Bayesian Minimum Expected Risk Estimation of Distributions for Statistical Learning

Santosh Srivastava

A dissertation submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

University of Washington

2007

Program Authorized to Offer Degree: Applied Mathematics

University of Washington
Graduate School

This is to certify that I have examined this copy of a doctoral dissertation by

Santosh Srivastava

and have found that it is complete and satisfactory in all respects,
and that any and all revisions required by the final
examining committee have been made.

Chair of the Supervisory Committee:

_____

Maya Gupta, Department of Electrical Engineering

Reading Committee:

_____

Maya Gupta, Department of Electrical Engineering

_____

K K Tung, Department of Applied Mathematics

_____

Marina Meila, Department of Statistics

Date: _____

In presenting this dissertation in partial fulfillment of the requirements for the doctoral degree at the University of Washington, I agree that the Library shall make its copies freely available for inspection. I further agree that extensive copying of this dissertation is allowable only for scholarly purposes, consistent with "fair use" as prescribed in the U.S. Copyright Law. Requests for copying or reproduction of this dissertation may be referred to Proquest Information and Learning, 300 North Zeeb Road, Ann Arbor, MI 48106-1346, 1-800-521-0600, to whom the author has granted "the right to reproduce and sell (a) copies of the manuscript in microform and/or (b) printed copies of the manuscript made from microform."

Signature_____

Date_____

University of Washington

Abstract

Bayesian Minimum Expected Risk Estimation of Distributions for Statistical Learning

Santosh Srivastava

Chair of the Supervisory Committee:
Maya Gupta, Department of Electrical Engineering

In this thesis, the principle of Bayesian estimation is applied directly to distributions such that the estimated distribution minimizes the expectation of some risk that is a functional of the distribution itself. Bregman divergence is considered as a risk function. An analysis of distribution-based Bayesian quadratic discriminant analysis (QDA) is presented, and a relationship is shown between the proposed approach and an existing regularized quadratic discriminant analysis approach. A functional definition of Bregman divergence is established and it is shown that Bayesian models are optimal in the expected functional Bregman divergence sense. Based on this analysis two practical classifiers are proposed. BDA7 uses a crossvalidated data dependent prior. Local BDA is a modification of Bayesian QDA to achieve flexible model-based classification, by restricting the inference to local neighborhoods of k samples from each class that are closest to the test sample.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

It is a pleasure to thank the many people who made this thesis possible.

It is difficult to overstate my gratitude to my Ph.D. supervisor, Prof. Maya Gupta. With her enthusiasm, her inspiration, and her great efforts to explain things clearly and simply, she helped to make statistics and machine learning fun for me. I owned my statistics knowledge to her. Throughout my study, she provided encouragement, sound advice, good teaching, good company, and lots of good ideas. I would have been lost without her. I thank my committee members, Prof. K K Tung, Prof. Marian Melia, Prof. Hong Qian and Prof. Mari Ostendorf for their helpful suggestions, feedbacks, ideas, and comments during my study

I am grateful to the applied mathematic department, for helping and assisting me in many different ways. Prof. R. O. Malley, and Prof. Mark Kot deserve very special mention. My colleagues from the Information Design Lab supported me in my research work. I want to thank them for all their help, support, interest and valuable hints. I wish to thank my best friends: Tracie Bartlebagh, Collen Altstock , Hilit Kletter, Anupam Sharma, Miguel Gomez, Bela Frigyik, Apurva Mishra, Hemant Kumar, Manish Tiwari, Sabrina Moss, Honora Lo, and Minna Kovanen for helping me get through the difficult times, and for all the emotional support, comraderie, entertainment, and caring they provided.

Especially, I would like to give my special thanks to a very beautiful girl Mariana Lopez whose patient love enabled me to complete this work. Lastly, and most importantly, I wish to thank my parents, and relatives. They bore me, raised me, supported me, taught me, and loved me. To them I dedicate this thesis.

Chapter 1

INTRODUCTION

As Tom Mitchell noted [1], "A scientific field is best defined by the central question it studies".

The central question of statistical classification is:

"*How to estimate probabilities of different class labels for a test sample given a set of labeled samples to learn from*?"

These questions cover a broad range of learning and classification tasks, such as how to design autonomous mobile robots that can train themselves from self-collected data, how to data mine historical medical records to learn which future patients will respond best to which treatments, and how to build search engines that automatically customize to user interests [1].

Many concepts and techniques in machine and statistical learning are illuminated by human and animal learning in psychology, neuroscience and related fields. The questions of how computers can learn and how humans learn most probably have highly intertwined answers. Human's ability to learn is a hallmark of intelligence. For example, in the field of visual category recognition humans can easily distinguish 30,000 or so categories and can be trained with very few examples, while the machine learning approach to digits and faces currently requires hundreds if not thousands of examples. Nevertheless as computers become more and more powerful, the idea that computers can imitate human learning is no longer science fiction. In fact, there has been a surge of interest to study machine and statistical learning paradigms that parallel human learning processes, such as efficient knowledge representation and visual recognition. These techniques have greatly influenced

the development of more intelligent computer interfaces that can recognize objects, understand human languages, predict weather and traffic, diagnose diseases, automatically sort letters containing hand written addresses in US post office, detect fraudulent financial transactions, learn models of gene expression in cells from high-throughput data, and even play chess or drive robots autonomously.

1.1 Challenges in Statistical Learning

The massive collections of data along with many new scientific problems create golden opportunities and significant challenges and has reshaped statistical thinking, data analysis, and theoretical studies. The challenges of high-dimensionality arise in diverse fields of sciences and the humanities, ranging from statistics, computational biology and health studies to financial engineering and risk management. High-dimensionality has significantly challenged traditional statistical theory and the intensive computational costs make traditional procedures infeasible for high-dimensional data analysis. As Donoho said [2] "many new insights need to be unveiled and many new phenomena need to be discovered in high dimensional data analysis, and it will be the most important research topic in machine learning and statistics in the 21st century."

As pointed out by Fan and Li [3], to optimize the performance of a portfolio or to manage the risk of portfolio, one needs to estimate the "covariance matrix of the returns of assets in the portfolio." Estimating covariance matrices in high-dimensional statistical problems poses challenges. Covariance matrices pervade every facet of statistical learning, from density estimation, to graphical models. They are also critical for studying genetic networks, as well as other statistical applications such as climatology.

There are two broad approaches to tackle problems when the dimension of the variables is comparable with the sample size. One approach to dimension reduction that is common in machine learning and data mining is to select reliable variables to minimize risk of prediction. Another approach is to employ a regularization method. Regularization is the class of methods that reduce estimation variance and can be used to modify maximum likelihood to give reasonable answers in unstable situations. Regularization techniques have been highly successful in the solution of ill-and poorly-posed inverse problems. Regularization is further

discussed in Chapter 5.

In this thesis a special case of the learning process is considered which is the supervised learning framework for classification. In this framework, the data consists of instance-label or feature-label $\{X_i, Y_i\}_{i=1}^n$ pairs, where the labels are $Y_i \in \{1, 2, 3, \ldots, G\}$. Given a set of such pairs, a learning algorithm constructs a function that maps instances to labels. This function should be such that it makes few errors when predicting the labels of unseen instances. For example in a wine classification problem [4], data consists of different wine samples made from the Pinot Noir (Burgundy) grapes. The wines are subjected to taste tests by 16 judges and graded with numerical scores on 14 sensory characteristics, which define a feature vector. These characteristics or features are clarity, color, aroma intensity, aroma character, undesirable odor, acidity, sugar, body, flavor intensity, flavor character, oakiness, astringency, undesirable taste, and overall quality. These wines originate from three different geographical regions, which defines the class label Y of the wine: California, Pacific Northwest, and France. The goal of supervised learning algorithm is to classify the geographical origin of the unseen wine sample x from 14 sensory characteristics.

1.2 Examples of Supervised Algorithms: k-NN, SVM

A variety of supervised machine learning algorithms have been studied in the past including k-Nearest-Neighbor (k-NN), support vector machine (SVM).

1.2.1 k-Nearest-Neighbor (k-NN) Classifiers,

These classifiers are *memory-based*, and require no model to be fit [5]. Given a query point or test point x, it finds the k training points $X_{(r)}$, $r = 1, \ldots, k$ closest in distance to x, and then classifies x using the majority vote among the k neighbors. Despite its simplicity, k-NN has been successful in a large number of classification problems, including handwritten digits, satellite image scenes, and EKG patterns.

1.2.2 Support Vector Machine (SVM)

Support vector machines (SVMs) are a useful classification method. The goal of the support vector machine (SVM) is to find the separating hyperplane in the input space with the largest *margin* [5, 6, 7]. It is based on the idea that the larger the *margin*, the better the generalization of the classifier. The *margin* of SVM has a nice geometric interpretation: it is defined informally as (twice) the smallest Euclidean distance between the decision surface and the closet training point. Non-linear SVMs usually use the "kernel trick" to first map the input space into a higher-dimension feature space with some non-linear transformation and build a maximum-margin hyperplane there. The "trick" is that this mapping is never computed directly, but implicity induced by a kernel. Support vector machines (SVM) were originally designed for binary classification and how to effectively extend it for multi-class classification is still an on-going research issue [8, 9].

1.3 Contributions and Organization

This dissertation makes contribution to the problem of statistical learning from the following aspects

- The theoretical contributions of this dissertation is that we defining and establishing functional Bregman divergence. It relates to square difference, square bias, and relative entropy. We have shown that functional Bregman divergence for functions and distributions generalizes vector and point-wise definitions of Bregman divergence. We extended Banerjee et. al.'s work to show that the mean of a set of functions minimizes the expected functional Bregman divergence. Furthermore, we extended Bayesian estimation using Bregman divergence risk function, and showed that using functional Bregman divergence one can directly estimate the underlying distribution instead of the parameters of the distribution.

- Application-wise, this dissertation gives an overview of the regularization of statistical learning problem in high dimensional feature space. The main research contribution from this dissertation is the Bayesian quadratic discriminant analysis for classification

and establishing its link to regularization.

- The algorithmic contributions come from the development of novel regularized data adaptive algorithms called BDA7 and local BDA for pattern classification tasks. Both of these algorithms are derived using inverted Wishart prior distribution and the Fisher information measure over the statistical manifold of Gaussian distributions. These algorithms perform remarkably well on a wide range of real datasets compared to other state-of-the-art classifiers in the literature.

The rest of the dissertation is organized as:

Chapter 2 reviews the basic principles of estimation including maximum likelihood (ML), maximum a posteriori (MAP), method of moments (MOM), and Bayesian mean square error estimation (BMSEE). It introduces the concept of Bregman divergence risk function for Bayesian estimation and shows that the mean of the posterior pmf is an optimal estimator for any Bregman divergence risk.

Chapter 3 discusses nearest-neighbor classifier model of constant class probabilities in the neighborhood of the test sample. A generalized form of Laplace smoothing for weighted k nearest-neighbors class probability estimates is derived, and it is shown that it is optimal in the sense of minimizing any expected Bregman divergence and leads to the class estimates that minimize expected misclassification cost.

Chapter 4 explains the theory of Bayesian quadratic discriminant analysis in many aspects. The Bayesian classifier is solved in terms of Gaussian distribution themselves, as opposed to the standard approach of Gaussian parameters. It explains that distribution-based Bayesian classifier based on minimizing the expected misclassification costs is equivalent to the classifier that minimizes the expected Bregman divergence estimates of the class conditional distribution.

Chapter 5 reviews approaches to cross-validated Bayesian QDA and regularized quadratic discriminant analysis (RDA). It explains how the distribution-based Bayesian classifier can be realized as RDA [4]. Results are presented on simulated and benchmark datasets and comparisons are made with RDA, Quadratic Bayes (QB) [10], model-selection discriminant analysis based on eigenvalue decomposition (EDDA) [11], and to maximum likelihood

estimated quadratic and linear discriminant analysis (LDA).

In Chapter 6 the local distribution-based Bayesian quadratic discriminant analysis (local BDA) classifier is proposed which applies to the neighborhood formed by the k samples from each class that are closest to the query. Performance of the local BDA classifier is compared with local nearest means [12], recently proposed local support vector machine (SVM-KNN) [13], Gaussian mixture models, k-NN, and local linear regression.

Chapter 7 discusses functional Bregman divergence, and establishes its relation with previously defined Bregman definition [14, 15]. After establishing properties and the main theorem, it discusses the role of functional Bregman divergence in Bayesian estimation of distributions.

Chapter 8 concludes, discusses open questions, and suggests directions for future work.

1.4 Conventions and Notations

For convenience, we present in Table 1.1 the important notation used in the rest of the dissertation. Less frequently used notation will be defined later when it is first introduced. Random variables are represented by upper-case alphabets, e.g., X, Y and their realization are represented by smaller-case alphabets, e.g., x, y. The symbol $\arg\min$ stands for the argument of the minimum, that is to say, the value of the given argument for which the value of the given expression attains its minimum value.

Table 1.1: Key notation

Notation	Description	Notation	Description
$\mathbb{R}$	set of real numbers	$\mathbb{R}^d$	d-dimensional real vector space
$X_i \in \mathbb{R}^d$	i^{th} random training sample	n	number of training samples
$Y_i \in G$	class label corresponding to X_i	n_h	number of training samples of class h
$\mathcal{G} = \{1, 2, \ldots, G\}$	set of class labels	$\bar{X}_h$	sample mean for class h
$X \in \mathbb{R}^d$	random test sample	S_h	$\sum_{i=1}^{n}(X_i - \bar{X}_h)(X_i - \bar{X}_h)^T I_{(Y_i=h)}$
$Y \in G$	class label corresponding to X	S	$\sum_{i=1}^{n}(X_i - \bar{X})(X_i - \bar{X})^T$
I, $I_{(.)}$	identity matrix, indicator function	$\|B\|$	determinant of B
diag(B)	diagonal of B	tr(B)	trace of B
arg min	argument of the minimum	$ri(\mathcal{S})$	relative interior of $\mathcal{S}$
$\nabla\phi(y)$	gradient of ϕ at y	$P(\mathcal{X})$	probability of $\mathcal{X}$
$E_f[\theta]$	expectation of θ w.r.t. f	$\|\cdot\|$	l_2 - norm
μ	mean	Σ	covariance matrix
$\phi[f]$	functional over $L^p(\nu)$	$\delta\phi[f;\cdot]$	Fréchet derivative of ϕ at f
$\Gamma(\cdot)$	gamma function	$\Gamma_d(\cdot)$	multivariate gamma function

Chapter 2

PRINCIPLES OF ESTIMATION AND BREGMAN DIVERGENCE

In this chapter, some principles of estimation and Bayesian estimation are reviewed. Then, a result is presented for Bayesian estimation with Bregman divergence risk function. These principles are used differently depending upon the information given. Maximum likelihood (ML), maximum a posteriori (MAP), methods of moments (MOM), are reviewed in Section 2.1. Bayesian approach to parameter estimation and the concept of Bayesian risk function are discussed in Section 2.2. Section 2.3 discusses Bregman divergence, followed by examples and a theorem of Banerjee et al., 2005 that states that the mean minimizes the expected Bregman divergence. Section 2.4 discusses Bayesian estimation using the Bregman divergence risk function, and a new result shows that the mean of the posterior pdf is the optimal Bayesian estimator for any Bregman divergence risk. Examples of the computation of the Bayesian estimator using Bregman divergence risk are included in Section 2.5. The results in this chapter have been published in the workshop [16].

2.1 ML, MAP, MOM

The maximum likelihood estimator (ML) estimates a pmf that maximizes the probability (likelihood) of the given data $\mathcal{X}$. To estimate a parameter $\theta \in \mathbb{R}^d$ of a parametric distribution given observations $\mathcal{X}$, the ML estimate solves

$$\max_{\theta \in \mathbb{R}^d} P(\mathcal{X}|\theta). \tag{2.1}$$

1. It is intuitively appealing.

2. It has good asymptotic properties.

3. It coincides with the relative frequency of the event in the sample.

For example, if three out of ten parts arrive broken, the ML estimate for the probability of a broken part is .3. ML estimates are unbiased for multinomial distributions but can be biased for other distributions; for instance estimating standard deviation in the Gaussian case.

A related principle of estimation is the maximum a posterior estimate (MAP), which chooses the distribution with maximum probability given the observations $\mathcal{X}$, and a prior $P(\theta)$,

$$\begin{aligned} & \max_{\theta \in \mathbb{R}^d} P(\theta|\mathcal{X}) \\ = & \max_{\theta \in \mathbb{R}^d} \frac{P(\mathcal{X}|\theta)P(\theta)}{P(\mathcal{X})}. \end{aligned} \tag{2.2}$$

For an estimate of parametric distributions, the method of moments (MOM) defines the estimated moments to be the sample moments. Another approach to parametric distribution estimation is to find the unbiased minimum variance estimate; this goes by various names such as MVUE [17], UMVU [18].

