tag{20}$$

where $R(\mathcal{N}_h, f)$ is the risk of guessing f if the truth is $\mathcal{N}_h$, the set of functions $\mathcal{A}$ is defined more precisely shortly, and dM is a probability measure on the set of Gaussians, M. Equation (20) is a distribution-based version of the standard parameter Bayesian estimate given in Ch. 4 of Lehmann and Casella (1998); for example, the standard parameter Bayesian estimate of the mean $\hat{\mu} \in \mathbb{R}^d$ would be formulated

$$\hat{\mu} = \operatorname*{argmin}_{\psi \in \mathbb{R}^d} \int R(\mu, \psi) d\Lambda(\mu),$$

where $\Lambda(\mu)$ is some probability measure.

Given estimates of the class-conditional distributions $\{\hat{f}_h\}$ from (20), one can solve for the class label as

$$\tilde{Y}^* = \operatorname*{argmin}_{g=1,\ldots,G} \sum_{h=1}^{G} C(g,h)\, \hat{f}_h(x) \hat{P}(Y = h). \tag{21}$$

In this section we show that the class estimate $\hat{Y}$ from minimizing the expected misclassification cost as defined in (5) is equivalent to the class estimate $\tilde{Y}^*$ from (21) if the risk function in (20) is a (functional) Bregman divergence. This result links minimizing expected misclassification cost and minimizing an expected Bregman divergence.

Bregman divergences form a set of distortion functions that include squared error, relative entropy, logistic loss, Mahalanobis distance, and the Itakura-Saito function, and are sometimes termed *Bregman loss functions* (Censor and Zenios, 1997). Bregman divergences act on pairs of vectors. Csiszár defined a Bregman divergence between two distributions (Csiszár, 1995), but Csiszár's definition acts pointwise on the input distributions, which limits its usefulness in analysis. A recent result showed that the mean minimizes the average Bregman divergence (Banerjee et al., 2005a,b). In order to extend this result to distributions and show how it links to Bayesian estimation, one must solve for minima over sets of functions. To this end, we define a new *functional Bregman*

divergence that acts on pairs of distributions. This allows us to extend the Banerjee et al. result to the Gaussian case and establish the equivalence between minimizing expected misclassification cost and minimizing the expected functional Bregman divergence.

This section makes use of functional analysis and the calculus of variations; the relevant definitions and results from these fields are provided for reference in Appendix B.

Let ν be some measure, and define the set of functions $\mathcal{A}_p$ to be

$$\mathcal{A}_p = \left\{ a : \mathbb{R}^d \to [0,1] \;\middle|\; a \in L^p(\nu),\ a > 0,\ \|a\|_{L^p(\nu)} = 1 \right\}.$$

5.1 Functional Definition of Bregman Divergence

Let $\phi : \mathcal{A}_p \to \mathbb{R}$ be a continuous functional. Let $\delta^2\phi[f;\cdot,\cdot]$, the second variation of ϕ, be strongly positive. The functional Bregman divergence $d_\phi : \mathcal{A}_p \times \mathcal{A}_p \to [0,\infty)$ is defined as

$$d_\phi(f,g) = \phi[f] - \phi[g] - \delta\phi[g; f-g], \tag{22}$$

where $\delta\phi[g;\cdot]$ is the Fréchet derivative or first variation of ϕ at g.

Different choices of the functional ϕ will lead to different Bregman divergences. As an example, we present the ϕ for squared error.

5.2 Example (Squared error)

Let $\phi : \mathcal{A}_2 \to \mathbb{R}$ be defined

$$\phi[g] = \int g^2 d\nu.$$

Perturbing g by a sufficiently nice function a (see Appendix B for more details about the perturbation function a) leads to the difference

$$\phi[g+a] - \phi[g] = \int \left(g^2 + 2ga + a^2 - g^2\right) d\nu.$$

Then, because

$$\begin{aligned} \frac{\|\phi[g+a] - \phi[g] - \int 2ga\, d\nu\|_{L^2(\nu)}}{\|a\|_{L^2(\nu)}} &= \frac{\int a^2\, d\nu}{(\int a^2\, d\nu)^{\frac{1}{2}}}, \\ &= \left(\int a^2\, d\nu\right)^{\frac{1}{2}} \end{aligned}$$

tends to zero as $\|a\|_{L^2(\nu)}$ tends to zero, the differential is

$$\delta\phi[g;a] = \int 2ga\, d\nu. \tag{23}$$

Using (23), the Bregman divergence (22) becomes

$$\begin{aligned} d_\phi(f,g) &= \int f^2 d\nu - \int g^2 d\nu - \int 2g(f-g)d\nu \\ &= \int f^2 d\nu - \int 2fg d\nu + \int g^2 d\nu \\ &= \int (f-g)^2 d\nu \\ &= \|f-g\|^2_{L^2(\nu)}, \end{aligned}$$

which is the integrated squared error between two functions f and g in $\mathcal{A}_2$.

5.3 Minimizing Expected Bregman Divergence

The functional definition (22) is a generalization of the standard vector Bregman divergence

$$d_{\tilde{\phi}}(x,y) = \tilde{\phi}(x) - \tilde{\phi}(y) - \nabla\tilde{\phi}(y)^T(x-y)$$

where $x, y \in \mathbb{R}^n$, and $\tilde{\phi} : \mathbb{R}^n \to \mathbb{R}$ is strictly convex and twice differentiable.

Theorem 2: Let $\phi : \mathcal{A}_1 \to \mathbb{R}$, $\phi \in \mathcal{C}^3$, and $\delta^2\phi[f;\cdot,\cdot]$ be strongly positive. Suppose the function $\hat{f}_h$ minimizes the expected Bregman divergence between a random Gaussian N_h and any probability density function $f \in \mathcal{A}$ where the expectation is taken with respect to the distribution $r(\mathcal{N}_h)$, such that

$$\hat{f}_h = \underset{f\in\mathcal{A}}{\operatorname{argmin}} E_{N_h}[d_\phi(N_h, f)]. \tag{24}$$

Then $\hat{f}_h$ is given by

$$\hat{f}_h = \int_M \mathcal{N}_h \, r(\mathcal{N}_h) dM = E_{N_h}[N_h(x)].$$

The proof of the theorem is given in Appendix A.

Corollary: The result of (5) is equivalent to the result of (21) where each $\hat{f}_h$ comes from (24).

The corollary follows directly from Theorem 2 where $r(\mathcal{N}_h)$ is the posterior distribution of $\mathcal{N}_h$ given the training samples.