2.2 Bayesian Minimum Expected Risk Estimation

For an unknown pmf parameterized by some θ, the Bayesian Mean Square Error Estimator (BMSEE) [17] (pages 310-316, 342-350) solves

$$\theta^* = \arg\min_{\hat{\theta} \in \mathbb{R}^d} \int_{\theta} (\theta - \hat{\theta})^2 f(\theta|\mathcal{X}) d\theta \tag{2.3}$$

where a prior distribution over the θ parameter, $f(\theta)$, has yielded a posterior pdf $f(\theta|\mathcal{X})$ based on knowledge or data $\mathcal{X}$. This is equivalent to solving

$$\theta^* = E_{f(\theta|\mathcal{X})}[\theta] \tag{2.4}$$

where the expectation is taken over the posterior $f(\theta|\mathcal{X})$. Thus the optimal estimator $\hat{\theta}$ in terms of minimizing the Bayesian Mean Square Error is the *mean* of the posterior pdf $f(\theta|\mathcal{X})$. The BMSEE estimator will in general depend on the prior knowledge as well the data $\mathcal{X}$. If the prior knowledge is weak relative to the knowledge of the data, then the estimator will ignore the prior knowledge. Otherwise, the estimator will be "biased"

towards the prior mean. On average, the use of relevant prior information always improves the estimation accuracy. More generally the Bayesian minimum expected risk principle [18, ch. 4] uses a risk function $R : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ to estimate the parameter as

$$\theta^* = \arg\min_{\hat{\theta}} \int R(\theta, \hat{\theta}) f(\theta|\mathcal{X}) d\theta \tag{2.5}$$

$$\equiv \arg\min_{\hat{\theta}} E_{f(\theta|\mathcal{X})}[R(\Theta, \hat{\theta})], \tag{2.6}$$

where $\Theta \in \mathbb{R}^d$ is a random variable with realization θ. The average risk or cost $E_{f(\theta|\mathcal{X})}[R(\Theta, \hat{\theta})]$ is termed as Bayes risk $\mathcal{R}$ or

$$\mathcal{R} = E_{f(\theta|\mathcal{X})}[R(\Theta, \hat{\theta})], \tag{2.7}$$

and measures the performance of the estimator. If $R(\theta, \hat{\theta}) = (\theta - \hat{\theta})^2$, then the risk function is quadratic and Bayes risk is just the mean square error (MSE). Other widely used risk functions are

$$R(\theta, \hat{\theta}) = |\theta - \hat{\theta}|, \tag{2.8}$$

$$R(\theta, \hat{\theta}) = \begin{cases} 0 & \text{if } |\theta - \hat{\theta}| < \delta \\ 1 & \text{if } |\theta - \hat{\theta}| > \delta. \end{cases} \tag{2.9}$$

The risk function (2.8) penalizes errors proportionally, while (2.9) penalizes with value 1 for error greater than the threshold $\delta > 0$. In all the above three cases the risk function is symmetric in $\theta - \hat{\theta}$, reflecting the implicit assumption that positive errors are just as bad negative errors. In general this need not be case. In the next section we will discuss the Bayesian estimation problem using a general risk function called Bregman divergence. Bregman divergences included a large number of useful risk or loss functions such squared loss, Kullback Leibler-divergence, logistic loss, Mahalanobis distance, Itakura-Saito distance, I-divergence, etc.

2.3 Bregman Divergence

This section defines the Bregman divergence [14], [15] corresponding to a strictly convex function and presents some examples.

2.3.1 Definition of Bregman Divergence

Let $\phi : \mathcal{S} \to \mathbb{R}$ be a strictly convex function defined on a convex set $\mathcal{S} \subset \mathbb{R}^d$ such that ϕ is differentiable on the relative interior of $\mathcal{S}$ $ri(\mathcal{S})$, assumed to be nonempty. The Bregman divergence $d_\phi : \mathcal{S} \times ri(\mathcal{S}) \to [0, \infty)$ is defined as

$$d_\phi(x, y) = \phi(x) - \phi(y) - (\nabla\phi(y))^T(x - y), \tag{2.10}$$

where $\nabla\phi(y)$ represent the gradient vector of ϕ evaluated at y.

Example 1: Squared Euclidean distance is the simplest and most widely used Bregman divergence. The underlying function $\phi(\mathbf{x}) = x^T x$ is strictly convex, differentiable on $\mathbb{R}^d$ and

$$\begin{aligned} d_\phi(x, y) &= x^T x - y^T y - (\nabla\phi(y))^T(x - y) \\ &= \|x\|^2 - \|y\|^2 - 2y^T(x - y) \\ &= \|x\|^2 - 2x^T y + \|y\|^2 \\ &= \|x - y\|^2 . \end{aligned}$$

Example 2: Another Bregman divergence is relative entropy or Kullback Leibler distance $D(p\|q)$ between two probability mass functions p and q. Relative entropy is a measure of the distance between two distributions. In statistics, it arises as the expected logarithm of the likelihood ratio. In information and coding theory, it is a measure of the inefficiency of assuming that the distribution is q when the true distribution is p. If p is a discrete probability distribution so that $\sum_{i=1}^d p_i = 1$, the negative entropy $\phi(p) = \sum_{i=1}^d p_i \log p_i$ is a

convex function. The corresponding Bregman divergence is

$$\begin{aligned} d_\phi(p,q) &= \sum_{i=1}^{d} p_i \log p_i - \sum_{i=1}^{d} q_i \log q_i - (\nabla\phi(q))^T(p-q) \\ &= \sum_{i=1}^{d} p_i \log p_i - \sum_{i=1}^{d} q_i \log q_i - \sum_{i=1}^{d} (\log q_i + 1)(p_i - q_i) \\ &= \sum_{i=1}^{d} p_i \log p_i - \sum_{i=1}^{d} q_i \log q_i - \sum_{i=1}^{d} (p_i - q_i) \log q_i \\ &= \sum_{i=1}^{d} p_i \log p_i - \sum_{i=1}^{d} p_i \log q_i \\ &= \sum_{i=1}^{d} p_i \log \frac{p_i}{q_i} \\ &= D(p\|q). \end{aligned}$$

Example 3: Itakura-Saito distance is another Bregman divergence that is widely used in signal processing. If $F(e^{i\theta})$ is the power spectrum of a signal $f(t)$, then the functional $\phi(F) = -\frac{1}{2\pi}\int_{-\pi}^{\pi} \log(F(e^{i\theta}))d\theta$ is convex in F and corresponds to the negative entropy rate of the signal assuming it was generated by a stationary Gaussian process [19], [20]. The Bregman divergence between $F(e^{i\theta})$ and $G(e^{i\theta})$ (the power spectrum of another signal $g(t)$) is given by

$$\begin{aligned} d_\phi(F,G) &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \left(-\log(F(e^{i\theta})) + \log(G(e^{i\theta})) - (F(e^{i\theta}) - G(e^{i\theta})) \left(-\frac{1}{G(e^{i\theta})} \right) \right) d\theta \\ &= \frac{1}{2\pi}\int_{-\pi}^{\pi} \left(-\log\left(\frac{F(e^{i\theta})}{G(e^{i\theta})}\right) + \frac{F(e^{i\theta})}{G(e^{i\theta})} - 1 \right) d\theta, \end{aligned}$$

which is exactly the Itakura-Saito distance between the power spectra $F(e^{i\theta})$ and $G(e^{i\theta})$ and can also be interpreted as the I-divergence [21] between the generating processes under the assumption that they are equal mean, stationary Gaussian process [22].

Table 2.1 contains a list of some common convex functions and their corresponding Bregman divergences.

The following theorem from Banerjee et al. 2005 [14] states that the mean of the random variable X minimizes the expected Bregman divergence and, surprisingly, *does not depend* on the choice of Bregman divergence.

Table 2.1: Bregman divergences generated from some convex functions.

Domain	$\phi(x)$	$d_\phi(x, y)$	**Divergences**
$\mathbb{R}$	x^2	$(x-y)^2$	Squared loss
$\mathbb{R}_+$	$x \log x$	$x \log\left(\frac{x}{y}\right) - (x-y)$	
$[0,1]$	$x \log x + (1-x)\log(1-x)$	$x \log\left(\frac{x}{y}\right) + (1-x)\log\left(\frac{1-x}{1-y}\right)$	Logistic loss
$\mathbb{R}_{++}$	$-\log x$	$\frac{x}{y} - \log\left(\frac{x}{y}\right) - 1$	Itakura-Saito distance
$\mathbb{R}$	e^x	$e^x - e^y - (x-y)e^y$	
$\mathbb{R}^d$	$\|x\|^2$	$\|x-y\|^2$	Square Euclidean distance
R^d	$x^T Ax$	$(x-y)^T A(x-y)$	Mahalanobis distance
d-Simplex	$\sum_{i=1}^d x_i \log x_i$	$x_i \log\left(\frac{x_i}{y_i}\right)$	KL-divergence
$\mathbb{R}_+^d$	$\sum_{i=1}^d x_i \log x_i$	$\sum_{i=1}^d x_i \left(\frac{x_i}{y_i}\right) - (x_i - y_i)$	Generalized I-divergence

Theorem 2.3.1. ***(Banerjee et al., 2005)*** *Let X be a random variable that takes values in $\mathcal{X} = \{x_i\}_{i=1}^n \subset \mathcal{S} \subseteq \mathbb{R}^d$ following a positive probability measure ν such that $E_\nu[X] \in ri(\mathcal{S})$. Given a Bregman divergence $d_\phi : \mathcal{S} \times ri(\mathcal{S}) \to [0, \infty)$, the problem*

$$\min_{s \in ri(\mathcal{S})} E_\nu[d_\phi(X, s)] \tag{2.11}$$

has a unique minimizer given by $s^* = \mu = E_\nu[X]$.

More examples of Bregman divergences and their properties can be found in [14, 23, 24].

2.4 Bayesian Estimation with Bregman Divergence

In this section, the class of Bregman divergences are considered for the risk functions for Bayesian estimation. The following theorem establishes the solution to (2.5) for Bregman divergence risk functions and a general form of likelihood.

Theorem 2.4.1. ***(Gupta, Srivastava, Cazzanti)*** *[25] Let the posterior $f(\theta)$ have the*

form

$$f(\theta) = \gamma \prod_{g=1}^{G} \theta_g^{\alpha_g}, \tag{2.12}$$

where γ *is a normalization constant, and* $\sum_{g=1}^{G} \alpha_g = 1$*, then for any Bregman divergence risk* $R(\theta, \phi) = d_\psi(\theta, \phi)$*,*

$$\theta_g^* = \frac{\alpha_g + 1}{\sum_{g=1}^{G} \alpha_g + G} \cdot \tag{2.13}$$

2.5 An Example: Binomial

As a simple example, suppose one orders five parts and when they arrive, three of the five parts are broken. We would like to estimate the probability of a part arriving broken based on this data. Let θ be the probability of a part arriving broken. Then the probability of the data $\mathcal{X}$ (three parts broken out of five) for a given θ is $P(\mathcal{X}|\theta) = 10(1-\theta)^2\theta^3$. In Figure 2.1, the likelihood $P(\mathcal{X}|\theta)$ for each θ is shown.

Based on the data $\mathcal{X}$, different estimates for pmf = *theta* are shown in Table 1. The probability of each of each pmf conditioned on the data is $P(\theta|\mathcal{X})$. Using Bayes' theorem, $P(\theta|\mathcal{X}) = P(\mathcal{X}|\theta)P(\theta)/P(\mathcal{X})$. Since the MER estimate solves a minimization problem, the $P(\mathcal{X})$ in the divisor is a constant and thus can be disregarded. If a prior or other information about $P(\theta)$ is available, then that information can be used. For this example we assume the prior $P(\theta)$ is uniform. Then, for the binomial case at hand, $f(\theta|\mathcal{X}) = 10(1-\theta)^2\theta^3$.

Method	Estimated θ
ML	.60
MOM	.57
Bayesian estimate with Bregman divergence	.57

Table 2.2: Estimated Bernoulli parameter θ, given that three parts out of five were broken.

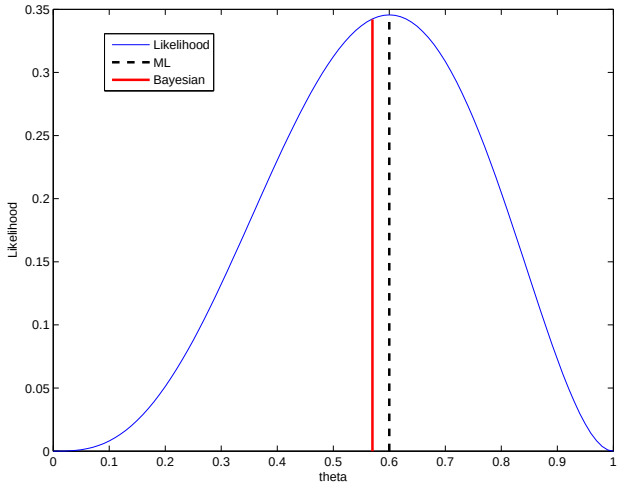

Figure 2.1: Probability of $\{\mathcal{X} = \text{three parts broken out of five}\}$, based on an iid Bernoulli distribution with parameter θ

As seen from Figure 2.1, the maximum probability $P(\mathcal{X}|\theta)$ occurs at $\theta = .60$, and thus this is the ML estimate. Which estimate is best? That depends on what one is trying to accomplish. Each estimate does exactly what its principle aims to do: the ML maximizes the probability of being right, but does not worry about how wrong the estimate could be. The Bayesian estimates minimize expected risk, and thus, on average, we expect these estimates to be more robust.

Chapter 3

BAYESIAN ESTIMATES FOR WEIGHTED NEAREST-NEIGHBOR CLASSIFIERS

In this chapter we derive minimum expected Bregman divergence estimates for weighted nearest-neighbor class probability estimates, and show that classifying with these class probability estimates minimizes the expected misclassification cost. Section 3.1 introduces supervised learning problem and notations. Section 3.2 discusses the k-nearest neighbor learning problem and a generalized form of Laplace smoothing for weighted k nearest-neighbors class probability estimates is derived. Section 3.3 discusses nearest neighbor classification with these probability estimates. The results in this chapter have been submitted for publication [25].

3.1 Introduction

The standard statistical learning problem is treated, with training pairs $\mathcal{T} = \{(X_i, Y_i)\}$ and test pair (X, Y) drawn independently and identically from a sufficiently nice joint distribution $P_{X,Y}$, where X_i and X are feature vectors in $\mathbb{R}^d$ and $Y_i \in \{1, 2, \ldots, G\}$, are the class labels. The problems are to estimate class label's probability $P_{Y|x} = P(Y|X = x)$ and Y given training pairs $\mathcal{T}$, test sample x, and a $G \times G$ misclassification cost matrix C, where $C(g, h)$ specifies the cost of classifying a test sample as class g when the truth is class h.

In this chapter, the unknown $P_{Y|x}$ is treated as a random vector Θ where Θ_g denotes the unknown $P(Y = g|x)$ for $g \in \{1, 2, \ldots, G\}$, a realization of Θ is the probability mass function θ, and Θ is distributed with density $f(\theta)$, which is constrained to have a particular formulation, as stated in (2.12).

3.2 Nearest-Neighbor Learning

Let the training samples be re-indexed by their distance to x, such that x_k is the kth nearest neighbor of x. Nonparametric nearest-neighbor methods assign a weight w_j to each x_j; the

present analysis is restricted to weights that satisfy $\sum_{j=1}^{k} w_j = 1$ and $w_j \geq 0$. The standard estimate for $P(Y = g|x)$ is [26]

$$\hat{\theta}_g = \sum_{j=1}^{k} w_j I_{(Y_j=g)}, \tag{3.1}$$

where $I_{(\cdot)}$ is an indicator function that equals one when its argument is true, and equals zero otherwise. Given the class pmf estimate $\hat{\theta} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_G)$, the standard classification of x is

$$\hat{Y} = \arg\min_{g} \sum_{h=1}^{G} C(g,h)\hat{\theta}_h. \tag{3.2}$$

The underlying model is that the k nearest neighbors of x are all drawn from the true $P_{Y|x}$, so that the likelihood of m_g nearest neighbors being labeled class g for every $g = 1, 2, \ldots, G$ is

$$f(\theta) = \left(\frac{k!}{\prod_{g=1}^{G} m_g!} \right) \prod_{g=1}^{G} \theta_g^{m_g}. \tag{3.3}$$

Under this model, the probability estimate (3.1) with $w_j = 1/k$ for all j maximizes the likelihood $f(\theta)$.

Similarly, define a *weighted likelihood* $f(\theta, w)$ to be the likelihood of the weighted neighbors:

$$f(\theta, w) = \gamma \prod_{g=1}^{G} \prod_{j=1}^{k} \theta_g^{w_j k I_{(Y_j=g)}} = \gamma \prod_{g=1}^{G} \theta_g^{k \sum_{j=1}^{k} w_j I_{(Y_j=g)}}, \tag{3.4}$$

where γ is the normalization constant. It can be shown that the estimate (3.1) maximizes the weighted likelihood (3.4).

ML estimates can yield high variance estimates because the maximum of the likelihood function can be quite unrepresentative of the complete likelihood distribution, particularly when sample sizes are small. For nearest neighbor classifiers the sample sizes are often very small in an effort to keep the neighborhood local, which under the compactness hypothesis improves the validity of the nearest-neighbor model assumption that the k nearest neighbors are drawn from $P_{Y|x}$. See [27, pgs. 300-310] and [28, ch. 15] for further discussions of the problems with maximum likelihood estimation. Other smoothing approaches have also been applied to statistical learning, most heuristic in nature [29].

From theorem 2.4.1, for any Bregman divergence risk the probability estimate θ_g^* is given as

$$\theta_g^* = \frac{\alpha_g + 1}{\sum_{g=1}^{G} \alpha_g + G}, \tag{3.5}$$

where

$$\alpha_g = k \sum_{j=1}^{k} w_j I_{(Y_j = g)}. \tag{3.6}$$

When the weights are uniform such that $w_j = 1/k$ for all j, the estimate θ^* is equivalent to *Laplace correction* for estimating multinomial distributions [30, pg. 272], also called *Laplace smoothing*. Appropriately, the history of Laplace correction goes back to Laplace himself; Jaynes offers historical information and more details about alternate derivations [31, pgs 154-165]. Laplace correction has been shown to be useful for class probability estimation in decision trees [32, 33, 34, 35], and with naive Bayes [36]. Laplace correction was incorporated in the CN2 rule learner [37], and Domingos used it to break ties in a unified instance-based and rule-based learner [38].