6. The BDA7 Classifier

Distribution-based QDA with a fixed degree of freedom as proposed by the authors in the conference paper (Srivastava and Gupta, 2006) does not require cross-validation. However, with cross-validation, one can generally do better if cross-validating a useful parameter. The question is, what parameters to cross-validate with what options? As done in the QB Bayesian QDA classifier, we believe that the degree of freedom of the prior should be cross-validated. Also, in preliminary experiments we found that the distribution-based performance could be enhanced by using the diagonal of $\hat{\Sigma}_{ML}$ rather than the trace for each prior seed matrix B_h; the diagonal encodes more information

but is still relatively stable to estimate. The diagonal of $\hat{\Sigma}_{ML}$ has been used to regularize a regularized discriminant analysis (Hoffbeck and Landgrebe, 1996) and model-based discriminant analysis (Bensmail and Celeux, 1996).

Note that setting $B_h = \text{diag}\left(\hat{\Sigma}_{\text{ML}}\right)$ places the maximum of the prior at $\frac{1}{q}\text{diag}\left(\hat{\Sigma}_{\text{ML}}\right)$. We hypothesize that in some cases it may be more effective to place the maximum of the prior at $\text{diag}\left(\hat{\Sigma}_{\text{ML}}\right)$; that requires setting $B_h = q\ \text{diag}\left(\hat{\Sigma}_{\text{ML}}\right)$.

We also consider moving the maximum of the prior closer to the zero matrix, effectively turning the prior into an exponential prior rather than a unimodal one. We expect that this will have the rough effect of shrinking the estimate toward zero. Shrinkage towards zero is a successful technique in other estimation scenarios: for example ridge and lasso regression shrink linear regression coefficients toward zero (Hastie et al., 2001), and wavelet denoising shrinks wavelet coefficients toward zero. To this end, we also consider setting the prior matrix seed to be $B_h = \frac{1}{q}\text{diag}\left(\hat{\Sigma}_{\text{ML}}\right)$.

These different choices for B_h will be more or less appropriate depending on the amount of data and the true generating distributions. Thus, for BDA7 the B_h is selected by crossvalidation from seven options:

$$B_h = \begin{cases} q\ \text{diag}\left(\hat{\Sigma}_{(\text{pooled ML})}\right), \text{q diag}\left(\hat{\Sigma}_{(\text{class ML,h})}\right), \\ \frac{1}{q}\ \text{diag}\left(\hat{\Sigma}_{(\text{pooled ML})}\right), \frac{1}{\text{q}}\ \text{diag}\left(\hat{\Sigma}_{(\text{class ML,h})}\right), \\ \text{diag}\left(\hat{\Sigma}_{(\text{pooled ML})}\right), \text{diag}\left(\hat{\Sigma}_{(\text{class ML,h})}\right), \\ \frac{1}{q}\,\text{tr}(\hat{\Sigma}_{(\text{pooled ML})})\, I. \end{cases} \tag{25}$$

To summarize: BDA7 uses the result (12) where q is cross-validated, and B_h is cross-validated as per (25).

7. Experiments with Benchmark Data Sets

We compared BDA7 to popular QDA classifiers on ten benchmark data sets from the UCI Machine Learning Repository. In the next section we use simulations to further analyze the behavior of each of the classifiers.

QDA classifiers are best-suited for data sets where there is relatively little data, assuming the class-conditional distribution is Gaussian model can be an appropriate model. For this reason, for data sets with separate training and test sets, Tables 3 and 4 show results on the test set given a randomly chosen 5% or 10% of the training samples. For data sets without separate training and test sets, 5% and 10% of the available samples of each class were randomly selected and used for training, and the rest of the available samples were used as test samples. Each result in Table 3 and 4 is the average test error of 100 trials with different randomly chosen training samples, with the exception of the Cover Type data set, for which only 10 random trials were performed due to its large size. Table 2 summarizes each of the benchmark data sets.

7.1 Experimental Details for Each Classifier

BDA7 is compared with QB (Brown et al., 1999), RDA (Friedman, 1989), eigenvalue decomposition discriminant analysis (EDDA) (Bensmail and Celeux, 1996), and maximum-likelihood estimated QDA, LDA, and the nearest-means classifier (NM). Code for the classifiers (and the simulations presented in the next section) is available at idl.ee.washington.edu.

The parameters for each classifier were estimated by leave-one-out crossvalidation, unless there exists a separate validation set. Some of the classifiers required the prior probability of each class

$P(Y = h)$; those probabilities were estimated based on the number of observations from each class using Bayesian estimation (see Section 3).

The RDA parameters λ and γ were calculated and crossvalidated as in Friedman's paper (Friedman, 1989) for a total of 25 joint parameter choices.

The BDA7 method is crossvalidated with the seven possible choices of B_h for the prior specified in (25). The scale parameter q of the inverse Wishart distribution is crossvalidated in steps of the feature space dimension d so that $q \in \{d, 2d, 3d, 4d, 5d, 6d\}$. Thus there are 42 parameter choices.

The QB method is implemented as described by Brown et al. (1999). QB uses a normal prior for each mean vector, and an inverse Wishart distribution prior for each covariance matrix. There are two free parameters to the inverse Wishart distribution: the scale parameter $q \geq d$ and the seed matrix B_h. For QB, B_h is restricted to be spherical: $B_h = kI$. The parameters q and k are trained by crossvalidation. In an attempt to be similar to the RDA and BDA7 crossvalidations, we allowed there to be 42 parameter choices, $q \in \{d, 2d, 3d, 4d, 5d, 6d\}$ and $k \in \{1, 2, \ldots 7\}$. Other standard Bayesian quadratic discriminant approaches use a uniform (improper) or fixed Wishart prior with a parameter-based classifier (Ripley, 2001; Geisser, 1964; Keehn, 1965); we have previously demonstrated that the inverse Wishart prior performs better than these other choices (Srivastava and Gupta, 2006).

The EDDA method of Bensmail and Celeux is run as proposed in their paper (Bensmail and Celeux, 1996). Thus, the crossvalidation selects one of fourteen models for the covariance, where the ML estimate for each model is used. Unfortunately, we found it computationally infeasible to run the 3rd and 4th model for EDDA for some of the problems. These models are computationally intensive iterative ML procedures that sometimes took prohibitively long for large n, and when $n < d$ these models sometimes caused numerical problems that caused our processor to exit with error. Thus, in cases where the 3rd and 4th model were not feasible, they were not used.

7.2 Results

The results are shown in Tables 3 and 4 for the UCI Machine Learning Repository benchmark data sets described in Table 2. The lowest mean error rate is in bold, as well as any other mean error rate that is not statistically significantly different from the lowest mean error rate, as determined by the Wilcoxon signed rank test for paired-differences at a significance level of .05.

BDA7 was the best or statistically insignificant from the best for seven of the ten tested data sets: Heart, Iris, Image Segmentation, Pen Digits, Sonar, Thyroid, and Wine. For the Pima Diabetes, Ionosphere, and Cover Type data sets, BDA7 is not the best QDA classifier, but performs competitively.