3.3 Classifying

For zero-one costs such that $C(g, g) = 0$ and $C(g, h) = 1$ for all $g \neq h$, it can easily be shown that using either the maximum likelihood estimate given in (3.1) or the minimum expected risk estimate given in (2.13) with the classification rule (3.2) will result in the same estimated class.

For more general costs, the Bayes classifier minimizes

$$\arg \min_{g \in \{1,2,\ldots,G\}} \sum_{h=1}^{G} C(g, h) P(Y = h|x).$$

In practice, $P(Y = h|x)$ is unknown and is estimated as some $\hat{\theta}$ as discussed in this paper. It is proposed that instead of first estimating the class pmf $P(Y = h|x)$ and then classifying, one should model the uncertain class pmf as a random variable Θ, and directly classify to minimize the expected misclassification cost, choosing the class $\hat{Y}$ that solves:

$$\hat{Y} = \arg \min_{g \in \{1,2,\ldots,G\}} E_\Theta \left[\sum_{h=1}^{G} C(g, h) \Theta_h \right]. \tag{3.7}$$

Corollary 3.3.1. Classifying as per (3.7) is equivalent to classifying as per (3.2) with the class pmf estimate given by (A.2).

Proof. The equivalence follows directly from applying the linearity of expectation to (3.7) and (A.1). □

This corollary explicitly links minimizing expected Bregman divergences of the class probability estimates to the optimal choice for the class label in terms of expected misclassification cost. For discrete random variables, it has been shown that for the expectation of a risk function to equal the expectation of the variable it is necessary that the risk function be a Bregman divergence [14, Theorem 4]. It is conjectured in this chapter that this is true for continuous random variables as well, such that the solution to (6) is $E_{\Theta}[\Theta]$ only if R is a Bregman divergence.

In summary, this chapter theoretically motivated the application of a generalized form of Laplace smoothing for weighted k nearest-neighbors class probability estimates as the solution to minimizing any expected Bregman divergence. Also, it established that minimizing expected Bregman divergence in the class pmf estimation is equivalent to resolving the uncertainty of the unknown class pmf so as to minimize the expected misclassification cost. This simple result is important because it establishes that minimizing the Bregman divergence of the class estimate has a direct link to minimizing the 0-1 misclassification cost, which is difficult to work with analytically.

Chapter 4

DISTRIBUTION-BASED BAYESIAN MINIMUM EXPECTED RISK FOR DISCRIMINANT ANALYSIS

This chapter considers a distribution-based Bayesian estimation for classification by quadratic discriminant analysis, instead of the standard parameter-based Bayesian estimation. This approach yields closed form solutions, but removes the parameter-based restriction of requiring more training samples than feature dimensions. Section 4.1 describes the motivations behind the distribution-based approach to Bayesian quadratic discriminant analysis. Section 4.2 reviews bias phenomenon in linear and quadratic discriminant analysis. Section 4.3 discusses related work on Bayesian approach to quadratic discriminant analysis. In Section 4.4 the criterion of minimizing expected misclassification cost is motivated and distribution-based approach to Bayesian quadratic discriminant analysis is proposed using idea of statistical manifold of Gaussian distributions. Section 4.4.2 investigates prior so that it has an adaptively regularizing effect. In Section 4.4.3 closed form solutions of distribution-based and parameter-based Bayesian discriminant analysis classifiers are established for different priors. In Section 4.5 performance of the various classifiers are compared on a suite of simulations. This chapter takes the first steps towards showing that the prior itself can act as an efficient regularizing force. The results in this chapter have been published [39, 40].

4.1 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 based on minimizing the expected misclassification cost [5]. This method is also known as predictive classification. The Gaussian parameters for each class can be estimated from training points with maximum likelihood (ML) estimation. The simple Gaussian model is best suited for cases

when one does not have much information to characterize a class. 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 [41], but this approach has not become popular, even though it minimizes the expected misclassification cost. Ripley [42], 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 [43] examines Bayesian QDA in detail, but does not show that in practice it can yield better performance than regularized QDA. The performance of Bayesian QDA classifiers is very sensitive to the choice of prior [39], and that priors suggested by Geisser [41] and Keehn [44] produce error rates similar to those yielded by ML.

This chapter considers two issues in using Bayesian estimation effectively for quadratic discriminant analysis. First, it considers directly integrating out the uncertainty over the domain of Gaussian probability distributions (pdfs), as opposed to the standard approach of integrating out the uncertainty over the domain of the parameters. The proposed distribution-based Bayesian discriminant analysis removes the parameter-based Bayesian analysis restriction of requiring more training samples than feature dimensions and also removes the question of invariance to transformations of the parameters, because the estimate is defined in terms of the Gaussian distribution itself. Comparative performance on the Friedman suite of simulations shows that the distribution-based Bayesian discriminant analysis is also advantageous in terms of average error. The second issue considered here is the choice of prior. Ideally, a prior should have an adaptively regularizing effect, yielding robust estimation when the number of training samples is small compared to the number of feature dimensions (and hence the number of parameters), but also converging as the number of data points grows large. In practice, there can be more informative features d than labeled training samples n. This situation has previously been addressed through regularization (such as regularizing quadratic discriminant analysis with linear discriminant

analysis [4]). This chapter takes the first steps towards showing that the prior itself can act as an efficient regularizing force. This would be more clear when we would move to Chapter 5.

4.2 Review of Bias in Linear and Quadratic Discriminant Analysis

This review section is based on [5], [4], and Michael D. Perlman class's note on multivariate statistic. The most common generative classification rules are based on the normal distribution

$$f_h(x) = \frac{1}{(2\pi)^{\frac{d}{2}}|\Sigma_h|^{\frac{1}{2}}} \exp\left(-\frac{1}{2}(x-\mu_h)^T\Sigma_h^{-1}(x-\mu_h)\right), \tag{4.1}$$

where μ_h and Σ_h are the class h $(1 \leq h \leq G)$ mean and covariance matrix. For the simple loss $(0-1)$ function and the uniform prior over the class labels, the classification rule for a given test sample $x \in \mathbb{R}^d$ becomes: classify as class $\hat{g}$ where

$$\hat{g} = \arg\min_{h\in\{1,2,\ldots,G\}} d_h(x) \tag{4.2}$$

$$\text{where } d_h(x) = (x-\mu_h)^T\Sigma_h^{-1}(x-\mu_h) + \log|\Sigma_h|. \tag{4.3}$$

The quantity $d_h(x)$ is called the discriminant function. The first term on the right side of (4.3) is the well-known Mahalanobis distance between x and μ_h.

The classification rule (4.2) and (4.3) is called quadratic discriminant analysis (QDA) since it separates the disjoint regions of the feature space corresponding to each class label by quadratic boundaries. When all the class covariance matrices are identical

$$\Sigma_h = \Sigma, \;\; 1 \leq h \leq G, \tag{4.4}$$

then classification rule (4.2) and (4.3) is called linear discriminant analysis (LDA). LDA results in linear decision boundaries as the quadratic terms associated with (4.2) and (4.3) get canceled.

Quadratic and linear discriminant analysis can be expected to work well if the class conditional densities are approximately normal and good estimates can be obtained for mean μ_h and covariance matrices Σ_h. Classification rules based on QDA are known to require generally larger samples than those based on LDA and seems to be more sensitive to violation of the basic assumptions.

In common application of linear and quadratic discriminant analysis the parameters associated with the class densities are estimated by their sample analogs

$$\hat{\mu}_h = \bar{X}_h \tag{4.5}$$

$$\hat{\Sigma}_h = \frac{1}{n_h}\sum_{i=1}^{n}(X_i - \bar{X}_h)(X_i - \bar{X}_h)^T I_{(y_i=h)}. \tag{4.6}$$

When the class sample size n_h $(1 \leq h \leq G)$ is small compared with the dimension of the feature space d, the covariance matrix estimates, especially, become highly variable. Moreover, when $n_h < d$ not all of their parameters are even identifiable. The effect this has on discriminant analysis can be seen by representing the class covariance matrices by their spectral decompositions

$$\Sigma_h = \sum_{i=1}^{d} \lambda_{ih} v_{ih} v_{ih}^T,$$

where λ_{ih} is the ith eigenvalue of Σ_h (ordered in decreasing value) and $v_{ih} \in \mathbb{R}^d$ is the corresponding eigenvector. The inverse in this representation is

$$\Sigma_h^{-1} = \sum_{i=1}^{d} \frac{v_{ih} v_{ih}^T}{\lambda_{ih}},$$

and the discriminant function (4.3) becomes

$$d_h(x) = \sum_{i=1}^{d} \frac{[v_{ih}^T(x - \mu_h)]^2}{\lambda_{ih}} + \sum_{i=1}^{d} \ln \lambda_{ih}. \tag{4.7}$$

The discriminant function (4.7) is heavily weighted by the smallest eigenvalues and the direction associated with their eigenvectors. When sample-based plug-in estimates are used, this becomes the eigenvalues and eigenvectors of $\hat{\Sigma}_h$. Moreover, writing the extremal representation of the largest and smallest eigenvalues of the hth class's estimated covariance matrix $\hat{\Sigma}_h$,

$$\lambda_{1h}(\hat{\Sigma}_h) = \max_{v^T v=1} v^T \hat{\Sigma}_h v$$

$$\lambda_{dh}(\hat{\Sigma}_h) = \min_{v^T v=1} v^T \hat{\Sigma}_h v.$$

Therefore, $\lambda_{1h}(\hat{\Sigma}_h)$ and $\lambda_{dh}(\hat{\Sigma}_h)$ are, respectively, convex and concave functions of $\hat{\Sigma}_h$. Thus

by Jensen's inequality,

$$E_{\Sigma_h}[\lambda_{1h}(\hat{\Sigma}_h)] \geq \lambda_{1h}(E_{\Sigma_h}[\hat{\Sigma}_h]) = \lambda_{1h}(\Sigma_h) \tag{4.8}$$

$$E_{\Sigma_h}[\lambda_{dh}(\hat{\Sigma}_h)] \leq \lambda_{dh}(E_{\Sigma_h}[\hat{\Sigma}_h]) = \lambda_{dh}(\Sigma_h). \tag{4.9}$$

Thus, even though the estimate $\hat{\Sigma}_h$ given by (4.6) is unbiased estimate of Σ_h, it produces the biased estimate of the eigenvalues: the largest eigenvalue is biased high (overestimated) (4.8), while the smallest eigenvalue is biased low (underestimated) (4.9). This holds for the other eigenvalues also. One way to attempt to mitigate this problem is to either try to obtain more reliable estimates of the eigenvalues: by shrinking the larger eigenvalues and expanding the lower ones. Moreover

$$E\left[\prod_{i=1}^{d}\lambda_{ih}(\hat{\Sigma}_h)\right] = \prod_{i=1}^{d}\lambda_{ih}(\Sigma_h)\prod_{j=1}^{d}\left(\frac{n_h-d+j}{n_h}\right) \tag{4.10}$$

$$= \prod_{i=1}^{d}\lambda_{ih}(\Sigma_h)\left(\prod_{j=1}^{d}\left(\frac{n_h-d+j}{n_h}\right)\right)^{\frac{d}{d}}$$

$$\leq \prod_{i=1}^{d}\lambda_{ih}(\Sigma_h)\left(\sum_{j=1}^{d}\left(\frac{1}{d}\frac{n_h-d+j}{n_h}\right)\right)^{d} \tag{4.11}$$

$$= \prod_{i=1}^{d}\lambda_{ih}(\Sigma_h)\left(1-\frac{d-1}{2n_h}\right)^{d}$$

$$\leq \prod_{i=1}^{d}\lambda_{ih}(\Sigma_h)\exp[-\frac{d(d-1)}{2n_h}], \tag{4.12}$$

where (4.11) follows from the fact that geometric mean is less than equal to arithmetic means, (4.12) follows from the inequality $1-x \leq \exp(-x)$. Thus, $\prod_{i=1}^{d}\lambda_{ih}(\hat{\Sigma}_h)$ will tend to underestimate $\prod_{i=1}^{d}\lambda_{ih}(\Sigma_h)$ unless $n_h \gg d^2$, which does not usually hold in applications. This suggest that shrinkage-expansion of the sample eigenvalues should not be done in a linear manner: the smaller $\lambda_{ih}(\hat{\Sigma}_h)'s$ should be expanded proportionally more than the larger $\lambda_{ih}(\hat{\Sigma}_h)'s$ should be shrunk.

With as many parameters in the model as training examples we would expect ML estimation to lead to severe over-fitting. To avoid this a common approach is to impose some additional constraint on the parameters, for example the addition of a penalty term

to the likelihood or error function. The other way is to adopt a Bayesian perspective and 'constrain' the parameters by defining an explicit prior probability distribution over them.

4.3 Related Work on Bayesian Approach to QDA

Discriminant analysis using Bayesian estimation was first proposed by Geisser [41] and Keehn [44]. 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 has a Wishart distribution. Work by Brown et al. [10] on this topic uses conjugate priors, and they proposed a hierarchical approach that compromises between the two extremes of linear and quadratic Bayes discriminant analysis, similar to Friedman's regularized discriminant analysis [4]. Raudys and Jain note that the Geisser and Keehn Bayesian discriminant analysis may be inefficient when the class sample sizes differ [45]. In all of this prior Bayesian work is *parameter-based* in that the mean μ and covariance Σ are treated as random variables, and the expectation of μ and of Σ are calculated with respect to Lebesgue measure over the domain of the parameters. In the next section *distribution-based* Bayesian QDA is solved, 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. The mathematics of such statistical manifolds needed for this approach has been investigated by Kass [46], Amari [47] and others.

4.4 Distribution-Based Bayesian Quadratic 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 Bayesian QDA is derived by formulating the problem in terms of the Gaussian distributions explicitly. This section extends work presented in a recent conference paper [39].

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

$$Y^* \stackrel{\triangle}{=} \arg \min_{g=1,\ldots,G} \sum_{h=1}^{G} C(g,h)p(x|Y=h)P(Y=h). \tag{4.13}$$

In practice the class conditional distributions and the class priors are usually unknown. Model each unknown distribution $p(x|h)$ by a random Gaussian distribution N_h, and model the unknown class priors by the random vector Θ, which has components $\Theta_h = P(Y = h)$ for $h = 1, \ldots, G$. Then, 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 (4.13) with their random counterparts and taking the expectation:

$$\hat{Y} \stackrel{\triangle}{=} \arg \min_{g=1,\ldots,G} E\left[\sum_{h=1}^{G} C(g,h)N_h(x)\Theta_h\right]. \tag{4.14}$$

In (4.14) 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 (4.14) can be rewritten as

$$\hat{Y} = \arg \min_{g=1,\ldots,G} \sum_{h=1}^{G} C(g,h)E_{N_h}[N_h(x)]E_{\Theta}[\Theta_h]. \tag{4.15}$$

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

4.4.1 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, [47, pp. 25–28].

Let the differential element over the set M be defined by the Riemannian metric [46, 47],

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

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

$$dM = \frac{d\mu}{|\Sigma|^{\frac{1}{2}}} \frac{d\Sigma}{|\Sigma|^{\frac{d+1}{2}}} = \frac{d\mu d\Sigma}{|\Sigma|^{\frac{d+2}{2}}}. \tag{4.16}$$

Let $\mathcal{N}_h(\mu_h, \Sigma_h)$ be a possible realization of the Gaussian pdf N_h. Using the measure defined in (4.16),

$$E_{N_h}[N_h(x)] = \int_M \mathcal{N}_h(x) r(\mathcal{N}_h) dM, \tag{4.17}$$

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,

$$r(\mathcal{N}_h) = \frac{\ell(\mathcal{N}_h, \mathcal{T}_h) p(\mathcal{N}_h)}{\alpha_h}, \tag{4.18}$$

where α_h is a normalization constant, $p(\mathcal{N}_h)$ is the prior probability of $\mathcal{N}_h$ (treated further in Section 4.4.2), and $\ell(\mathcal{N}_h, \mathcal{T}_h)$ is the likelihood of the data $\mathcal{T}_h$ given $\mathcal{N}_h$, that is,

$$\ell(\mathcal{N}_h(\mu_h,\Sigma_h), \mathcal{T}_h) = \frac{\exp[-\frac{1}{2}\operatorname{tr}\left(\Sigma_h^{-1} S_h\right) - \frac{n_h}{2}\operatorname{tr}\left(\Sigma_h^{-1}(\mu_h - \bar{X}_h)(\mu_h - \bar{X}_h)^T\right)]}{(2\pi)^{\frac{dn_h}{2}} |\Sigma_h|^{\frac{n_h}{2}}}. \tag{4.19}$$

4.4.2 Priors

A prior probability distribution of the Gaussians, $p(\mathcal{N}_h)$, is needed to solve the classification problem given in (4.14). A common interpretation of Bayesian analysis is that the prior represents information that one has prior to seeing the data [31]. 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, consider 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

$$p(\mathcal{N}_h) = p(\mu_h)p(\Sigma_h) = \gamma_0 \frac{\exp[-\frac{1}{2}\operatorname{tr}\left(\Sigma_h^{-1}B_h\right)]}{|\Sigma_h|^{\frac{q}{2}}}, \tag{4.20}$$

where B_h is a positive definite matrix and γ_0 is a normalization constant. The prior (4.20) is equivalent to a noninformative prior for the mean μ, and an inverted Wishart prior with q degrees of freedom over Σ. One can also note that if $B_h = 0$ and $q = d + 1$, the prior (4.20) reduces to an improper, invariance non-informative prior.

$$p(\mathcal{N}_h) = \frac{1}{|\Sigma_h|^{\frac{d+1}{2}}}. \tag{4.21}$$

To meet the goal of minimizing bias, encode some coarse information about the data into B_h. Setting $B_h = kI$ is reminiscent of Friedman's RDA [4], where the covariance estimate is regularized by the trace: $\frac{\operatorname{tr}(\hat{\Sigma}_{ML})}{d}I$. The trace of the ML covariance estimate is stable, and provides coarse information about the scale of the data samples. As pointed out by Friedman [4], this term $\frac{\operatorname{tr}(\hat{\Sigma}_{ML})}{d}I$ has the effect of decreasing the larger eigenvalues and increasing the smaller ones, thereby counteracting the biasing inherent in sample-based estimation of eigenvalues.