The Ionosphere data set is a two-class problem with 351 samples described by 34 features. One of the features has the value zero for all samples of class two. This causes numerical difficulties in estimating the maximum likelihood covariance estimates. BDA7 chooses the identity prior seed matrix $B_h = \frac{1}{q}\,\mathrm{tr}(\hat{\Sigma}_{(\text{pooled ML})})\,I$ every time for this data set. Given that this is BDA7's choice, BDA7 is at a disadvantage compared to QB, which cross-validates a scaled identity seed matrix kI. In fact, our experiments have showed that shrinking the prior closer to zero improves the performance on this data set, for example using $B_h = \frac{1}{qd}\,\mathrm{tr}(\hat{\Sigma}_{(\text{pooled ML})})\,I$.

For the Cover Type problem, the maximum likelihood estimated matrices for QDA and LDA contain many zeros, which result in a zero determinant; this is in part because 44 of the 54 features are binary. Thus the QDA and LDA error rate is solely due to the prior class probabilities. Because of the zero determinant, the best BDA7 model for the prior's seed matrix is the identity-based model,

	Number of Classes	Number of Features	Number of Total Samples (Training/Test) (Training/Validation/Test)
Cover Type	8	54	11,340/3,780/565,892
Heart	2	10 (nominal features excluded)	270
Image Segmentation	7	19	2310
Ionosphere	2	34	351
Iris	3	4	150
Pen Digits	10	16	7,494/3,498
Pima	2	8	768
Sonar	2	60	238
Thyroid	3	5	215
Wine	3	13	178

Table 2: Information About the Benchmark Data Sets

	BDA7	QB	RDA	EDDA	NM	LDA	QDA
Cover Type	48.3	**44.6**	59.2	83.1	78.5	62.9	62.9
Heart	**30.6**	38.5	38.2	33.9	39.9	35.9	44.5
Image Segmentation	**12.7**	**12.9**	18.3	29.7	29.5	85.7	85.7
Ionosphere	16.9	16.1	**12.5**	26.0	27.1	35.8	35.8
Iris	**6.9**	**7.6**	**8.1**	9.4	9.3	8.8	66.6
Pen Digits	**6.1**	8.1	8.5	7.7	23.9	17.8	23.8
Pima	29.7	32.7	29.4	30.7	35.8	**27.6**	33.4
Sonar	**36.8**	40.4	40.4	39.8	42.6	46.7	46.7
Thyroid	**11.7**	14.8	17.0	14.7	19.2	15.0	30.0
Wine	**9.6**	33.1	34.2	**11.2**	32.0	60.1	60.1

Table 3: Average Error Rate Using Randomly Selected 5% of Training Samples

	BDA7	QB	RDA	EDDA	NM	LDA	QDA
Cover Type	48.0	**43.7**	57.4	82.1	77.9	62.9	62.9
Heart	**27.4**	32.0	31.8	**28.3**	38.6	**28.1**	41.8
Image Segmentation	**10.9**	**10.8**	16.4	27.9	27.7	85.7	85.7
Ionosphere	12.5	11.1	**8.7**	23.3	24.2	35.9	35.9
Iris	**6.2**	**5.9**	**6.2**	7.4	8.3	**5.3**	44.9
Pen Digits	**5.3**	6.0	7.5	5.6	22.9	17.3	16.9
Pima	28.4	29.7	27.7	29.0	36.1	**25.8**	29.8
Sonar	**31.2**	33.7	32.8	34.8	38.8	46.8	46.8
Thyroid	**7.9**	9.1	10.0	**8.6**	16.2	10.5	30.0
Wine	**7.9**	16.9	25.0	**8.2**	30.1	21.6	60.3

Table 4: Average Error Rate Using Randomly Selected 10% of Training Samples

$B_h = \frac{1}{q}\,\mathrm{tr}(\hat{\Sigma}_{(\text{pooled ML})})\,I$. As in the Ionosphere data set, this puts BDA7 at a disadvantage compared to QB, and it is not surprising that on this data set QB does a little better. Despite the numerical difficulties inherent in this problem, the two Bayesian QDA classifiers do better than the other QDA classifiers.

8. Simulations

In order to further analyze the behavior of the different QDA classifiers, we compared them in ten simulations. For each of the ten simulations, the data are drawn iid from three Gaussian class conditional distributions. Six of the simulations were originally created by Friedman and used to show that RDA performs favorably compared to ML linear discriminant analysis (LDA) and ML QDA (Friedman, 1989). Friedman's six simulations all have diagonal generating class covariance matrices, corresponding to independent classification features. In those cases, the constrained diagonal models used in RDA and EDDA are correct, and so RDA and EDDA's performance may be optimistic compared to real data. For a fuller picture, we add two full covariance matrix simulations to Friedman's six diagonal Gaussian simulations (Cases 1–6), and for each of the full covariance matrix simulations we consider the cases of classes with the same means (Case 7 and 9), and classes with different means (Case 8 and 10).

All of the simulation results are presented for 40 training and 100 test samples, drawn iid. The parameters for each classifier were estimated by leave-one-out crossvalidation, as described in Section 7.1. Each simulation was run 100 times. Thus, each result presented in Tables 5, 6, and 7 is the average error over $10,000$ test samples.

8.1 Simulation Details and Results

The results show that when the true generating distributions are diagonal (Cases 1–6), then EDDA performs the best. This can be expected because many of the 14 models EDDA cross-validates are

diagonal, and are thus a good match to the true generating distributions. When the true generating distributions are drawn from full covariance matrices (Cases 7–10), EDDA performs well when there are only 10 features, but the maximum likelihood estimates used in EDDA fail for 50 and 100 feature dimensions. RDA similarly does well when the true generating distribution matches the RDA models, and like EDDA is unable to learn when the features are highly correlated (Cases 9–10).

BDA7 is rarely the best classifier on the simulations, but never fails to produce relatively reasonable error rates. In particular, BDA7 is shown to be a more robust classifier than QB, achieving much lower error rates for Cases 5–6, and lower error rates for Cases 2–4 (with the exception of Case 4 for 100 feature dimensions).

A more detailed analysis and description of each of the simulation cases follows.

8.1.1 CASE 1: EQUAL SPHERICAL COVARIANCE MATRICES

Each class conditional distribution is normal with identity covariance matrix I. The mean of the first class μ_1 is the origin, and the second class has zero mean, except that the first component of the second class mean is 3. Similarly, the third class has zero mean, except the last component of the third class mean is 3.

The performance here is bounded by the nearest-means classifier, which is optimal for this simulation. EDDA also does well, because one of the 14 models available for it to choose is exactly correct: scalar times the identity matrix. Similarly, RDA strongly shrinks towards the trace times the identity. Though this is an unrealistic case, it shows that BDA7 can perform relatively well even though it does not explicitly model the true generating distribution.