This chapter shows that by setting $B_h = \frac{\operatorname{tr}(\hat{\Sigma}_{ML})}{d}I$ and $q = d + 3$ a distribution-based discriminant analysis outperforms Geisser's or Keehn's parameter-based Bayesian discriminant methods, and does not require crossvalidation [39]. Next, the closed form result with a prior of the form given in (4.20) is described. Chapter 5 returns to the question of data-dependent definitions for B_h when we propose the BDA7 classifier and there it is shown that an effect this prior has on discriminant analysis is that it regularizes the likelihood covariance estimate towards the maximum of the prior.

4.4.3 Closed-Form Solutions

In Theorem 4.4.1 the closed-form solution for the proposed distribution-based Bayesian discriminant analysis classifier is established. The closed-form solution for the parameter-based classifier with the same prior is given in Corollary 4.4.2.

Theorem 4.4.1. ***(Srivastava, Gupta 2006)*** *The classifier (4.15) using the inverted Wishart prior (4.20) is equivalent to*

$$\hat{Y} = \arg\min_{g} \sum_{h=1}^{G} C(g,h) \frac{(n_h)^{\frac{d}{2}} \Gamma(\frac{n_h+q+1}{2}) \left|\frac{S_h+B_h}{2}\right|^{\frac{n_h+q}{2}}}{(n_h+1)^{\frac{d}{2}} \Gamma(\frac{n_h+q-d+1}{2}) |A_h|^{\frac{n_h+q+1}{2}}} \hat{P}(Y=h), \tag{4.22}$$

where

$$A_h = \frac{1}{2}\left(S_h + \frac{n_h (X-\bar{X}_h)(X-\bar{X}_h)^T}{(n_h+1)} + B_h\right). \tag{4.23}$$

The proof is given in the Appendix A. Because $q \geq d$, the solution (4.22) is valid for any $n_h > 0$ and any feature space dimension d.

Corollary 4.4.2. *The parameter-based Bayesian discriminant analysis solution using the inverted Wishart prior given in (4.20) is to classify test point X as class label*

$$\hat{Y} \stackrel{\triangle}{=} \arg\min_{g} \sum_{h=1}^{\mathcal{G}} C(g,h) \frac{n_h^{\frac{d}{2}} \Gamma(\frac{n_h+q-d-1}{2})}{(n_h+1)^{\frac{d}{2}} \Gamma(\frac{n_h+q-2d-1}{2})} \frac{|\frac{S_h+B_h}{2}|^{\frac{n_h+q-d-2}{2}}}{|A_h|^{\frac{n_h+q-d-1}{2}}} \hat{P}(Y=h). \tag{4.24}$$

The proof of this corollary follows the same steps as the proof of the presented theorem 4.4.1 by replacing the Fisher information measure $\frac{d\mu d\Sigma}{|\Sigma|^{\frac{d+2}{2}}}$ with Lebesgue measure. One can also get parameter-based Bayesian classifier (4.24) by replacing q equal to $q-d-2$ in distribution-based Bayesian classifier (4.22). Notably, the parameter-based Bayesian discriminant solution (4.24) will not hold if $n_h \leq 2d-q+1$.

Theorem 4.4.3. *The distribution-based Bayesian discriminant analysis solution using the noninformative prior*

$$p(\mathcal{N}_h) = p(\mu_h)p(\Sigma_h) = \frac{1}{|\Sigma_h|^{\frac{d+1}{2}}}, \tag{4.25}$$

is to classify test point X as class label

$$\hat{Y} \stackrel{\triangle}{=} \arg\min_{g} \sum_{h=1}^{\mathcal{G}} C(g,h) \frac{n_h^{\frac{d}{2}} \Gamma(\frac{n_h+d+2}{2})}{(n_h+1)^{\frac{d}{2}} \Gamma(\frac{n_h+2}{2})} \frac{|\frac{S_h}{2}|^{\frac{n_h+d+1}{2}}}{|T_h|^{\frac{n_h+d+2}{2}}} \hat{P}(Y=h), \tag{4.26}$$

where

$$T_h = \frac{1}{2}\left(S_h + \frac{n_h(X-\bar{X}_h)(X-\bar{X}_h)^T}{(n_h+1)}\right). \tag{4.27}$$

The proof is given in the Appendix A. Again, this distribution-based Bayesian discriminant solution (4.26) will hold for any $n_h > 0$ and any d. Also note that one gets (4.26) by setting $B_h = 0$ and $q = d+1$ in (4.22)

A parameter-based Bayesian discriminant analysis given by Geisser [41] using the non-informative prior over Σ and μ is also given for comparison.

Theorem 4.4.4. ***(Geisser 1964)*** *The parameter-based Bayesian discriminant analysis solution using the noninformative prior (4.25) is to classify test point X as class label*

$$\hat{Y} \stackrel{\triangle}{=} \arg\min_{g} \sum_{h=1}^{\mathcal{G}} C(g,h) \frac{n_h^{\frac{d}{2}} \Gamma(\frac{n_h}{2})}{(n_h+1)^{\frac{d}{2}} \Gamma(\frac{n_h-d}{2})} \frac{|\frac{S_h}{2}|^{\frac{n_h-1}{2}}}{|T_h|^{\frac{n_h}{2}}} \hat{P}(Y=h), \tag{4.28}$$

where T_h is given by (4.27).

Note that Geisser's parameter-based Bayesian classifier requires at least d number of training samples from each class, for (4.28) to holds. Also Geisser's formula (4.28) can be directly obtained from (4.26) by substituting n_h as $n_h - d - 2$.

4.5 Simulation

The performance of the various estimators was compared using simulations similar to those proposed by Friedman to evaluate regularized discriminant analysis [4]. The comparison is between parameter-based Bayesian estimation, distribution-based Bayesian estimation, quadratic discriminant analysis, and nearest-means classification. Furthermore, for the Bayesian perspective, the non-informative prior was compared to the inverted Wishart prior with $d+3$ degree of freedom for the covariance (the non-informative prior was used throughout for the mean). The class label is randomly drawn to be class 1 ($Y = 1$) with probability half, and class 2 ($Y = 2$) with probability half.

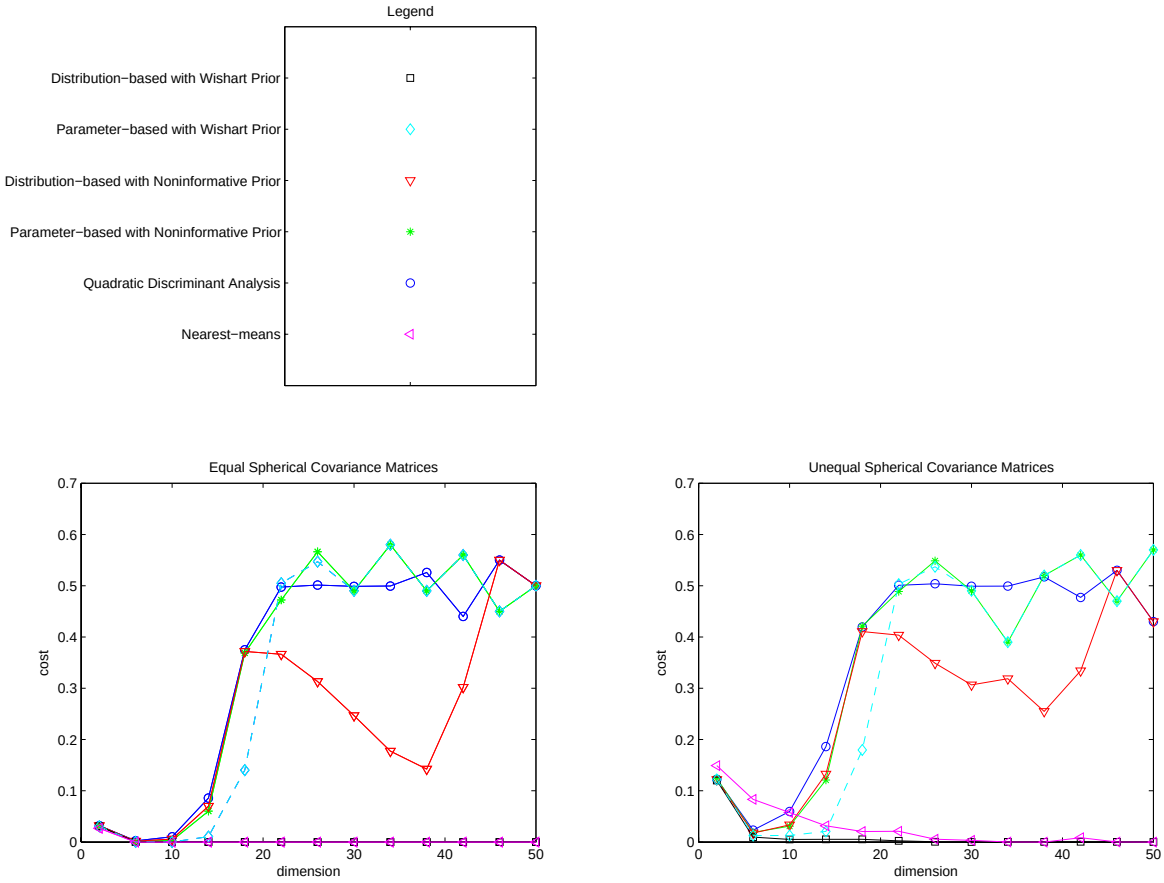

Figure 4.1: Below: Results of equal (bottom left) and unequal (bottom right) spherical covariance matrix simulation

Equal Spherical Covariance Matrices

Each class conditional distribution was normal with identity covariance matrix I. The mean of the first class μ_1 was the origin. Each component of the mean μ_2 of the second class was 3. Results are shown in Figure 4.5 (bottom left).

Unequal Spherical Covariance Matrices

Conditioned on class 1, the distribution was normal with identity covariance matrix I and mean at the origin. Conditioned on class 2, the distribution was normal with covariance matrix $2I$ and each component of the mean was 3. Results are shown in Figure 4.5 (bottom

right).

Equal Highly Ellipsoidal Covariance Matrices

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

$$e_i = [\frac{9(i-1)}{d-1} + 1]^2, \quad 1 \leq i \leq d, \tag{4.29}$$

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

A first case was that the class mean differences were concentrated in a low-variance subspace. The mean of class 1 was located at the origin and i^{th} component of the mean of class 2 was given by

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

Results are shown in Figure 4.5 (top left).

A second case was that the class mean differences were concentrated in a high-variance subspace. The mean of the class 1 was again located at the origin and the i^{th} component of the mean of class 2 was given by

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

Result is shown in Figure 4.5 (top right).

Unequal Highly Ellipsoidal Covariance Matrices

Covariance matrices were highly ellipsoidal and different for each class. The eigenvalues of the class 1 covariance were given by equation (4.29) and those of class 2 were given by

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

A first case was that the class means were identical. A second case was that the class means were different, where the mean of class 1 was located at the origin and the i^{th} component of the mean of class 2 was given by $\mu_{2i} = \frac{14}{\sqrt{d}}$. Results are shown in Figure 4.5 (bottom left) and Figure 4.5 (bottom right) respectively.

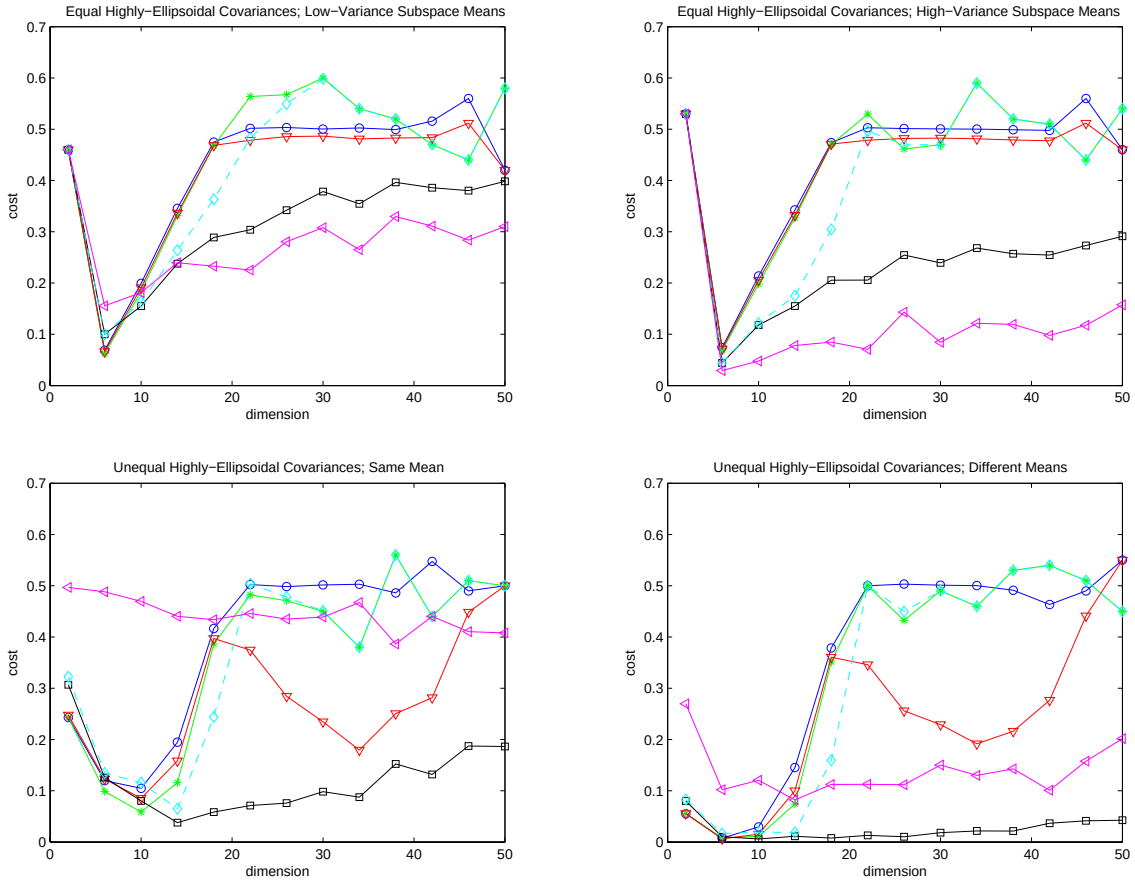

Figure 4.2: Above: Results of equal highly ellipsoidal covariance matrix with low variance (top left) and high variance (top right) subspace mean differences simulations. Below: Results of unequal highly ellipsoidal covariance matrix with the same (bottom left) and the different (bottom right) means simulations.

Experimental Procedure

For each of the above described choices of class conditional covariance matrix and mean, the figures show the average misclassification costs from 1000 replications of the following procedure: First $n = 40$ training sample pairs were drawn iid. Each classifier used the training samples to estimate its parameters. For all the classifiers, the prior probability of each of the two classes was estimated based on the number of observations from each class using Bayesian minimum expected risk estimation. Then, 100 test samples were drawn iid,

and classified by each estimator.

4.6 Conclusion

The distribution-based Bayesian discriminant analysis is seen to perform better in almost all cases of the simulations. In particular, using the adaptive inverted Wishart prior led to significantly better performance in some cases. We hypothesize that this choice of prior has a regularizing effect, and that using a well-designed adaptive prior could be an effective regularization strategy for discriminant analysis without the need for cross-validation to find regularization parameters as in regularized discriminant analysis [4].

Acknowledgment

The work in this chapter was funded in part by the Office of Naval Research, Code 321, Grant # N00014-05-1-0843. We thank Béla Frigyik and Richard Olshen for helpful discussions on this work.

Chapter 5

BAYESIAN QUADRATIC DISCRIMINANT ANALYSIS

This chapter proposes 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 twelve benchmark datasets and ten simulations, performance of BDA7 is compared to that of Friedman's regularized quadratic discriminant analysis (RDA) [4], to a model-selection discriminant analysis (EDDA) [11], to a modern cross-validated Bayesian QDA (QB) [10], and to ML-estimated QDA, LDA, and the nearest-means classifier. 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 better than the other approaches on average for the real datasets. The simulations help analyze the methods under controlled conditions.

This chapter also contributes to the theory of Bayesian QDA in several aspects. First, it is shown that the Bayesian distribution-based classifier that minimizes the expected misclassification cost is equivalent to the classifier that minimizes the expected Bregman divergence of the class conditional distributions. Second, using a series approximation, it is shown how the Bayesian QDA solution acts like Friedman's regularized QDA, which provides insight into determining effective prior distributions.

Chapter 4 has already discussed that the distribution-based Bayesian classifier performance is superior to the parameter-based Bayesian classifier given the same prior if no cross-validation is allowed. Section 5.1 reviews approaches to cross-validated Bayesian QDA and regularized QDA. An approximate relationship between Bayesian QDA and Friedman's regularized QDA is given in Section 5.2. Section 5.3 establishes 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, it turns to the practical matter of classification: proposes a cross-validated Bayesian QDA classifier BDA7 in Section 5.4. In Section 5.5, benchmark dataset results compare BDA7 to other QDA classifiers, followed by further analysis using simulation results in Section 5.6. The chapter concludes with a discussion of the results. The results in this chapter have been submitted for journal publication [40]

5.1 Prior Research on Ill-Posed QDA

5.1.1 Bayesian Approaches to QDA

In Section 4.3 Bayesian approaches to QDA were reviewed.

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 [10], 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.

5.1.2 Regularizing QDA

Friedman [4] 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:

$$\hat{\Sigma}_h(\lambda) = \frac{(1-\lambda)S_h + \lambda S}{(1-\lambda)n_h + \lambda n} \tag{5.1}$$

$$\hat{\Sigma}_h(\lambda, \gamma) = (1-\gamma)\hat{\Sigma}_h(\lambda) + \frac{\gamma}{d}\mathrm{tr}\left(\hat{\Sigma}_\mathrm{h}(\lambda)\right)\mathrm{I}. \tag{5.2}$$

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 quadratic discriminant analysis 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 [5].

Hoffbeck and Landgreb [48] proposed a similar regularized covariance estimate for classification of the form

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

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 [48]. Another restricted model used to regularize covariance matrix estimation is a banded covariance matrix [49].

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. Besnmail and Celeux [11] propose eigenvalue decomposition discriminant analysis (EDDA), in which fourteen different models are considered. Each model is the reparameterization of the class covariance matrix Σ_h in terms of eigenvalues decomposition $\Sigma_h = \lambda_h A_h D_h A_h^T$, where λ_h specifies the volume of density contours of the class, diagonal matrix of eigenvalues D_h specifies its shape, and the eigenvectors A_h specifies its orientation. Variations on constraints concerning volumes, shape, and orientations, λ_h, D_h and A_h leads to fourteen discrimination models 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, and 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 discriminant analysis considers a continuous set of possible models for Σ, and averages the continuous set with respect to each model's posterior probability.

5.1.3 Other Approaches to Quadratic Discriminant Analysis

Other approaches have been developed for ill-posed quadratic discriminant analysis. Friedman [4] 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 [50]. One of the most recent algorithms of this type is orthogonal linear discriminant analysis Ye [51], which was shown by the author of that work to perform similarly to Friedman's regularized linear discriminant analysis on six real data sets.

5.1.4 Priors and Regularization

Prior (4.20) is used where B_h is a positive definite matrix (further specified in (5.13)) and γ_0 is a normalization constant. The prior (4.20) 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 [52] 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 $\Sigma_{h,max}$,

$$\begin{aligned} -\frac{1}{2}\frac{\partial}{\partial \Sigma_h}\operatorname{tr}\left(\Sigma_{h,max}^{-1}B_h\right) - \frac{q}{2}\frac{\partial}{\partial \Sigma_h}\log|\Sigma_{h,max}| &= 0 \\ \Sigma_{h,max}^{-1}B_h\Sigma_{h,max}^{-1} - q\Sigma_{h,max}^{-1} &= 0 \\ \Sigma_{h,max} &= \frac{B_h}{q}. \end{aligned} \tag{5.3}$$

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 (5.3). To meet the goal of minimizing bias, some coarse information about the data is encoded into B_h. In QB [10], 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 [4], where the covariance estimate is regularized by the trace: $\frac{\operatorname{tr}\left(\hat{\Sigma}_{ML}\right)}{d}I$.

Chapter 4 has shown that setting $B_h = \frac{\operatorname{tr}\left(\hat{\Sigma}_{ML}\right)}{d}I$ for a distribution-based discriminant analysis outperforms Geisser's or Keehn's parameter-based Bayesian discriminant methods, and does not require crossvalidation [39]. 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 [53, 54, 18]. Data-dependent priors are also used to form proper priors that act similarly to improper priors, or to match frequentist goals [55].

It is interesting to see that the only difference between the parameter-based Bayesian QDA (4.24) and the distribution-based Bayesian QDA (4.22) is a shift of the degree of freedom q; this is true whether the distribution-based formula (4.22) is solved for using either the Fisher or Lebesgue measure. QB, which is a modern parameter-based Bayesian QDA classifier (discussed further in Section 5.1.1), 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.

5.2 Relationship Between Regularized QDA and Bayesian QDA

In this section it is shown that Friedman's regularized form for the covariance matrix [4] 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 (4.22) can be simplified to

$$\begin{aligned} E_{N_h}[N_h] &= \frac{n_h^{\frac{d}{2}}\Gamma\left(\frac{n_h+q+1}{2}\right)}{(2\pi)^{\frac{d}{2}}(n_h+1)^{\frac{d}{2}}\Gamma\left(\frac{n_h+q-d+1}{2}\right)} \frac{\left|\frac{D_h}{2}\right|^{\frac{n_h+q}{2}}}{\left|\frac{D_h}{2}+\frac{n_h}{2(n_h+1)}Z_h Z_h^T\right|^{\frac{n_h+q+1}{2}}} \\ &= \frac{n_h^{\frac{d}{2}}\Gamma\left(\frac{n_h+q+1}{2}\right)\left|\frac{D_h}{2}\right|^{\frac{n_h+q}{2}}}{(2\pi)^{\frac{d}{2}}(n_h+1)^{\frac{d}{2}}\Gamma\left(\frac{n_h+q-d+1}{2}\right)\left|\frac{D_h}{2}\right|^{\frac{n_h+q+1}{2}}} \left|I+\frac{n_h}{n_h+1}Z_h Z_h^T D_h^{-1}\right|^{-\frac{n_h+q+1}{2}} \\ &= \frac{\Gamma\left(\frac{n_h+q+1}{2}\right)}{(\pi)^{\frac{d}{2}}\left|\left(\frac{n_h+1}{n_h}\right)D_h\right|^{\frac{1}{2}}\Gamma\left(\frac{n_h+q-d+1}{2}\right)}\left(1+\frac{n_h}{n_h+1}Z_h^T D_h^{-1}Z_h\right)^{-\frac{n_h+q+1}{2}}, \quad (5.4) \end{aligned}$$

where (5.4) follows by rearranging terms and applying the identity $|I + Z_h Z_h^T D_h^{-1}| = 1 + Z_h^T D_h^{-1} Z_h$ [56]. It is easy to see using identity [57]

$$\int_{\mathbf{R}^d}\left[1+(x-\mu)^T D^{-1}(x-\mu)\right]^{-\frac{\nu+d}{2}}dx = \frac{\Gamma(\frac{1}{2})^d\Gamma(\frac{\nu}{2})}{\Gamma(\frac{\nu+d}{2})}|D|^{\frac{1}{2}}, \quad (5.5)$$

$E_{N_h}[N_h]$ (5.4) is a proper probability density function for X.

Approximate $n_h/(n_h+1) \approx 1$ in (5.4). Recall the series expansion $e^r = 1 + r + r^2/2\ldots$, so if r is small, $1 + r \approx e^r$. Apply this approximation to the term $1 + Z_h^T D_h^{-1} Z_h$ in (5.4), 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 $|Z_h^T D_h^{-1} Z_h| \leq \|Z_h\|^2/\lambda_{min}$. Then (5.4) becomes

$$\begin{aligned} E_{N_h}[N_h] &\approx \frac{\Gamma\left(\frac{n_h+q+1}{2}\right)\left(\exp\left[\frac{n_h}{n_h+1}Z_h^T D_h^{-1}Z_h\right]\right)^{-\frac{n_h+q+1}{2}}}{(\pi)^{\frac{d}{2}}\left|\left(\frac{n_h+1}{n_h}\right)D_h\right|^{\frac{1}{2}}\Gamma\left(\frac{n_h+q-d+1}{2}\right)} \\ &= \frac{\Gamma\left(\frac{n_h+q+1}{2}\right)\exp\left[-\frac{1}{2}Z_h^T\left[\frac{n_h+1}{n_h+q+1}\left(\frac{D_h}{n_h}\right)\right]^{-1}Z_h\right]}{(\pi)^{\frac{d}{2}}\left|\left(\frac{n_h+1}{n_h}\right)D_h\right|^{\frac{1}{2}}\Gamma\left(\frac{n_h+q-d+1}{2}\right)}. \quad (5.6) \end{aligned}$$

Let

$$\tilde{\Sigma}_h \triangleq \frac{n_h+1}{n_h+q+1}\frac{D_h}{n_h}. \tag{5.7}$$

The approximation (5.6) resembles a Gaussian distribution, where $\tilde{\Sigma}_h$ plays the role of the covariance matrix. Rewrite (5.7),

$$\begin{aligned}\tilde{\Sigma}_h &= \frac{n_h+1}{n_h+q+1}\left(\frac{S_h+B_h}{n_h}\right)\\ &= \frac{n_h+1}{n_h+q+1}\left(\frac{S_h}{n_h}\right)+\frac{n_h+1}{n_h+q+1}\left(\frac{B_h}{n_h}\right)\\ &= \left(1-\frac{q}{n_h+q+1}\right)\frac{S_h}{n_h}+\frac{1}{n_h+q+1}\left(\frac{n_h+1}{n_h}\right)B_h. \end{aligned}\tag{5.8}$$

In (5.8), make the approximation $\frac{n_h+1}{n_h} \approx 1$, then multiply and divide the second term of (5.8) by q,

$$\tilde{\Sigma}_h \approx \left(1-\frac{q}{n_h+q+1}\right)\frac{S_h}{n_h}+\left(\frac{q}{n_h+q+1}\right)\frac{B_h}{q}. \tag{5.9}$$

The right-hand side of (5.9) 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 [4], re-stated in this chapter in equations (5.1) and (5.2). 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 (5.3) that $\frac{B_h}{q}$ is the maximum of the prior probability distribution. Equation (5.9) 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.

5.3 Bregman Divergences and Bayesian Quadratic Discriminant Analysis

In (4.15) the Bayesian QDA class estimate that minimizes the expected misclassification cost was defined. Then, assuming a Gaussian class-conditional distribution, the expected class-conditional distribution is given in (4.22). 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

$$\hat{f}_h = \underset{f \in \mathcal{A}}{\operatorname{argmin}} \int_M R(\mathcal{N}_h, f) dM, \tag{5.10}$$

where $R(\mathcal{N}_h, f)$ is the risk of guessing f if the truth is $\mathcal{N}_h$, the set of functions $\mathcal{A}$ more precisely is defined shortly, and dM is a probability measure on the set of Gaussians, M. Equation (5.10) is a distribution-based version of the standard parameter Bayesian estimate given in [18, ch. 4]; for example, using the standard parameter Bayesian estimate, estimating a mean $\mu \in \mathbb{R}^d$ would be formulated

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

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

Given estimates of the class-conditional distributions $\{\hat{f}_h\}$ from (5.10), one can solve for the class label as

$$\tilde{Y}^* = \underset{g=1,\ldots,G}{\operatorname{argmin}} \sum_{h=1}^{G} C(g, h) \hat{f}_h(x) \hat{P}(Y = h). \tag{5.11}$$

In this section it is shown that the class estimate $\hat{Y}$ from minimizing the expected misclassification cost as defined in (4.15) is equivalent to the class estimate $\tilde{Y}^*$ from (5.11) if the risk function in (5.10) 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* [23]. Bregman divergences act on pairs of vectors. Csiszár defined a Bregman divergence between two distributions [15], 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 [14, 24]. 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, a new *functional Bregman divergence* that acts on pairs of distributions is defined in Chapter 7. 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.

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

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

Theorem 5.3.1. *(**Srivastava Gupta and Frigyik 2006**) 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 d_ϕ 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*

$$\hat{f}_h = \underset{f \in \mathcal{A}}{\operatorname{argmin}} E_{N_h}[d_\phi(N_h, f)]. \tag{5.12}$$

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

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

This theorem is a special case of Theorem 7.4.1 in Chapter 7.

Corollary 5.3.2. ***Corollary:*** *The result of (4.15) is equivalent to the result of (5.11) where each $\hat{f}_h$ comes from (5.12).*

The corollary follows directly from Theorem 5.3.1 where $r(\mathcal{N}_h)$ is the posterior distribution of $\mathcal{N}_h$ given the training samples.

5.4 The BDA7 Classifier

Distribution-based QDA with a fixed degree of freedom (as proposed by the authors in the conference paper [39]) does not require cross-validation. 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, it is proposed that that the degree of freedom of the prior should be cross-validated. Also, in preliminary experiments it was 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 also been used for regularized discriminant analysis [48] and model-based discriminant analysis [11].

Note that setting $B_h = \text{diag}\left(\hat{\Sigma}_{\text{ML}}\right)$ places the maximum of the prior at $\frac{1}{q}\text{diag}\left(\hat{\Sigma}_{\text{ML}}\right)$. It is based on intuition 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 $B_h = q\ \text{diag}\left(\hat{\Sigma}_{\text{ML}}\right)$.

The heavy tail of the prior can add too much bias to estimates. One way to reduce the tail's effect is to move the maximum of the prior closer to the zero matrix, effectively turning the prior into an exponential prior rather than a unimodal one. 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 [5], 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 cross-validation from seven options:

$$B_h = \begin{cases} q\ \text{diag}\left(\hat{\Sigma}_{(\text{pooled ML})}\right),\ \text{q diag}\left(\hat{\Sigma}_{(\text{class ML,h})}\right), \\ \frac{1}{q}\ \text{diag}\left(\hat{\Sigma}_{(\text{pooled ML})}\right),\ \frac{1}{\text{q}}\ \text{diag}\left(\hat{\Sigma}_{(\text{class ML,h})}\right), \\ \text{diag}\left(\hat{\Sigma}_{(\text{pooled ML})}\right),\ \text{diag}\left(\hat{\Sigma}_{(\text{class ML,h})}\right), \\ \frac{1}{qd}\,\text{tr}(\hat{\Sigma}_{(\text{pooled ML})})\ I. \end{cases} \tag{5.13}$$

To summarize: BDA7 uses the result (4.22) where q is cross-validated, and B_h is cross-validated as per (5.13).

5.5 Results on Benchmark Datasets

We compared BDA7 to popular QDA classifiers on nine benchmark datasets from the UCI Machine Learning Repository. In the next section simulations are used to further analyze the behavior of each of the classifiers.

QDA classifiers are best-suited for datasets where there is relatively little data, such that a simple Gaussian model is fitting. For this reason, the error rate is tracked as the percentage of samples used for training is increased. A percentage of the samples for each

class is selected randomly to use as training; when this percentage results in a fraction of training samples, the number of training samples is rounded up.

For datasets with separate training and test sets, the tables show results on the test set given different percentages of the training samples randomly drawn to train the classifier. For datasets that do not have separate training and test sets, the tables show results based on using the stated percentage of the dataset as training data, and the rest of the dataset as test data. In each case except the Cover Type dataset, each result is the average of 100 trials with different randomly chosen training samples. The average results for one hundred random trials are give in Tables 2–9. Due to the immense size of the Cover Type dataset, the average results for it, shown in Table 10, are for only 10 random trials.

5.5.1 Experimental Details for Each Classifier

BDA7 is compared with QB [10], RDA [4], eigenvalue decomposition discriminant analysis (EDDA) [11], 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 (such as for the Pen Digits dataset). 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.

The RDA parameters λ and γ were calculated and crossvalidated as in Friedman's paper [4] 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 (5.13). 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. [10]. 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 [42, 41, 44]; previous Chapter 4 has demonstrated that the inverse Wishart prior performs better than these other choices [39].

The EDDA method of Bensmail and Celeux is run as proposed in their paper [11]. Thus, the crossvalidation selects one of fourteen models for the covariance, and then the ML estimate for that model is calculated. 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 was not feasible, they were not used.

5.5.2 Results

The results are shown in Tables 2–10. 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.

5.5.3 BDA7 Performs Well

BDA7 was the best or statistically insignificant from the best for six of the nine tested datasets: Pen Digits, Thyroid, Heart Disease, Wine, Image Segmentation, and Sonar. BDA7 performed competitively for the other three datasets.

The Pen Digits dataset has 7,494 training samples and 3,498 test samples spread over 10 classes described by 34 features. Here, BDA7 does significantly better than the other QDA classifiers, followed by EDDA, then QB and then RDA.

The Thyroid dataset is a three-class problem with 215 samples described by 5 features. BDA7 performs the best.

The Heart Disease dataset from the Statlog Project Databases is a two-class problem with 270 samples described by 13 features, but three of the feature are nominal, which are removed for this experiment, leaving 10 features.

The Wine dataset has 178 samples split between three classes and described by 13 features. For most fractions of the dataset, BDA7 has less than 50% of the error of QB, but its performance is generally not significantly different than EDDA. RDA does not perform well on this dataset.

The Image Segmentation dataset from the Statlog Project Datbases is a seven-class problem with 2,310 samples described by 19 features. BDA7 and QB perform the best.

The Sonar dataset has two classes, 208 training samples and 132 test samples described by 60 features. BDA7 performs consistently better than QB or RDA, but only by a tiny amount. EDDA has more difficulty with this dataset.

For all of these datasets, the features are continuous, with the exception of four features of the Heart Disease dataset, which are ordered or binary.

5.5.4 BDA7 Performs Competitively, But Not Best

For the Pima Diabetes, Ionosphere, and Cover Type datasets, BDA7 is not the best QDA classifier, but performs competitively.

The Pima Diabetes dataset is a two-class problem with 768 samples described by 8 features. BDA7, QB, RDA, and EDDA perform similarly, but LDA achieves the lowest error on this dataset.

The Waveform dataset is three class problem with 5000 samples described by 21 features. BDA7 and RDA perform better than other classifiers.

The Ionosphere dataset 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 $B_h = I$ every time for this dataset. Given that the identity

matrix is BDA7's best choice, it is at a disadvantage compared to QB, which cross-validates a scaled identity seed matrix kI. Thus, QB does slightly better for this dataset. RDA achieves the lowest error.