8.1.2 CASE 2: UNEQUAL SPHERICAL COVARIANCE MATRICES

The class one conditional distribution is normal with identity covariance matrix I and mean at the origin. The class two conditional distribution is normal with covariance matrix $2I$ and has zero mean except the first component of its mean is 3. The class three conditional distribution is normal with covariance matrix $3I$ and has zero mean except the last component of its mean is 4. Like Case 1, EDDA and RDA use the fact that the true generating distributions match their models to outperform BDA7, which is significantly better than QB.

8.1.3 CASES 3 AND 4: EQUAL HIGHLY ELLIPSOIDAL COVARIANCE MATRICES

Covariance matrices of each class distribution are the same, and highly ellipsoidal. The eigenvalues of the common covariance matrix are given by

$$e_i = \left(\frac{9(i-1)}{d-1}+1\right)^2, \quad 1 \le i \le d, \tag{26}$$

so the ratio of the largest to smallest eigenvalue is 100.

For Case 3 the class means are concentrated in a low-variance subspace. The mean of class one is located at the origin and the i^{th} component of the mean of class two is given by

$$\mu_{2i} = 2.5\sqrt{\frac{e_i}{d}}\frac{d-i}{(\frac{d}{2}-1)}, \quad 1 \le i \le d.$$

	BDA7	QB	RDA	EDDA	NM	LDA	QDA
Case 1	13.2	19.2	12.4	11.2	**9.7**	14.7	56.1
Case 2	21.3	27.4	17.4	**16.1**	21.0	26.0	54.6
Case 3	**10.4**	35.0	14.6	**9.1**	30.3	12.2	56.4
Case 4	**12.6**	32.8	**12.6**	**11.6**	**11.0**	13.5	56.9
Case 5	**4.1**	15.0	18.9	**4.4**	59.1	60.0	45.8
Case 6	5.2	7.9	7.2	**1.7**	36.1	36.9	47.7
Case 7	**19.5**	22.8	29.0	**19.7**	66.4	63.1	46.5
Case 8	**3.7**	**2.7**	**4.0**	5.1	40.6	6.7	37.5
Case 9	1.5	**0.9**	10.8	**1.0**	65.4	58.5	32.5
Case 10	0.4	**0.1**	2.8	3.4	60.7	7.2	36.5

Table 5: Average Error Rate for 10 Features

	BDA7	QB	RDA	EDDA	NM	LDA	QDA
Case 1	27.9	33.3	25.4	**21.7**	**21.5**	67.0	67.0
Case 2	26.8	42.6	15.4	**12.5**	32.0	66.8	66.8
Case 3	27.2	55.7	44.8	**21.2**	47.3	66.6	66.6
Case 4	22.5	30.9	17.9	17.0	**15.5**	67.0	67.0
Case 5	1.2	30.6	11.8	**0.0**	52.4	66.6	66.6
Case 6	0.5	26.5	8.0	**0.0**	43.2	66.6	66.6
Case 7	34.7	**30.9**	41.9	63.9	66.5	66.0	66.0
Case 8	9.2	**3.5**	**2.8**	25.5	40.9	66.6	66.6
Case 9	1.3	**0.9**	62.7	32.5	66.3	66.5	66.5
Case 10	1.7	**0.9**	60.6	32.4	66.2	66.2	66.2

Table 6: Average Error Rate for 50 Features

	BDA7	QB	RDA	EDDA	NM	LDA	QDA
Case 1	35.8	31.1	28.1	**24.8**	**24.7**	68.0	68.0
Case 2	20.8	41.9	11.1	**9.0**	37.2	66.8	66.8
Case 3	46.9	56.4	50.5	**27.7**	51.1	64.9	64.9
Case 4	37.6	32.1	23.9	**21.1**	**20.3**	66.7	66.7
Case 5	**0.2**	38.3	14.5	**0.1**	55.8	66.6	66.6
Case 6	**0.1**	29.4	7.2	**0.0**	40.9	67.3	67.3
Case 7	40.0	**35.2**	44.3	64.8	66.6	65.8	65.8
Case 8	17.3	**8.1**	**7.4**	55.2	54.1	66.3	66.3
Case 9	**2.9**	**2.8**	67.0	67.0	66.6	66.6	66.6
Case 10	**2.2**	**2.4**	64.0	64.0	66.3	65.9	65.9

Table 7: Average Error Rate for 100 Features

The mean of class three is the same as the mean of class two except every odd numbered dimension of the mean is multiplied by -1.

Here, BDA7 rivals the performance of EDDA, which makes half the errors of RDA. In contrast, the error rate of QB is twice as high as BDA7 for 10 and 50 dimensions.

Case 4 is that the class means are concentrated in a high-variance subspace. The mean of class one is again located at the origin and the i^{th} component of the mean of class two is given by

$$\mu_{2i} = 2.5\sqrt{\frac{e_i}{d}}\frac{i-1}{(\frac{d}{2}-1)}, \ 1 \leq i \leq d.$$

The mean of class three is the same as the mean of class two except every odd numbered dimension of the mean is multiplied by -1.

8.1.4 CASES 5 AND 6: UNEQUAL HIGHLY ELLIPSOIDAL COVARIANCE MATRICES

For these cases, the covariance matrices are highly ellipsoidal and different for each class. The eigenvalues of the class one covariance are given by Equation (26), and those of class two are given by

$$e_{2i} = \left(\frac{9(d-i)}{d-1}+1\right)^2, \ \ 1 \leq i \leq d.$$

The eigenvalues of class three are given by

$$e_{3i} = \left(\frac{9(i-\frac{d-1}{2})}{d-1}\right)^2, \ \ 1 \leq i \leq d.$$

For Case 5, the class means are identical. For Case 6 the class means are different, with the class one mean located at the origin and the i^{th} component of the class two mean given by $\mu_{2i} = \frac{14}{\sqrt{d}}$. The

mean of class three is the same as the mean of class two except every odd numbered dimension of the mean is multiplied by -1.

In both these cases the BDA7 error falls to zero as the number of feature dimensions rise, whereas RDA plateaus around 10% error, and QB has substantially higher error. Case 5 and Case 6 present more information to the classifiers than the previous cases because the covariance matrices are substantially different. BDA7 and EDDA are able to use this information effectively to discriminate the classes.

8.1.5 CASES 7 AND 8: UNEQUAL FULL RANDOM COVARIANCES

Let R_1 be a $d \times d$ matrix where each element is drawn independently and identically from a uniform distribution on $[0,1]$. Then let the class one covariance matrix be $R_1^T R_1$. Similarly, let the class two and class three covariance matrices be $R_2^T R_2$ and $R_3^T R_3$, where R_2 and R_3 are constructed in the same manner as R_1. For Case 7, the class means are identical. For Case 8, the class means are each drawn randomly, where each element of each mean vector is drawn independently and identically from a standard normal distribution.