The Cover Type dataset is an eight-class problem with 581,012 samples and described by 54 features, 44 of which are binary features. As is standard, the first 11,340 samples are used for training, the next 3,780 samples as a validation set, and the last 565,892 samples as test samples. In the maximum likelihood estimated matrices for QDA and LDA there are many zeros, which result in a zero determinant. Thus their 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, $B_h = I$. As in the Ionosphere dataset, this puts BDA7 at a disadvantage compared to QB, and it is not surprising that on this dataset QB does a little better. Still, the two Bayesian QDA classifiers do better than the other QDA classifiers.

5.5.5 Summary of Results

To summarize, BDA7 performed best, or statistically insignificantly different from the best, for those datasets where no class had a sample covariance with determinant zero, with the exception of the Pima dataset, where BDA7 performed around 10% worse than LDA. For datasets with zero determinants, like Ionosphere and Cover Type, BDA7 did slightly worse than QB, but both Bayesian classifiers performed relatively well in both those cases.

5.6 Simulations

In order to further analyze the behavior of the different QDA classifiers, ten simulations are compared. For each of the ten simulations, the data are drawn iid from three Gaussian class conditional distributions. Six of the simulations were originally used by Friedman to show that RDA performs favorably compared to ML linear discriminant analysis (LDA) and ML QDA [4]. 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, two full covariance matrix

Table 5.1: Pen Digits mean error rate

	Percentage Used as Training Data								
	2	3	4	5	6	7	8	9	10
BDA7	**8.82**	**7.57**	**6.64**	**6.32**	**5.91**	**5.57**	**5.34**	**5.25**	**5.16**
QB	15.64	11.48	8.97	8.22	7.34	6.76	6.35	6.03	5.97
RDA	11.50	10.10	9.16	8.64	8.24	7.87	7.86	7.56	7.45
EDDA	19.73	13.86	8.94	7.66	6.86	6.30	5.84	5.60	5.55
NM	24.75	24.26	24.24	23.56	23.38	23.20	23.18	22.99	23.05
LDA	19.33	18.38	17.99	17.68	17.51	17.58	17.42	17.51	17.50
QDA	89.62	89.62	89.62	89.62	89.62	89.62	89.62	89.62	89.62

simulations are added to Friedman's six diagonal Gaussian simulations, and for each of the full covariance matrix simulations cases of classes with the same means, and classes with different means are considered.

All of the simulation results are presented for 40 training and 100 test samples, drawn iid. The number of feature dimensions ranges from 6 to 100. The parameters for each classifier were estimated by leave-one-out crossvalidation. Each simulation was run 100 times. Thus, each result presented in this section is the average error over $10,000$ test samples.

Each classifier was trained using the cross-validation parameters described in Section 5.5.1. Because the goal in this section is analysis, when the EDDA results were uncomputable (when $n < d$), we simply marked those entries with a *uc*, rather than removing the 3rd and 4th EDDA model, as done for the benchmark results when those models were infeasible to compute.

Table 5.2: Thyroid mean error rate

	Percentage Used as Training Data								
	4	5	6	7	8	9	10	15	20
BDA7	**11.74**	**10.80**	**9.33**	**8.99**	**8.59**	**7.62**	**7.39**	**6.47**	**5.67**
QB	14.33	14.09	11.52	10.82	10.92	9.10	8.89	**6.36**	**5.60**
RDA	19.27	17.27	12.86	12.06	10.43	10.26	10.20	8.29	8.28
EDDA	15.37	15.54	10.80	10.32	**9.56**	9.75	8.56	5.97	**5.74**
NM	19.99	18.83	18.00	16.17	17.11	16.67	17.02	15.05	14.17
LDA	18.06	14.42	12.19	11.22	10.71	11.36	10.44	9.79	9.43
QDA	29.76	30.04	29.89	29.80	29.96	29.85	30.02	29.77	19.73

Table 5.3: Heart disease mean error rate

	Percentage Used as Training Data								
	2	3	4	5	6	7	8	9	10
BDA7	**38.97**	**36.17**	**32.96**	**31.65**	**28.56**	**28.89**	**27.90**	**27.49**	**25.99**
QB	43.74	42.93	41.31	37.90	34.96	34.21	32.91	31.63	31.24
RDA	45.95	46.20	43.88	37.46	35.92	34.89	34.86	32.27	31.86
EDDA	43.62	40.70	38.89	34.02	**30.58**	**29.72**	30.61	**27.90**	**26.76**
NM	41.85	41.55	42.65	40.03	39.45	39.68	38.37	37.87	38.17
LDA	44.32	44.44	44.40	36.94	33.48	32.47	29.55	28.90	28.09
QDA	44.32	44.44	44.40	44.53	44.28	44.40	44.38	42.66	41.31

Table 5.4: Image segmentation mean error rate

	Percentage Used as Training Data								
	2	3	4	5	6	7	8	9	10
BDA7	**18.65**	**14.61**	**12.94**	**12.23**	**11.78**	**11.31**	**11.10**	**11.13**	**10.83**
QB	19.44	15.30	13.78	12.82	12.35	11.70	11.39	11.45	**10.87**
RDA	21.74	20.31	18.89	18.29	17.20	16.87	16.71	16.64	16.17
EDDA	33.31	31.69	30.33	29.44	28.83	29.12	28.44	28.05	28.35
NM	33.07	31.44	29.74	29.33	28.68	28.95	28.37	27.95	28.18
LDA	85.71	85.72	85.72	85.71	85.71	85.71	85.72	85.71	85.71
QDA	85.71	85.72	85.72	85.71	85.71	85.71	85.72	85.71	85.71

Table 5.5: Wine mean error rate

	Percentage Used as Training Data							
	3	4	5	6	7	8	9	10
BDA7	**23.83**	**14.54**	**9.77**	**9.89**	**10.35**	**8.72**	8.53	7.57
QB	43.88	41.49	33.36	30.48	25.73	22.82	18.13	17.45
RDA	90.23	50.53	33.74	33.39	32.10	33.23	28.24	24.07
EDDA	49.02	30.31	**12.95**	**11.85**	**10.61**	**8.13**	**7.88**	**6.94**
NM	33.31	35.63	30.78	31.80	31.55	32.93	30.13	28.76
LDA	60.24	67.06	60.13	60.23	67.08	60.14	25.65	20.03
QDA	60.24	67.06	60.13	60.23	67.08	60.14	59.97	60.32

Table 5.6: Iris mean error rate

	Percentage Used as Training Data								
	3	4	5	6	7	8	9	10	20
BDA7	24.56	23.86	**7.38**	**8.27**	**7.21**	**6.91**	6.27	6.43	4.22
QB	**8.84**	**9.76**	**7.29**	**8.35**	**6.47**	**7.03**	6.11	**6.05**	4.13
RDA	72.41	79.20	9.54	**9.17**	**7.36**	**7.94**	6.05	**6.11**	4.37
EDDA	53.17	37.03	9.88	9.21	7.93	9.03	8.14	7.91	5.79
NM	9.85	**10.01**	8.88	9.82	8.54	9.10	8.50	8.70	7.98
LDA	66.65	66.65	9.77	**9.02**	**6.45**	**7.32**	**5.07**	**5.63**	**3.64**
QDA	66.66	66.65	66.65	66.67	66.62	66.70	46.96	42.95	7.41

Table 5.7: Sonar mean error rate

	Percentage Used as Training Data									
	5	10	15	20	25	30	35	40	45	50
BDA7	**34.63**	**31.98**	**30.14**	**27.87**	**27.39**	**25.52**	**25.67**	**24.37**	**22.92**	**22.54**
QB	38.96	35.40	**30.80**	**28.27**	**27.04**	**25.48**	**25.33**	25.02	23.69	23.17
RDA	40.05	34.36	31.06	**28.04**	**27.26**	**25.64**	**25.64**	25.67	25.49	24.80
EDDA	38.59	35.17	32.83	33.04	32.20	31.86	33.09	35.19	35.14	34.05
NM	41.26	40.00	38.00	36.64	36.41	36.45	36.73	36.76	34.88	34.76
LDA	46.68	46.69	46.61	46.56	46.65	46.34	44.47	40.13	37.68	36.19
QDA	46.68	46.69	46.61	46.56	46.65	46.34	46.35	46.82	46.87	46.91

Table 5.8: Waveform mean error rate

	Percentage Used as Training Data								
	2	3	4	5	6	7	8	9	10
BDA7	**18.01**	**17.24**	**16.80**	**16.78**	**16.32**	**16.31**	16.23	16.03	15.90
QB	26.01	24.34	22.85	21.77	20.97	20.35	19.92	19.38	19.05
RDA	18.35	**17.21**	**16.88**	**16.62**	**16.09**	**16.00**	**16.00**	**15.72**	**15.55**
EDDA	20.13	20.13	20.00	19.92	19.70	19.82	19.37	18.92	18.46
NM	20.35	20.01	20.23	20.21	20.21	20.08	20.11	20.04	20.04
LDA	28.64	24.82	22.40	21.08	19.75	19.40	18.78	18.25	17.85
QDA	32.76	26.38	23.87	22.40	21.35	20.66	20.12	19.53	19.18

Table 5.9: Pima mean error rate

	Percentage Used as Training Data								
	2	3	4	5	6	7	8	9	10
BDA7	**32.04**	**31.53**	30.65	29.52	28.90	28.94	28.05	28.03	27.87
QB	35.05	35.03	33.30	32.19	31.43	30.95	30.25	29.79	29.52
RDA	33.76	33.43	30.42	29.79	28.78	28.37	27.95	27.15	27.28
EDDA	35.62	33.70	30.72	30.91	30.03	29.17	28.71	28.16	28.85
NM	39.22	38.19	37.39	37.57	36.25	36.86	36.26	35.57	35.57
LDA	**32.74**	**31.66**	**28.81**	**27.77**	**27.22**	**26.43**	**26.21**	**25.75**	**25.56**
QDA	34.84	35.83	34.40	32.72	31.92	31.37	30.31	29.93	29.73

Table 5.10: Ionosphere mean error rate

	Percentage Used as Training Data								
	2	3	4	5	6	7	8	9	10
BDA7	**24.51**	**19.20**	**16.01**	**12.78**	**11.49**	**10.58**	11.58	10.14	10.28
QB	26.02	22.37	18.42	16.07	15.45	13.61	12.82	11.70	11.19
RDA	27.79	21.92	14.73	12.92	11.45	10.25	**9.63**	**9.12**	**9.16**
EDDA	32.72	30.03	27.54	26.14	25.46	29.15	24.95	24.11	24.40
NM	34.47	30.99	27.93	27.49	25.58	27.14	25.48	25.21	25.13
LDA	35.86	35.87	35.72	35.84	35.88	35.90	35.72	35.88	35.86
QDA	35.86	35.87	35.72	35.84	35.88	35.90	35.72	35.88	35.86

Table 5.11: Cover type mean error rate

	Percentage Used as Training Data								
	2	3	4	5	6	7	8	9	10
BDA7	**48.34**	47.87	48.32	48.31	47.88	48.79	47.70	47.47	47.98
QB	**49.47**	**45.27**	**44.80**	**44.58**	**43.64**	**44.41**	**42.67**	**42.57**	**43.74**
RDA	60.90	61.84	61.83	59.20	58.45	60.02	58.59	58.43	57.38
EDDA	81.71	79.48	81.21	83.13	82.59	83.34	82.73	80.86	82.10
NM	78.09	76.50	77.99	78.52	78.90	79.74	78.07	78.12	77.90
LDA	62.94	62.94	62.94	62.94	62.94	62.94	62.94	62.94	62.94
QDA	62.94	62.94	62.94	62.94	62.94	62.94	62.94	62.94	62.94

Table 5.12: Letter Recognition mean error rate

	Percentage Used as Training Data								
	2	3	4	5	6	7	8	9	10
BDA7	**31.21**	26.51	23.70	21.99	20.67	19.80	18.94	18.54	17.76
QB	31.57	**26.36**	**23.47**	**21.67**	**20.46**	**19.54**	**18.79**	**18.35**	**17.62**
RDA	32.92	28.11	25.73	24.51	23.64	21.89	20.45	19.55	18.54
EDDA	37.94	35.95	29.95	25.31	22.73	21.10	19.87	19.09	18.15
NM	55.89	53.56	52.26	51.16	50.61	49.56	49.43	49.01	48.78
LDA	37.79	35.74	34.77	34.01	33.63	33.22	33.00	32.80	32.47
QDA	96.10	77.36	40.65	26.94	23.68	21.79	20.43	19.54	18.53

5.6.1 Simulation Results

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. Results are shown in Table 5.13.

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 when the simulation uses a simple model built into EDDA and RDA. BDA7's performance is statistically significantly better than QB for dimensions 6 through 60, above that, the differences not significant.

Table 5.13: Case 1: Equal spherical covariance

	Number of Feature Dimensions										
	6	10	20	30	40	50	60	70	80	90	100
BDA7	12.77	11.73	13.70	21.51	24.01	26.66	27.48	27.55	34.73	33.91	34.01
QB	15.22	15.99	23.00	27.90	30.00	30.43	31.64	29.36	33.36	31.78	33.01
RDA	12.41	11.36	14.08	19.31	20.82	22.33	23.15	23.38	27.29	27.07	28.53
EDDA	11.40	10.64	11.32	**17.21**	19.00	**19.72**	**20.80**	**20.94**	**24.98**	**23.71**	**25.16**
NM	**10.16**	**9.80**	**10.03**	**16.33**	**18.45**	**19.79**	**20.40**	**20.73**	**24.14**	**23.70**	**25.28**
QDA	26.03	54.04	66.78	66.71	67.31	67.12	66.28	66.76	66.90	67.33	66.84
LDA	13.15	12.60	23.44	37.78	67.31	67.12	66.28	66.76	66.90	67.33	66.84

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. Results are shown in Table 5.14. This simulation allows EDDA and RDA to use their built-in shrinkage towards class-independent covariance matrices to outperform the Bayesian methods. BDA7 makes roughly half the errors of QB.

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

$$e_i = \left(\frac{9(i-1)}{d-1} + 1\right)^2, \quad 1 \le i \le d, \tag{5.14}$$

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

Table 5.14: Case 2: Unequal spherical covariance

	Number of Feature Dimensions										
	6	10	20	30	40	50	60	70	80	90	100
BDA7	20.35	15.46	17.95	19.78	20.16	20.23	19.61	20.79	19.22	18.76	19.94
QB	21.73	21.47	34.76	35.93	39.41	40.02	39.56	38.14	41.26	39.66	43.94
RDA	17.64	11.35	12.80	10.79	11.99	11.86	9.96	11.03	10.34	9.58	10.54
EDDA	**16.44**	**10.18**	**10.47**	**8.92**	**9.45**	**9.88**	**8.09**	**9.71**	**8.25**	**7.57**	**8.33**
NM	19.66	12.54	18.46	22.94	26.69	30.80	28.56	29.90	34.09	32.59	38.29
QDA	28.64	54.83	66.78	66.71	67.31	67.12	66.28	66.76	66.90	67.33	66.84
LDA	21.43	17.28	31.77	41.90	67.31	67.12	66.28	66.76	66.90	67.33	66.84

Table 5.15: Case 3: Equal highly ellipsoidal covariance, low-variance subspace means

	Number of Feature Dimensions										
	6	10	20	30	40	50	60	70	80	90	100
BDA7	4.30	7.89	13.24	17.29	21.01	26.42	**27.49**	**33.08**	**34.57**	**39.11**	**44.45**
QB	11.60	31.65	49.83	52.19	53.13	52.89	54.51	56.40	56.09	55.56	57.43
RDA	3.94	11.56	23.42	35.00	37.69	40.70	43.55	49.05	49.53	49.32	51.33
EDDA	**1.99**	**6.80**	**12.46**	**15.23**	**18.18**	**22.00**	uc	uc	uc	uc	uc
NM	16.73	24.65	30.76	39.68	43.77	43.95	46.32	50.05	50.31	49.45	51.93
QDA	14.71	50.59	76.07	78.62	68.07	76.00	65.00	67.00	64.00	65.00	70.00
LDA	2.89	9.13	21.14	38.00	74.54	77.22	78.32	67.00	64.00	65.00	70.00

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

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

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.

Results are shown in Table 5.15 for Case 3. 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 many 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

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

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.

Results are shown in Table 5.16 for Case 4. As in Cases 1 and 2, RDA and EDDA perform similarly with BDA7 performing slightly worse and QB performing generally worse, particularly for low dimensions. Nearest-means does the best, and RDA and EDDA use their identity covariance models to mimic nearest-means.