Case 7 is a difficult case because the means do not provide any information and the covariances may not be very different. However, BDA7 and QB only lose classification performance slowly as the dimension goes up, while EDDA jumps from 20% error at 10 feature dimensions to 64% error at 50 dimensions. For most runs of this simulation, the best EDDA model is the full covariance model, but because EDDA uses ML estimation, its estimation of the full covariance model is ill-conditioned.

Case 8 provides more information to discriminate the classes because of the different class means. EDDA again does relatively poorly because its best-choice model is the full-covariance which it estimates with ML. QB and RDA do roughly equally as well, with BDA7 at roughly twice their error.

8.1.6 CASES 9 AND 10: UNEQUAL FULL HIGHLY ELLIPSOIDAL RANDOM COVARIANCE

Let R_1, R_2, R_3 be as described for Cases 7 and 8. Then the Cases 9 and 10 covariance matrices are $R_i R_i^T R_i^T R_i$ for $i = 1,2,3$. These covariance matrices are highly ellipsoidal, often with one strong eigenvalue and many relatively small eigenvalues. This simulates the practical classification scenario in which the features are all highly correlated. For Case 9, the class means are the same. For Case 10, the class means are each drawn randomly, where each element of the mean vector is drawn independently and identically from a standard normal distribution. The two Bayesian methods capture the information and achieve very low error rates. The other classifiers do not model this situation well, and have high error rates for 50 and 100 feature dimensions.

9. Conclusions

In this paper, we have shown how a distribution-based formulation of the Bayesian quadratic discriminant analysis classifier relates to the standard parameter-based formulation, established an analytic link between Bayesian discriminant analysis and regularized discriminant analysis, and presented a functional equivalence between minimizing the expected misclassification cost and minimizing the expected Bregman divergence of class conditional distributions. A side result was the establishment of a functional definition of the standard vector Bregman divergence.

The practical contribution of this paper is the classifier BDA7, which has been shown to perform generally better than RDA, EDDA and QB on ten benchmark data sets. Key aspects of BDA7 are that the seed matrix in the inverse Wishart prior uses a coarse estimate of the covariance matrix to peg the maximum of the prior to a relevant part of the distribution-space.

The simulations presented are helpful in analyzing the different classifiers. Comparisons on simulations show that RDA and EDDA perform well when the true Gaussian distribution matches one of their regularization covariance models (e.g., diagonal, identity), but can fail when the generating distribution has a full covariance matrix, particularly when features are correlated. In contrast, the Bayesian methods BDA7 and QB can learn from the rich differentiating information offered by full covariance matrices.

We hypothesize that better priors exist, and that such priors will also be data-dependent and make use of a coarse estimate of the covariance matrix for the prior. Although modeling each class by one Gaussian distribution has too much model bias to be a general purpose classifier, Gaussian mixture model classifiers are known to work well for a variety of problems. It is an open question as to how to effectively integrate the presented ideas into a mixture model classifier.

Acknowledgments

This work was funded in part by the United States Office of Naval Research. We thank Inderjit Dhillon and Richard Olshen for helpful discussions.

Appendix A.

This appendix contains the proofs of Theorem 1 and Theorem 2.

A.1 Proof of Theorem 1

The proof employs the following identities (Box and Tiao, 1973; Gupta and Nagar, 2000),

$$\int_{\mu} \exp\left[-\frac{n_h}{2}\operatorname{tr}\left(\Sigma^{-1}(\mu-\bar{x})(\mu-\bar{x})^T\right)\right] d\mu = \left(\frac{2\pi}{n_h}\right)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}, \tag{27}$$

$$\int_{\Sigma>0} \frac{1}{|\Sigma|^{\frac{q}{2}}} \exp[-\operatorname{tr}\left(\Sigma^{-1}B\right)] d\Sigma = \frac{\Gamma_d\left(\frac{q-d-1}{2}\right)}{|B|^{\frac{q-d-1}{2}}}, \tag{28}$$

where $\Gamma_d(\cdot)$ is the multivariate gamma function,

$$\Gamma_d(a) = \left(\Gamma\left(\frac{1}{2}\right)\right)^{\frac{d(d-1)}{d}} \prod_{i=1}^{d} \Gamma\left(a-\frac{d-i}{2}\right). \tag{29}$$

To expand (5), we first simplify (7). The posterior (8) requires calculation of the normalization constant α_h,

$$\alpha_h = \int_{\Sigma_h}\int_{\mu_h} \ell(\mathcal{N}_h, \mathcal{T}_h) p(\mathcal{N}_h) \frac{d\mu d\Sigma_h}{|\Sigma_h|^{\frac{d+2}{2}}}.$$

Substitute $\ell(\mathcal{N}_h, \mathcal{T}_h)$ from (9) and $p(\mathcal{N}_h)$ from (10),

$$\alpha_h = \int_{\Sigma_h}\int_{\mu_h} \frac{\exp[-\frac{n_h}{2}\operatorname{tr}\left(\Sigma_h^{-1}(\mu_h-\bar{x_h})(\mu_h-\bar{x_h})^T\right)]}{(2\pi)^{\frac{n_h d}{2}}|\Sigma_h|^{\frac{n_h+q}{2}}} \exp\left[-\frac{1}{2}\operatorname{tr}\left(\Sigma^{-1}(S_h+B_h)\right)\right] \frac{d\Sigma_h d\mu_h}{|\Sigma_h|^{\frac{d+2}{2}}}.$$

Integrate with respect to μ_h using identity (27):

$$\alpha_h = \frac{1}{(2\pi)^{\frac{n_h d}{2}}}\left(\frac{2\pi}{n_h}\right)^{\frac{d}{2}} \int_{\Sigma_h} \frac{\exp[-\frac{1}{2}\operatorname{tr}\left(\Sigma_h^{-1}(S_h+B_h)\right)]}{|\Sigma_h|^{\frac{n_h+q+d+1}{2}}} d\Sigma_h.$$

Next, integrate with respect to Σ_h using identity (28):

$$\alpha_h = \frac{1}{(2\pi)^{\frac{n_h d}{2}}}\left(\frac{2\pi}{n_h}\right)^{\frac{d}{2}} \frac{\Gamma_d\left(\frac{n_h+q}{2}\right)}{\left|\frac{S_h+B_h}{2}\right|^{\frac{n_h+q}{2}}}. \tag{30}$$

Therefore the expectation $E_{N_h}[N_h(x)]$ is

$$\begin{aligned} E_{N_h}[N_h(x)] &= \int_M \mathcal{N}_h(x) f(\mathcal{N}_h) dM_h \\ &= \frac{1}{\alpha_h(2\pi)^{\frac{n_h d}{2}}} \int_{\Sigma_h}\int_{\mu_h} \frac{\exp\left[-\frac{1}{2}\operatorname{tr}\left(\Sigma_h^{-1}(x-\mu_h)(x-\mu_h)^T\right)\right]}{(2\pi)^{\frac{d}{2}}|\Sigma_h|^{\frac{1}{2}}} \\ &\quad \cdot \frac{\exp\left[-\frac{1}{2}\operatorname{tr}\left(\Sigma_h^{-1}(S_h+B_h)\right)\right]}{|\Sigma_h|^{\frac{n_h+q}{2}}} \exp\left[-\frac{n_h}{2}\operatorname{tr}\left(\Sigma_h^{-1}(\mu_h-\bar{x_h})(\mu_h-\bar{x_h})^T\right)\right] \frac{d\mu_h d\Sigma_h}{|\Sigma_h|^{\frac{d+2}{2}}}. \end{aligned}$$