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 (5.14), and those of class two are given by

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

The eigenvalues of class three are given by

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

Table 5.16: Case 4: Equal highly ellipsoidal covariance, high-variance subspace means

	Number of Feature Dimensions										
	6	10	20	30	40	50	60	70	80	90	100
BDA7	9.35	**12.48**	**16.04**	19.67	25.59	25.65	28.49	29.47	33.57	37.66	42.27
QB	15.32	32.00	42.29	34.92	35.84	32.83	31.40	32.05	31.42	33.08	32.17
RDA	8.95	**13.03**	16.21	14.71	19.88	20.91	19.41	21.38	20.99	24.31	22.78
EDDA	**7.76**	**12.35**	**15.66**	13.83	18.03	18.12	17.27	19.82	19.07	22.82	21.44
NM	**7.76**	**12.33**	**15.37**	**12.98**	**16.82**	**16.71**	**16.39**	**18.56**	**16.88**	**21.69**	**19.80**
QDA	20.55	54.03	66.68	67.77	66.07	68.04	66.85	66.19	67.39	67.39	65.97
LDA	8.71	13.97	25.91	36.80	66.07	68.04	66.85	66.19	67.39	67.39	65.97

Table 5.17: Case 5: Unequal highly ellipsoidal covariance, same means

	Number of Feature Dimensions										
	6	10	20	30	40	50	60	70	80	90	100
BDA7	**15.46**	**4.35**	1.14	0.79	1.57	0.83	0.62	0.78	**0.29**	0.55	0.68
QB	20.78	16.57	25.87	28.73	25.91	33.29	32.43	40.28	38.27	40.44	36.98
RDA	25.02	18.95	11.54	15.33	11.20	13.77	10.90	14.59	13.64	14.92	14.10
EDDA	**15.66**	4.72	**0.37**	**0.12**	**0.01**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
NM	58.22	59.42	54.34	56.49	48.05	56.74	50.74	61.12	57.42	58.69	55.10
QDA	23.32	46.04	66.90	65.76	66.05	67.14	67.14	66.00	66.99	68.01	67.11
LDA	59.20	61.20	57.66	59.17	66.05	67.14	67.14	66.00	66.99	68.01	67.11

Table 5.18: Case 6: Unequal highly ellipsoidal covariance, different means

	Number of Feature Dimensions										
	6	10	20	30	40	50	60	70	80	90	100
BDA7	**6.92**	5.44	2.17	0.87	1.84	0.90	0.62	1.16	0.38	0.31	0.48
QB	7.82	9.37	19.41	21.12	23.32	22.32	26.38	29.36	31.76	29.09	26.04
RDA	10.06	11.50	7.34	10.70	10.05	6.48	8.72	10.70	9.80	9.86	8.25
EDDA	**6.29**	**3.16**	**0.39**	**0.01**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**	**0.00**
NM	39.23	40.61	42.08	40.83	39.27	36.67	41.38	45.57	45.14	43.13	38.77
QDA	11.23	48.09	66.69	66.70	67.04	67.44	67.65	67.44	66.10	66.44	65.67
LDA	39.32	40.52	45.85	49.77	67.04	67.44	67.65	67.44	66.10	66.44	65.67

Table 5.19: Case 7: Unequal full random covariance, same means

	Number of Feature Dimensions										
	6	10	20	30	40	50	60	70	80	90	100
BDA7	10.01	19.64	18.63	24.79	32.08	35.27	40.23	40.22	40.58	40.96	42.42
QB	20.44	24.47	**15.91**	**22.19**	**27.57**	**30.66**	**34.38**	**32.52**	**33.17**	**34.56**	**34.89**
RDA	**6.68**	30.29	29.44	34.58	37.49	41.54	44.16	43.99	42.96	43.84	45.73
EDDA	9.98	**18.25**	53.08	60.53	61.84	63.83	65.81	66.24	66.85	66.88	65.08
NM	64.16	66.25	65.61	66.24	65.99	66.51	66.05	67.16	65.67	66.43	66.32
QDA	7.02	46.31	67.95	67.70	67.00	66.80	66.28	66.84	67.42	66.6200	67.04
LDA	61.60	63.10	62.41	64.65	67.00	66.80	66.28	66.84	67.42	66.62	67.04

Table 5.20: Case 8: Unequal full random covariance, different means

	Number of Feature Dimensions										
	6	10	20	30	40	50	60	70	80	90	100
BDA7	4.46	1.30	5.34	9.68	8.56	13.19	12.99	14.30	18.15	17.19	16.43
QB	**2.51**	**0.27**	**2.08**	3.91	3.34	4.86	5.09	7.38	**8.72**	**7.75**	**6.47**
RDA	**2.81**	**0.23**	2.71	**3.58**	**3.90**	**4.27**	**4.11**	**6.10**	**8.89**	**6.78**	**6.19**
EDDA	5.23	0.65	6.39	14.58	17.35	27.67	48.25	50.16	52.66	52.85	50.7200
NM	17.44	24.14	28.01	38.35	45.70	47.60	47.86	49.71	51.17	51.88	49.05
QDA	7.24	33.55	67.03	66.94	64.72	67.10	66.47	66.75	66.97	66.05	66.46
LDA	3.20	0.62	6.09	19.68	64.72	67.10	66.47	66.75	66.97	66.05	66.46

Table 5.21: Case 9: Unequal full highly ellipsoidal random covariance, same means

	Number of Feature Dimensions										
	6	10	20	30	40	50	60	70	80	90	100
BDA7	2.12	5.25	0.93	1.26	**1.10**	**1.18**	**2.87**	**1.84**	**1.89**	**2.57**	**2.13**
QB	**2.21**	**4.21**	**0.61**	**0.65**	**1.01**	**0.88**	**2.64**	**1.68**	**1.84**	**2.58**	**1.89**
RDA	1.13	12.36	12.57	39.01	61.34	59.93	63.04	61.22	65.22	64.00	69.00
EDDA	0.34	2.80	25.48	27.22	37.23	32.33	62.54	66.18	69.00	64.00	69.00
NM	64.86	66.57	66.73	65.93	65.68	65.67	66.76	66.25	66.92	65.68	66.94
QDA	1.60	35.97	67.07	66.38	66.33	67.62	66.01	66.60	67.69	66.12	67.25
LDA	56.26	58.22	58.15	60.92	66.33	67.62	66.01	66.60	67.69	66.12	67.25

Table 5.22: Case 10: Unequal full highly ellipsoidal random covariance, different means

	Number of Feature Dimensions										
	6	10	20	30	40	50	60	70	80	90	100
BDA7	0.36	**0.16**	0.67	0.69	0.82	**1.01**	**1.73**	**1.58**	**0.96**	**1.38**	**1.56**
QB	**0.05**	**0.06**	**0.27**	**0.30**	**0.69**	**0.98**	**1.47**	**1.37**	**1.11**	**1.43**	**1.40**
RDA	**0.00**	2.68	5.81	30.38	59.29	62.30	63.35	64.36	62.43	61.00	62.00
EDDA	0.79	2.70	17.20	21.96	25.48	31.09	67.32	68.04	70.00	61.00	62.00
NM	51.34	60.57	66.48	65.24	66.30	66.74	66.57	66.31	66.41	66.37	64.99
QDA	**1.35**	27.71	66.57	65.65	66.88	66.73	66.93	66.99	66.86	65.06	65.74
LDA	2.05	10.92	21.98	41.31	66.88	66.73	66.93	66.99	66.86	65.06	65.74

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.

Results are shown in Tables 5.18 and 5.19. In both 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 appears to be able to use this information effectively to discriminate the classes. EDDA also achieves very low error but breaks when run for high dimensions.

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.

Results are shown in Tables 5.19 and 5.20. 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, with QB doing slightly better than BDA7 from 20 dimensions and higher. In contrast, EDDA jumps from 15% error at 10 feature dimensions to 55% error at 20 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. In this case for dimensions greater than twenty, BDA7 chooses the prior seed matrix to be the trace-scaled identity matrix, whereas QB cross-validates the identity matrix as its prior seed matrix which gives it the edge.

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.

Results are shown in Tables 5.21 and 5.22. The two Bayesian methods perform similarly, but RDA, EDDA, and nearest-means have a very difficult time discriminating the classes.

5.7 Conclusions

In this chapter, it has been shown how a distribution-based formulation of the Bayesian quadratic discriminant analysis classifier relates to the standard parameter-based formula-

tion, 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 chapter is the classifier BDA7, which has been shown to perform generally better than RDA, EDDA and QB over nine benchmark datasets. Key aspects of BDA7 are that the seed matrix in the inverse Wishart prior defines the maximum of the prior, and that using a coarse estimate of the covariance matrix as the seed matrix pegs 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.

In cases where BDA7 chooses the identity matrix for the prior seed, $B_h = \frac{\mathrm{tr}(\hat{\Sigma}_{ML})}{d} I$, BDA7 usually performs a little worse than QB, because QB cross-validates a scaled identity matrix kI for the prior seed. This is the case in the Ionosphere and Cover Type benchmark datasets, where QB performs better than BDA7. This issue can be fixed in BDA7 by cross-validating k in $B_h = kI$ as the seventh model, rather than using the $B_h = \frac{\mathrm{tr}(\hat{\Sigma}_{ML})}{d} I$.

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. QDA has too much model bias to be a general purpose classifier, but 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

The work in this chapter was funded in part by the Office of Naval Research, Code 321, Grant # N00014-05-1-0843. We thank Richard Olshen and Inderjit Dhillon for helpful

discussions.

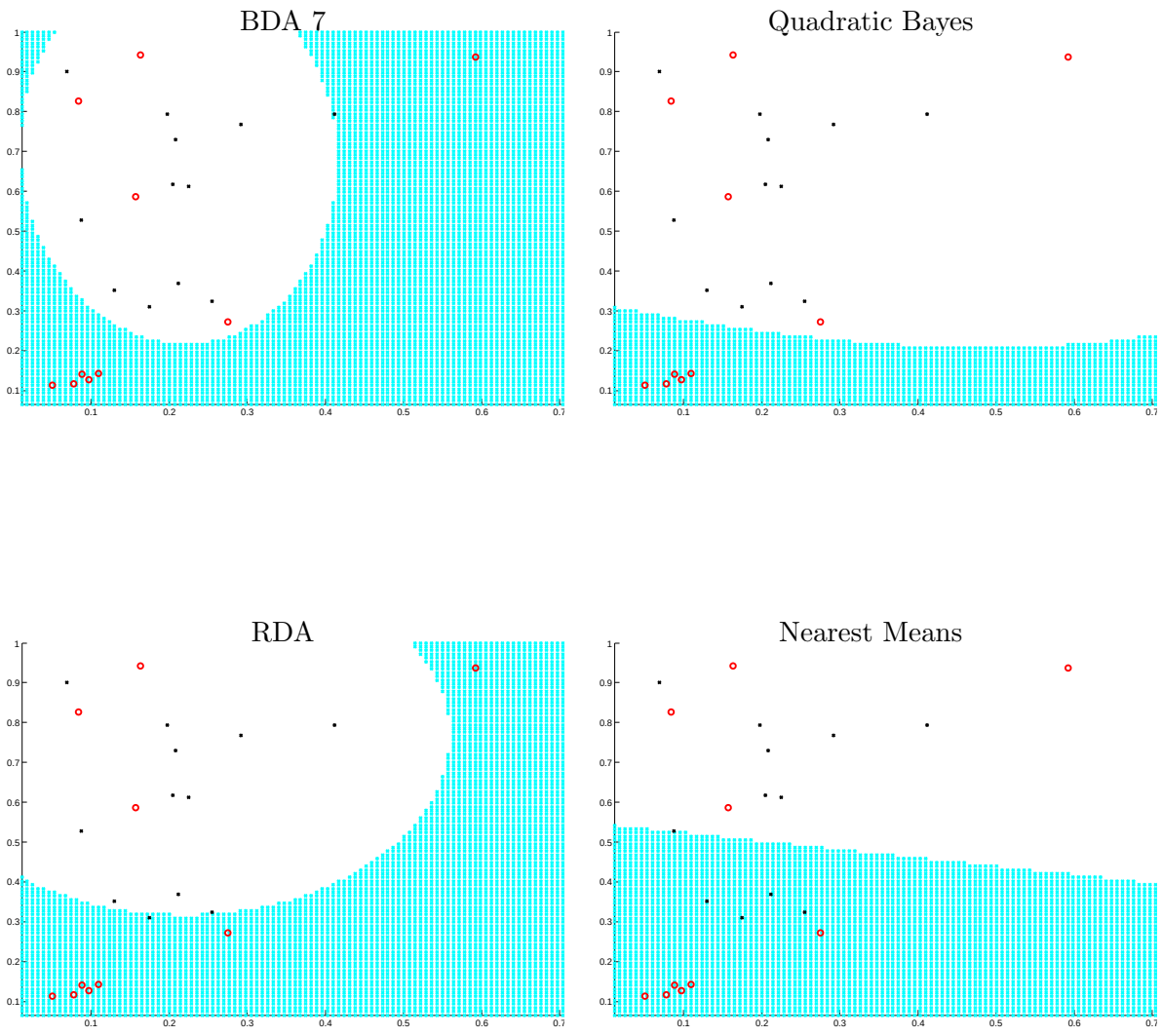

Figure 5.1: Examples of two-class decision regions for different classifiers. Features 11 and 21 from the Sonar UCI dataset were used to create this two-dimensional classification problem for the purpose of visualization; the training samples from class 1 are marked by red 'o' and the training samples from class 2 are marked by black '·'.

LDA QDA

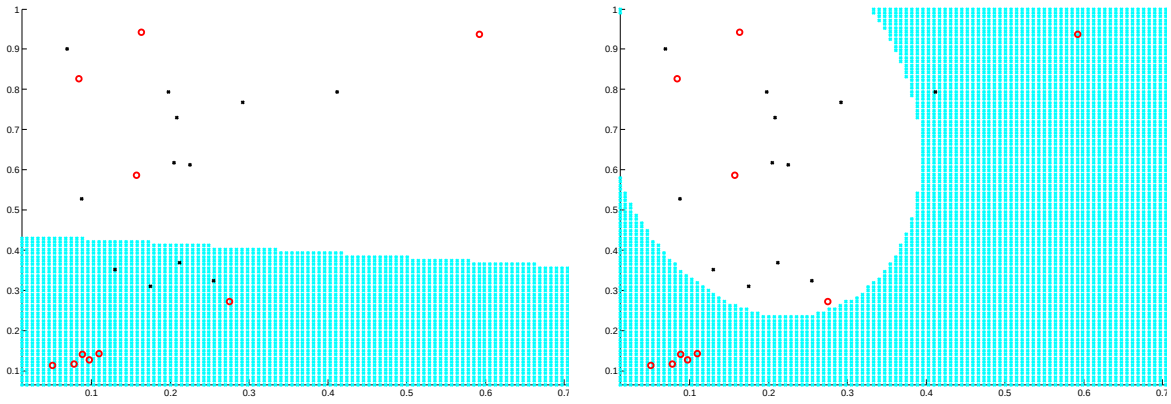

Figure 5.2: Examples of two-class decision regions for different classifiers LDA and QDA. Features 11 and 21 from the Sonar UCI dataset were used to create this two-dimensional classification problem for the purpose of visualization; the training samples from class 1 are marked by red 'o' and the training samples from class 2 are marked by black '·'.

Chapter 6

LOCAL BAYESIAN QUADRATIC DISCRIMINANT ANALYSIS

Local classifiers, such as k-NN, base decisions only on a local neighborhood of the test sample, and do not need to make assumptions about the smoothness of the entire feature space. This chapter explores model-based local classifiers and proposes a different solution to achieve flexible model-based classification. It applies the distribution-based Bayesian QDA classifier locally to the neighborhood that is formed by the k samples from each class that are closest to the test sample. Only the neighborhood size k must be cross-validated. It is shown that a local distribution-based Bayesian QDA classifier can achieve low error and is a simple alternative to Gaussian mixture models. First, related work in local model-based classification is reviewed in Section 6.1. In Section 6.2 the proposed local BDA classifier is described. Experimental results in Section 6.3 show that the proposed local Bayesian quadratic discriminant analysis classifier can achieve higher accuracy than the local nearest means classifier, a local support vector machine (SVM), a Gaussian mixture model, k-NN, and classifying by local linear regression. A discussion section concludes the chapter. The results in this chapter have been submitted for publication [58].

6.1 Background

Local classifiers, such as k-NN, make decisions based only on a subset of training data that are local to the test sample. Local classifiers can achieve competitive error rates on practical problems [5, 59, 60, 61], can automatically incorporate additional training datasets, do not require training a classifier on the entire labeled dataset, do not require smoothness assumptions about the feature space, and are generally easy to implement and train. Additionally, there is some evidence in psychology literature that people make decisions based on examples, and that learning is a function of local examples [62]. It is observed in psychophysics that humans can perform coarse categorization quite quickly: when presented

with an image, human observers can answer coarse queries such as presence or absence of an animal in as little as 150ms, and of course can tell what animal it is, given enough time [63]. The process of a coarse and quick categorization, followed by successive finer but slower discrimination motivates the modeling of such process in the setting of statistical learning. We use k-NN as an initial pruning stage and perform local BDA on the smaller but more relevant set of examples that require careful discrimination. Traditionally, local classifiers have been nonparametric weighted nearest-neighbor voting methods [64, 5], though recently, model-based local classifiers have been shown to be promising.

6.1.1 Related Work in Local Model-based Classification

The most closely related algorithm to the proposed local BDA is the idea of Mitani and Hamamoto to classify a test sample as the class whose local mean is nearest [65, 12]. To ensure that one always has samples from each class, they define the neighborhood to be the k nearest samples of each class, where k is cross-validated. Compared to the standard nearest-means classifier, their local nearest means classifier drastically reduces the immense bias inherent in modeling each class as being characterized by its mean feature vector. Compared to standard k-NN, their motivation is to reduce the classification variance due to outlying samples. The local nearest means classifier creates a locally linear decision boundary.

Another related classifier is local similarity discriminant analysis (local SDA) [66], which may be the first classifier to locally model the class-conditional distributions. Local SDA is a similarity-based classifier, but its model for the local class-conditional distributions is the exponential distributions, which is the maximum entropy distribution given mean constraints. In contrast, the proposed local BDA acts on metric spaces, models class-conditional distributions as Gaussians (which are the maximum entropy distributions given mean and covariance constraints), and uses Bayesian rather than maximum likelihood estimation.