Integrate with respect to μ_h and Σ_h using identities (27) and (28), and equation (13) to yield

$$E_{N_h}[N_h(x)] = \frac{1}{\alpha_h(2\pi)^{\frac{dn_h}{2}}(n_h+1)^{\frac{d}{2}}} \frac{\Gamma_d\left(\frac{n_h+q+1}{2}\right)}{|A_h|^{\frac{n_h+q+1}{2}}},$$

where A_h is given by Equation (13). After substituting the value of α_h from (30),

$$E_{N_h}[N_h(x)] = \frac{1}{(2\pi)^{\frac{d}{2}}} \frac{n_h^{\frac{d}{2}}\Gamma_d\left(\frac{n_h+q+1}{2}\right)\left|\frac{S_h+B_h}{2}\right|^{\frac{n_h+q}{2}}}{(n_h+1)^{\frac{d}{2}}\Gamma_d\left(\frac{n_h+q}{2}\right)|A_h|^{\frac{n_h+q+1}{2}}}.$$

Simplify the multivariate gamma function Γ_d using (29):

$$E_{N_h}[N_h(x)] = \frac{1}{(2\pi)^{\frac{d}{2}}} \frac{n_h^{\frac{d}{2}}\Gamma\left(\frac{n_h+q+1}{2}\right)}{(n_h+1)^{\frac{d}{2}}\Gamma\left(\frac{n_h+q-d+1}{2}\right)} \frac{\left|\frac{S_h+B_h}{2}\right|^{\frac{n_h+q}{2}}}{|A_h|^{\frac{n_h+q+1}{2}}}. \tag{31}$$

Substituting (31) into (5) proves the theorem. Because $q \geq d$, the result (31) is valid for any $n_h > 0$ and any feature space dimension d.

□

A.2 Proof of Theorem 2

Let

$$\begin{aligned} J[f] &= E_{N_h}[d_\phi(N_h, f)] \\ &= \int_M d_\phi(\mathcal{N}_h, f) r(\mathcal{N}_h) dM \\ &= \int_M (\phi[\mathcal{N}_h] - \phi[f] - \delta\phi[f; \mathcal{N}_h - f]) r(\mathcal{N}_h) dM, \end{aligned} \tag{32}$$

where (32) follows by substituting the definition of functional Bregman divergence (22). Consider the increment

$$\begin{aligned} \triangle J[f,a] &= J[f+a] - J[f] \\ &= -\int_M (\phi[f+a] - \phi[f])\, r(\mathcal{N}_h) dM \\ &\quad -\int_M (\delta\phi[f+a; \mathcal{N}_h - f - a] - \delta\phi[f; \mathcal{N}_h - f])\, r(\mathcal{N}_h) dM, \end{aligned} \tag{33}$$

$$\tag{34}$$

where (34) follows from substituting (32) into (33). Using the definition of the differential of a functional (46), the first integrand in (34) can be written as

$$\phi[f+a] - \phi[f] = \delta\phi[f;a] + \varepsilon[f,a] \|a\|_{L^1(\nu)}. \tag{35}$$

Take the second integrand of (34), and subtract and add $\delta\phi[f; \mathcal{N}_h - f - a]$,

$$\begin{aligned} &\delta\phi[f+a; \mathcal{N}_h - f - a] - \delta\phi[f; \mathcal{N}_h - f] \\ &= \delta\phi[f+a; \mathcal{N}_h - f - a] - \delta\phi[f; \mathcal{N}_h - f - a] + \delta\phi[f; \mathcal{N}_h - f - a] - \delta\phi[f; \mathcal{N}_h - f] \\ &\overset{(a)}{=} \delta^2\phi[f; \mathcal{N}_h - f - a, a] + \varepsilon[f,a] \|a\|_{L^1(\nu)} + \delta\phi[f; \mathcal{N}_h - f] - \delta\phi[f;a] - \delta\phi[f; \mathcal{N}_h - f] \\ &\overset{(b)}{=} \delta^2\phi[f; \mathcal{N}_h - f, a] - \delta^2\phi[f; a, a] + \varepsilon[f,a] \|a\|_{L^1(\nu)} - \delta\phi[f;a], \end{aligned} \tag{36}$$

where (a) follows from the definition of the second differential (see (47) in Appendix B) and from the fact that

$$\delta\phi[f+a;a] - \delta\phi[f;a] = \delta^2\phi[f;a,a] + o(\|a\|^2_{L^1(\nu)}), \tag{37}$$

where $o(\|a\|^2_{L^1(\nu)})$ denotes a function which goes to 0 even if it is divided by $\|a\|^2_{L^1(\nu)}$; and (b) follows from linearity of the first term. Substitute (35) and (36) into (34),

$$\triangle J[f,a] = -\int_M \left(\delta^2\phi[f; \mathcal{N}_h - f, a] - \delta^2\phi[f;a,a] + \varepsilon[f,a] \|a\|_{L^1(\nu)}\right) r(\mathcal{N}_h) dM.$$

Note that the term $\delta^2\phi[f;a,a]$ is of order $\|a\|^2_{L^1(\nu)}$, that is, $\left\|\delta^2\phi[f;a,a]\right\|_{L^1(\nu)} \leq m \|a\|^2_{L^1(\nu)}$ for some constant m. Therefore,

$$\lim_{\|a\|_{L^1(\nu)} \to 0} \frac{\|J[f+a] - J[f] - \delta J[f;a]\|_{L^1(\nu)}}{\|a\|_{L^1(\nu)}} = 0,$$

where

$$\delta J[f;a] \quad = \quad -\int_M \delta^2\phi[f;\mathcal{N}_h - f, a] r(\mathcal{N}_h) dM. \tag{38}$$

For fixed a, $\delta^2\phi[f;\cdot,a]$ is a bounded linear functional in the second argument, so the integration and the functional can be interchanged in (38). Thus the above equation becomes

$$\delta J[f;a] = -\delta^2\phi\left[f; \int_M (\mathcal{N}_h - f)\, r(\mathcal{N}_h) dM, a\right].$$

Here we tacitly assumed that both $\int_M r(\mathcal{N}_h) dM$ and $\int_M \mathcal{N}_h\, r(\mathcal{N}_h) dM$ exist. Under the functional optimality conditions (stated in Appendix B), $J[f]$ has an extremum for $f = \hat{f}$ if, for all admissible $a \in \mathcal{A}_1$,

$$\delta^2\phi\left[f; \int_M \left(\mathcal{N}_h - \hat{f}\right) r(\mathcal{N}_h) dM, a\right] = 0. \tag{39}$$

Set $a = \int_M \left(\mathcal{N}_h - \hat{f}\right) r(\mathcal{N}_h) dM$ in (39) and use the assumption that the functional $\delta^2\phi[f;\cdot,\cdot]$ is strongly positive, which implies that the above functional can be zero if and only if $a = 0$, that is,

$$0 \quad = \quad \int_M (\mathcal{N}_h - \hat{f}) r(\mathcal{N}_h) dM, \tag{40}$$

$$\hat{f} \quad = \quad E_{N_h}[N_h]. \tag{41}$$

Equation (41) is well-defined with the measure given in (6). Because a Bregman divergence is not necessarily convex in its second argument, it is not yet established that the above unique extremum is a minimum. To show that (41) is in fact a minimum of J, from the functional optimality conditions it is enough to show that $\delta^2 J[\hat{f};\cdot,\cdot]$ is strongly positive. To show this, for $b \in \mathcal{A}_1$, consider $\delta J[f+b;a] - \delta J[f;a]$ to be the increment in the first differential of J at f. Then

$$\begin{aligned}
&\delta J[f+b;a] - \delta J[f;a] \\
&\quad \overset{(c)}{=} -\int_M \left(\delta^2\phi[f+b;\mathcal{N}_h - f - b, a] - \delta^2\phi[f;\mathcal{N}_h - f, a]\right) r(\mathcal{N}_h) dM \\
&\quad \overset{(d)}{=} -\int_M (\delta^2\phi[f+b;\mathcal{N}_h - f - b, a] - \delta^2\phi[f;\mathcal{N}_h - f - b, a] + \delta^2\phi[f;\mathcal{N}_h - f - b, a] \\
&\qquad -\delta^2\phi[f;\mathcal{N}_h - f, a]) r(\mathcal{N}_h) dM \\
&\quad \overset{(e)}{=} -\int_M (\delta^3\phi[f;\mathcal{N}_h - f - b, a, b] + \varepsilon[f,a,a,b]\, \|b\|_{L^1(\nu)} + \delta^2\phi[f;\mathcal{N}_h - f, a] - \delta^2\phi[f;b,a] \\
&\qquad -\delta^2\phi[f;\mathcal{N}_h - f, a]) r(\mathcal{N}_h) dM \\
&\quad \overset{(f)}{=} -\int_M (\delta^3\phi[f;\mathcal{N}_h - f, a, b] - \delta^3\phi[f;b,a,b] + \varepsilon[f,a,a,b]\, \|b\|_{L^1(\nu)} \\
&\qquad -\delta^2\phi[f;b,a]) r(\mathcal{N}_h) dM,
\end{aligned} \tag{42}$$

where (c) follows from using integral (38); (d) from subtracting and adding $\delta^2\phi[f;\mathcal{N}_h - f - b, a]$; (e) from using the definition of the second differential and from the fact (37); and (f) follows from the linearity of the first term and cancelation of the third and fifth terms. Note that in (42), for fixed

a, the term $\delta^3\phi[f;b,a,b]$ is of order $\|b\|^2_{L^1(\nu)}$, while the first and the last terms are of order $\|b\|_{L^1(\nu)}$. Therefore,

$$\lim_{\|b\|_{L^1(\nu)}\to 0} \frac{\left\|\delta J[f+b;a]-\delta J[f;a]-\delta^2 J[f;a,b]\right\|_{L^1(\nu)}}{\|b\|_{L^1(\nu)}} = 0,$$

where

$$\delta^2 J[f;a,b] = -\int_M (\delta^3\phi[f;\mathcal{N}_b - f,a,b]-\delta^2\phi[f;b,a])r(\mathcal{N}_b)dM. \tag{43}$$

Substitute $b=a$ and $f=\hat{f}$, and interchange integration and the functional $\delta^3\phi$ in the first integral of (43) to yield

$$\begin{aligned}
\delta^2 J[\hat{f};a,a] &= -\delta^3\phi\left[\hat{f};\int_M(\mathcal{N}_b-\hat{f})r(\mathcal{N}_b)dM,a,a\right]+\int_M\delta^2\phi[\hat{f};a,a]r(\mathcal{N}_b)dM \\
&= \int_M \delta^2\phi[\hat{f};a,a]r(\mathcal{N}_b)dM \qquad (44)\\
&\geq \int_M k\|a\|^2_{L^1(\nu)}\, r(\mathcal{N}_b)dM \\
&= K\|a\|^2_{L^1(\nu)} > 0, \qquad (45)
\end{aligned}$$

where (44) follows from (40), and (45) follows from the strong positivity of $\delta^2\phi[\hat{f};\cdot,\cdot]$ and $K \stackrel{\triangle}{=} k\int_M dM > 0$. Therefore from (45) and the functional optimality conditions, $\hat{f}$ is the minimum. The result $\hat{f}$ is given in (31).

□

Appendix B. Relevant Definitions and Results from Functional Analysis

The following survey of the relevant definitions and results from functional analysis is based on Gelfand and Fomin (2000). Let a function space $\mathcal{A}_1$ be defined

$$\mathcal{A}_1 = \left\{ a:\mathbb{R}^d \to [0,1] \;\middle|\; a\in L^1(\nu),\ a>0,\ \|a\|_{L^1(\nu)}=1 \right\}$$

where ν is some measure. Note that $\mathcal{A}_1$ is a convex subspace of $L^1(\nu)$: if $a_1, a_2 \in \mathcal{A}_1$ and $0\leq\omega\leq 1$, then $\omega a_1 + (1-\omega)a_2 \in \mathcal{A}_1$.

B.1 Definition of Continuous Linear Functionals

The functional $\psi : L^1(\nu)\to\mathbb{R}$ is linear and continuous if

1. $\psi[\omega a_1 + a_2] = \omega\psi[a_1]+\psi[a_2]$ for any $a_1, a_2 \in L^1(\nu)$ and any real number ω (linearity); and
2. there is a constant C such that $|\psi[a]|\leq C\|a\|$ for all $a\in L^1(\nu)$.

B.2 Functional Derivatives

1. Let ϕ be a real functional over the normed space $L^1(\nu)$. The bounded linear functional $\delta\phi[f;\cdot]$ is the Fréchet derivative of ϕ at $f \in L^1(\nu)$ if

$$\begin{aligned}\phi[f+a]-\phi[f] &= \triangle\phi[f;a] \\ &= \delta\phi[f;a]+\varepsilon[f,a]\,\|a\|_{L^1(\nu)} \end{aligned} \tag{46}$$

   for all $a \in L^1(\nu)$, with $\varepsilon[f,a] \to 0$ as $\|a\|_{L^1(\nu)} \to 0$.

2. When the second variation $\delta^2\phi$ and the third variation $\delta^3\phi$ exist, they are described by

$$\begin{aligned}\triangle\phi[f;a] &= \delta\phi[f;a]+\frac{1}{2}\delta^2\phi[f;a,a]+\varepsilon[f,a]\,\|a\|^2_{L^1(\nu)} \\ &= \delta\phi[f;a]+\frac{1}{2}\delta^2\phi[f;a,a]+\frac{1}{6}\delta^3\phi[f;a,a,a]+\varepsilon[f,a]\,\|a\|^3_{L^1(\nu)},\end{aligned}$$

   where $\varepsilon[f,a] \to 0$ as $\|a\|_{L^1(\nu)} \to 0$. The term $\delta^2\phi[f;a,b]$ is bilinear with respect to arguments a and b, and $\delta^3\phi[f;a,b,c]$ is trilinear with respect to a,b, and c.

3. Suppose $\{a_n\},\{f_n\} \subset L^1(\nu)$, moreover $a_n \to a$, $f_n \to f$,where $a, f \in L^1(\nu)$. If $\phi \in \mathcal{C}^3(L^1(\nu);\mathbb{R})$ and $\delta\phi[f;a]$, $\delta^2\phi[f;a,a]$, and $\delta^3[f;a,a,a]$ are defined as above, then $\delta\phi[f_n;a_n] \to \delta\phi[f;a]$, $\delta^2\phi[f_n;a_n,a_n] \to \delta^2\phi[f;a,a]$, and $\delta^3\phi[f_n;a_n,a_n,a_n] \to \delta^3\phi[f;a,a,a]$, respectively.

4. The functional $\delta^2\phi[f;a,a]$ defined on normed linear space $L^1(\nu)$ is quadratic; it is also ***strongly positive*** if there exists a constant $k>0$ such that $\delta^2\phi[f;a,a] \geq k\,\|a\|^2_{L^1(\nu)}$ for all $a \in \mathcal{A}_1$. In a finite-dimensional space, strong positivity of a quadratic form is equivalent to the quadratic form being positive definite.

B.3 Functional Optimality Conditions

A necessary condition for the differential functional $J[f]$ to have an extremum (minimum) at $f=\hat{f}$ is that $\delta J[f;a]=0$ and $\delta^2 J[f;a,a] \geq 0$, for $f=\hat{f}$ and all admissible functions $a \in \mathcal{A}_1$. A sufficient condition for a functional $J[f]$ to have a minimum for $f=\hat{f}$ is that the first variation $\delta J[f;a]$ vanishes for $f=\hat{f}$, and its second variation $\delta^2 J[f;a,a]$ is strongly positive for $f=\hat{f}$.

References

S. Amari and H. Nagaoka. *Methods of Information Geometry*. American Mathematical Society, USA, 2000.

T. W. Anderson. *An Introduction to Multivariate Statistical Analysis*. Wiley-Interscience, USA, 2003.

A. Banerjee, X. Guo, and H. Wang. On the optimality of conditional expectation as a Bregman predictor. *IEEE Trans. on Information Theory*, 51(7):2664–2669, 2005a.

A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh. Clustering with Bregman divergences. *Journal of Machine Learning Research*, 6:1705–1749, 2005b.

H. Bensmail and G. Celeux. Regularized Gaussian discriminant analysis through eigenvalue decomposition. *Journal of the American Statistical Association*, 91:1743–1748, 1996.

P. J. Bickel and B. Li. Regularization in statistics. *Test*, 15(2):271–344, 2006.

G. E. P. Box and G. C. Tiao. *Bayesian Inference in Statistical Analysis*. Addison-Wesley, Reading, Massachusetts, 1973.

P. J. Brown, T. Fearn, and M. S. Haque. Discrimination with many variables. *Journal of the American Statistical Association*, 94(448):1320–1329, 1999.

S. Censor and Y. Zenios. *Parallel Optimization: Theory, Algorithms, and Applications*. Oxford University Press, Oxford, England, 1997.

I. Csiszár. Generalized projections for non-negative functions. *Acta Mathematica Hungarica*, 68: 161–185, 1995.

P. S. Dwyer. Some applications of matrix derivatives in multivariate analysis. *Journal of the American Statistical Association*, 333:607–625, 1967.

B. Efron and C. Morris. Multivariate empirical Bayes and estimation of covariance matrices. *The Annals of Statistics*, 4:22–32, 1976.

J. H. Friedman. Regularized discriminant analysis. *Journal of the American Statistical Association*, 84(405):165–175, 1989.

S. Geisser. Posterior odds for multivariate normal distributions. *Journal of the Royal Society Series B Methodological*, 26:69–76, 1964.

S. Geisser. *Predictive Inference: An Introduction*. Chapman & Hall, New York, 1993.

I. M. Gelfand and S. V. Fomin. *Calculus of Variations*. Dover, USA, 2000.

A. K. Gupta and D. K. Nagar. *Matrix Variate Distributions*. Chapman and Hall/CRC, Florida, 2000.

L. R. Haff. Empirical Bayes estimation of the multivariate normal covariance matrix. *The Annals of Statistics*, 8(3):586–597, 1980.

T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning*. Springer-Verlag, New York, 2001.

J. P. Hoffbeck and D. A. Landgrebe. Covariance matrix estimation and classification with limited training data. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18:763–767, 1996.

E. T. Jaynes and G. T. Bretthorst. *Probability Theory: the Logic of Science*. Cambridge University Press, Cambridge, 2003.

R. E. Kass. The geometry of asymptotic inference. *Statistical Science*, 4(3):188–234, 1989.

D. G. Keehn. A note on learning for Gaussian properties. *IEEE Trans. on Information Theory*, 11: 126–132, 1965.

E. L. Lehmann and G. Casella. *Theory of Point Estimation*. Springer, New York, 1998.

S. J. Raudys and A. K. Jain. Small sample size effects in statistical pattern recognition: Recommendations for practitioners. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 13(3): 252–264, 1991.

B. Ripley. *Pattern Recognition and Neural Nets*. Cambridge University Press, Cambridge, 2001.

S. Srivastava and M. R. Gupta. Distribution-based Bayesian minimum expected risk for discriminant analysis. *Proc. of the IEEE Intl. Symposium on Information Theory*, pages 2294–2298, 2006.

D. L. Swets and J. Weng. Using discriminant eigenfeatures for image retrieval. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(8):831–836, 1996.

L. Wasserman. Asymptotic inference of mixture models using data-dependent prior. *Journal of the Royal Statistical Society Series B*, 62(1):159–180, 2000.

J. Ye. Characterization of a family of algorithms for generalized discriminant analysis of undersampled problems. *Journal of Machine Learning Research*, 6:483–502, 2005.