Another set of algorithms models the decision boundary as locally linear. For example, a recent paper proposed applying a SVM locally [13]. Here the primary motivation is to reduce the difficulty in training SVMs. Training an SVM only on a local neighborhood uses

fewer samples, and for large multi-class problems there tend to be fewer different classes in the neighborhood, which reduces the number of pairwise classifications needed for multiclass SVM. In fact, the results showed that the accuracy of the SVM-kNN was very similar to the standard DAG-SVM accuracy on USPS, CUReT, and Caltech-101 datasets. The authors of the SVM-KNN paper note that while nonlinear kernels could be used, they used a linear kernel. Based on experimentation with a few small datasets, we found that the linear kernel worked better than a radial basis function kernel. In general, one expects simpler classifiers to work better locally because estimation variance tends to be a more serious problem than bias when only a few samples are used for learning. Using a linear kernel means that the SVM-KNN locally fits a hyperplane to the given feature space, where the fit maximizes the margin for the training data in the neighborhood.

Locally linear decision boundaries can also be created by a least-squares local linear regression, which was shown to be a statistically consistent classifier in 1977 [26]. Local linear regression is usually formulated for the two-class problem as calculating the discriminant $f(x) = \beta^T x$, where β is the slope vector for the least-squares hyperplane fit to the training sample pairs in the neighborhood of the test sample x. For multiclass classification it is computationally easier to formulate local linear regression as a weighted nearest-neighbor classifier with weight vector $w^* = x(X^T X)^{-1} X^T$, where the ith row of X is the ith training feature vector x_i, and one classifies as the class that receives the most weight [26]. This formula for w^* uses the Moore-Penrose pseudoinverse of X and in the case of non-uniqueness yields the w^* solution with minimum ℓ_2 norm. In the experiments section we call this psuedoinverse implementation *pinv*. A related approach is local ridge regression, which regularizes the least-squares error with a squared norm penalty, but this requires a penalty parameter to be chosen or cross-validated; Bottou and Vapnik showed that local ridge regression can achieve high accuracy [61].

Vincent and Bengio proposed a modified nearest neighbor algorithm called k-local hyperplane distance nearest neighbor (HKNN) [7]. The basic HKNN algorithm uses k nearest samples of each class as the neighborhood for a given test sample, and for each class calculates the linear subspace that passes through all the training samples of that class, then classifies as the class with the nearest hyperplane. The intuition behind HKNN is the linear

subspaces generate many "fantasy" training examples. Like pinv and SVM-KNN, HKNN creates a locally linear decision boundary. Flexible locally linear decision boundaries can also be created with decision trees; for example, the recent FlexTrees classifier [67].

Local models have also been used to adapt the metric used for local classification. For example, Hastie and Tibshirani iteratively fit a Gaussian model to the neighborhood training samples of the test sample using a local linear discriminant analysis, and then a uniform or weighted k-NN classifier is applied using the metric implied by the locally-fitted Gaussian model [68]. Other efforts have used non-Gaussian assumptions to locally adapt the metric [69, 70].

6.2 Local BDA Classifier

In this section the local BDA classifier is proposed and new analysis of (4.22) is given in order to better analyze local BDA.

For local BDA, we will estimate one local model per class from k neighbors of each class, and so we treat the local class prior probabilities as equal: $P(Y = h) = P(Y = g)$ for all $g, h \in \{1, \ldots, G\}$. Define the local BDA classifier to be the classifier that results from applying the BDA classifier given in (4.22) to the neighborhood that is composed of the k samples of each class that are nearest to the test sample x with fixed parameters $q = d + 3$ and $B_h = I$. Then the BDA classifier (4.22) has a closed-form solution:

$$\begin{aligned}
\hat{Y} &= \arg\min_g \sum_{h=1}^{G} C(g,h) \frac{\Gamma\left(\frac{k+d+4}{2}\right)\left(1+\frac{k}{k+1}(x-x_h)^T(S_h+I)^{-1}(x-x_h)\right)^{-\frac{k+d+4}{2}}}{(\pi)^{\frac{d}{2}}\Gamma\left(\frac{k+4}{2}\right)\left|\left(\frac{k+1}{k}\right)(S_h+I)\right|^{\frac{1}{2}}} \\
&\equiv \arg\min_g \sum_{h=1}^{G} C(g,h) \frac{\left(1+\frac{k}{k+1}(x-\bar{x}_h)^T(S_h+I)^{-1}(x-\bar{x}_h)\right)^{-\frac{k+d+4}{2}}}{|S_h+I|^{\frac{1}{2}}}, \qquad (6.1)
\end{aligned}$$

where $\bar{x}_h$ is the hth class sample mean, and $S_h = \sum_{i=1}^{n}(x_i - \bar{x}_h)(x_i - \bar{x}_h)^T I_{(y_i=h)}$, where $I_{(\cdot)}$ is the indicator function.

For the local BDA, the number of neighbors k is chosen by cross-validation, such that for G different classes, there are k neighbors from each class in total. Note, only one parameter k is cross-validated irrespective of the number of classes. Thus, local BDA forms a flexible

generative classifier with only one parameter (the neighborhood size) to estimate for a given classification problem.

For a test sample x, the local BDA classifier is implemented in two steps. **Step 1:** Find the k nearest training samples of class h to x using Euclidean distance. **Step 2:** Estimate the class label $\hat{Y}$ as per (6.1). Note that (6.1) is a closed-form solution and no parameters must be cross-validated. For local BDA the number of neighbors k must be cross-validated, such that if there are G different classes, then there are $k \times G$ neighbors used to classify a test sample x.

There are three main motivations for local BDA. First, we hypothesize that many practical learning problems do have local structure, and that a generative local model can be an effective classifier.

The second motivation is that one expects that each class-conditional distribution will differ, and thus we model each class's local class-conditional distribution individually. Our assumption that the local class-conditional distributions have different shapes is different than the work of Hastie and Tibshirani, as they assume the shape of the class posterior distributions is locally the same (but with different means) [68]. There has been recent research into classification of samples that lie on lower-dimensional manifolds within a feature space. Such methods may first attempt to estimate the lower-dimensional manifold, then classify on the manifold. Local BDA has the capability to adapt to the local shape of such a manifold. Because each class is modeled by its own Gaussian, there is no assumption that different classes lie on the same lower-dimensional manifold.

The third motivation for the local BDA method is the same as Hamamoto et al.'s motivation in their work [12, 71]: to reduce the effect of outlier samples. The local means classifier will be more robust to outlier samples than the local BDA, but at the cost of an increased model bias. Fitting a full Gaussian to each class, as done by local BDA, is a more flexible model. However, the robust Bayesian estimation used to fit the covariance matrix restricts the sensitivity of the classifier to any one training sample.

6.2.1 The Local BDA Decision Boundary

Figure 6.1 shows an example of the local BDA decision boundary for a two-class toy problem with a two-dimensional feature space. In general, the estimated distributions $E_{\mu_h,\Sigma_h}[\mathcal{N}(x;\mu_h,\Sigma_h)]$ in Bayesian QDA are not Gaussians [39], and so the shape of the Bayesian QDA decision boundary is not quadratic. However, we show here that the local Bayesian QDA decision boundary is in fact quadratic, which happens because we use the same number of samples from each class. Assume a two-class problem and zero-one costs; then the decision boundary for the Bayesian QDA classifier is defined by the x such that $E[N_1] = E[N_2]$. From (6.1) the decision boundary can be described as

$$[1 + \frac{k}{k+1}(x - \bar{x}_1)^T D_1^{-1}(x - \bar{x}_1)]^{\frac{\tilde{k}}{2}} = \gamma_{db}[1 + \frac{k}{k+1}(x - \bar{x}_2)^T D_2^{-1}(x - \bar{x}_2)]^{\frac{\tilde{k}}{2}}, \qquad (6.2)$$

where $\tilde{k} = k + d + 4$, $D_h = S_h + I$, and γ_{db} is a constant that depends on the training samples and the number of training samples, but γ_{db} does not depend on the test sample x. Raise both sides of (6.2) to the power $\frac{2}{\tilde{k}}$, and because the exponentiated terms must always be positive, it must be that the decision boundary is completely described by the x that solve,

$$\left(1 + \frac{k}{k+1}(x - \bar{x}_1)^T D_1^{-1}(x - \bar{x}_1)\right) = \tilde{\gamma}_{db}\left(1 + \frac{k}{k+1}(x - \bar{x}_2)^T D_2^{-1}(x - \bar{x}_2)\right), \qquad (6.3)$$

where $\tilde{\gamma}_{db} = \gamma_{db}^{\frac{2}{\tilde{k}}}$. Since (6.3) is quadratic in x, the local BDA decision boundary is locally quadratic over any region of the feature space where the neighborhood is constant.

6.3 Experiments

We compared the local BDA classifier with the local nearest-means classifier, k-NN, the pinv classifier (that is, the local linear regression classifier), the SVM-KNN, and a GMM. All of the datasets come from the UCI Machine Learning Repository except the standard USPS recognition set with 7291 training samples and 2709 test samples which is from cervisia.org/machine_learning_data.php. For all the experiments, misclassification costs are taken to be $C(g,h) = 1$ if $g \neq h$ and $C(g,h) = 0$ for $g = h$.

For all of the classifiers except the GMM, the only parameter cross-validated is a neighborhood size parameter k, where $k \in \{1, 2, \ldots, 20, 30, 40, \ldots, 100\}$, unless there were fewer

neighbors of one class, in which case the maximum k was taken to be the maximum number of neighbors in the smallest class. For k-NN, pinv, and SVM-KNN the k is the number of neighbors, while for local nearest means and local BDA, the number of neighbors is $k \times G$. All neighbors were calculated in terms of Euclidean distance. There are many ways to implement a GMM. We chose the number of mixture components by cross-validation, where the maximum number of components was $c = \min_h \text{floor}(n_h/d)$, and each class was modeled as a mixture with c components. The mixture weights, means, and full covariances were estimated using the EM algorithm. Occasionally, the EM algorithm produces estimated Gaussians with ill-posed covariance matrices, in these cases we regularized the covariance matrix by adding $10^{-6}I$. The SVM-KNN was implemented with libsvm [72] using a linear kernel and the DAG approach for multiclass classification as in [13], and all other options set as the defaults. Unless marked *raw*, the datasets were normalized before classification in the standard way by shifting each feature to have zero-mean, then dividing each feature by its standard deviation. If there were both training and test datasets, the normalizing means and standard deviations were calculated only from the training data but applied to both the training and test data.

We used a randomized 10-fold cross-validation: For each of 100 runs, the training dataset was randomly divided into a set with 9/10 of the data, and a set with 1/10 of the data. The 9/10 data set was used to build models for each of the choices of the parameter k or number of mixture components for the GMM, and each model was then tested on the remaining 1/10 data set. For each parameter setting, the cross-validation error is the average error on the 100 randomly drawn 1/10 datasets. The best cross-validation error is reported in Table 1 if only one dataset was available. If a separate test set was available, then the test error is reported in Table 2, where the choice of k or the number of mixture components was chosen using the above-described randomized 10-fold cross-validation on the training set.

6.3.1 Results

Results are shown in Tables (6.1) and (6.2) and in Figure 6.1, which illustrates the different decision boundaries for a toy example.

Table 6.1: 10-fold randomized cross-validation errors

	k-NN	pinv	Local Nearest Means	Local BDA	SVM-KNN	GMM
Wine	3.25	2.94	1.86	**0.44**	0.94	7.06
Iris	3.40	2.46	3.27	**2.13**	3.93	2.93
Glass	26.89	26.89	25.78	**23.00**	26.89	47.39
Heart	20.15	22.07	**19.41**	21.18	21.41	27.70
Sonar	13.10	10.35	12.05	**10.05**	10.15	22.90
Ionosphere	12.24	9.65	8.59	**7.41**	7.85	13.21
Thyroid	3.86	3.86	3.24	**2.90**	3.86	7.80

Figure (6.1) shows the local BDA and local nearest means for $k = 5$, which corresponds to a neighborhood of 10 training samples. For comparison, we show the k-NN, pinv, and SVM-KNN decision regions for both $k = 5$ neighbors and $k = 10$ neighbors. Classifying based on larger neighborhoods can be expected to yield smoother decision boundaries, which is true for this pinv and SVM-KNN example, but only somewhat the case for the k-NN classifier. Comparing the smoothness to the local nearest means and local BDA classifiers is difficult because the neighborhood definitions are different, but it was expected that the model-based classifiers would be more robust to outliers, and in fact the decision boundaries appear smooth, suggesting robustness to small changes in the neighbors.

For the seven datasets in Table (6.1), local BDA or local nearest means achieved the lowest cross-validation error, and in every case local BDA performed better than pinv or SVM-KNN, though sometimes only by a small amount. The test results in Table 2 show a more mixed story, with local BDA achieving the best error in 3 of the 7 cases, including an 18% improvement in Letter Recognition over the second best performer (SVM-KNN) and

an 11% improvement in Pen Digits over the second best performer (pinv). We have included results on the Vowel data set without normalization (marked *raw*) for ease of comparison with the Vowel results in the Hastie et al. book [5, p. 396]. Their results match ours for k-NN, and show that the best performance on this dataset is 39% and is achieved by a reduced-dimension version of flexible discriminant analysis.

Although the local nearest means performs very well on some datasets, on other datasets it is the worst local classifier, for example on Image Segmentation. The closed-form pinv classifier is only worse than k-NN for two datasets, and we propose that it is a simple classifier that is a more formidable baseline than k-NN. The SVM-KNN only loses to k-NN for the Iris dataset, and often offers a large gain over k-NN. For many of these cases, the GMM classifier was an unfortunate mix of too much model bias and too much estimation variance due to the number of parameters that have to be learned.

While we have tried to make the neighborhood cross-validation choices equitable, the two different neighborhood definitions are simply not directly comparable. For a fixed choice of k, the k-NN, pinv and SVM-KNN classifiers have fewer training samples to learn from than local nearest means or local BDA. However the k-nearest neighbors gives finer control over the size of the neighborhood, and guarantees that all of their neighborhood training samples are actually local, and hence more likely to be relevant to the test sample. The local nearest means and local BDA classifiers can be expected to have lower estimation variance because their neighborhood sizes are bigger, but this can also increase bias. One dataset where the different neighborhood definitions may be a factor is the Letter Recognition dataset (26 classes), on which local BDA performs well but chooses $k = 20$, giving it $20 \times 26 = 520$ training samples. The pinv classifier and k-NN are content with small neighborhoods of $k = 6$ and $k = 1$ respectively, but the SVM-KNN cross-validates to $k = 100$, suggesting it might have preferred an even larger neighborhood size. On other datasets where local BDA performs best it does not appear to be a factor. For example on the Vowel dataset local BDA uses 55 samples, but the other local classifiers do not choose a comparable k, rather they choose $k = 1$ or 2. Lastly, due to the model bias of the local nearest means and local BDA classifiers we expect they already have relatively low estimation variance (as seen, for example, in Figure (6.1)), and thus the variance-reducing advantage of their larger

Table 6.2: Test errors using 10-fold randomized cross-validated neighborhood size/number of components

	k-NN	pinv	Local Nearest Means	Local BDA	SVM-KNN	GMM
Letter Rec.	5.20	4.58	4.43	**3.23**	3.93	12.20
Optical Digits	3.23	2.78	2.73	2.78	**2.67**	9.29
Pen Digits	2.77	2.06	2.29	**1.89**	2.14	13.95
Image Seg.	12.76	**10.52**	13.67	10.95	11.52	16.86
USPS (raw)	5.88	**4.53**	4.68	**4.53**	4.88	50.88
Vowel	49.78	49.78	**43.72**	44.59	49.78	62.34
Vowel (raw)	43.72	48.05	**38.53**	39.18	43.72	62.99

neighborhood size choices may not be important.

6.4 Discussion

The local BDA classifier is a closed-form, flexible classifier that employs robust local Gaussian modeling. The experiments showed that local BDA can achieve high-accuracy. Of the local classifiers compared in this chapter, the simplest are k-NN and the local means classifier, followed by the closed-form classifiers pinv and local BDA, and then the local SVM, which requires optimization of the SVM objective and pairwise comparisons for multiclass classification. In terms of computation time, we found that the k-NN, local means, pinv and local BDA classifiers completed almost instantly, while the local SVM required nontrivial processing time.

If a generative classifier is desired, then local BDA is a simple but effective alternative to GMM classifiers and to flexible discriminant analysis [5]. GMMs can have high variance due to their sensitivity to the number of Gaussian components used. In contrast, the local BDA classifier specified requires only the neighborhood size to be trained, and is fairly robust to changes in the neighborhood.

Table 6.3: Cross-validated k or number of components

	k-NN	pinv	Local Nearest	Local BDA BDA	SVM-KNN	GMM # comp's
Glass (6 classes)	1	1	6	6	1	1
Heart (2 classes)	60	100	70	100	13	5
Image Seg. (7 classes)	6	4	8	20	18	1
Ionosphere (2 classes)	1	100	6	90	70	2
Iris (3 classes)	13	14	3	40	16	2
Letter Rec. (26 classes)	1	6	4	20	100	2
Optical Rec. (10 classes)	3	11	2	19	80	2
Pen Digits (10 classes)	1	60	2	14	40	2
Sonar (2 classes)	1	6	2	7	30	2
Thyroid (3 classes)	1	1	3	10	2	2
USPS (10 classes)	1	15	4	30	90	1
Vowel (11 classes)	1	1	2	5	1	2
Vowel (raw)	1	5	2	2	1	2
Wine (3 classes)	18	40	6	40	16	2

The term "local classifier" implies that the training samples used to make the classification decision will be in a tight region about the test sample. However, even for the 1-NN classifier this is not necessarily the case for high-dimensional feature spaces, as randomly drawn points in high-dimensions tend to be equally spread apart [73, 5]. For the model-based classifiers compared in this work, some of the cross-validated neighborhood sizes are large portions of the available training data. Rather than interpreting local classifiers as including relevant near samples, it may be more accurate to frame these classifiers as excluding irrelevant far samples.

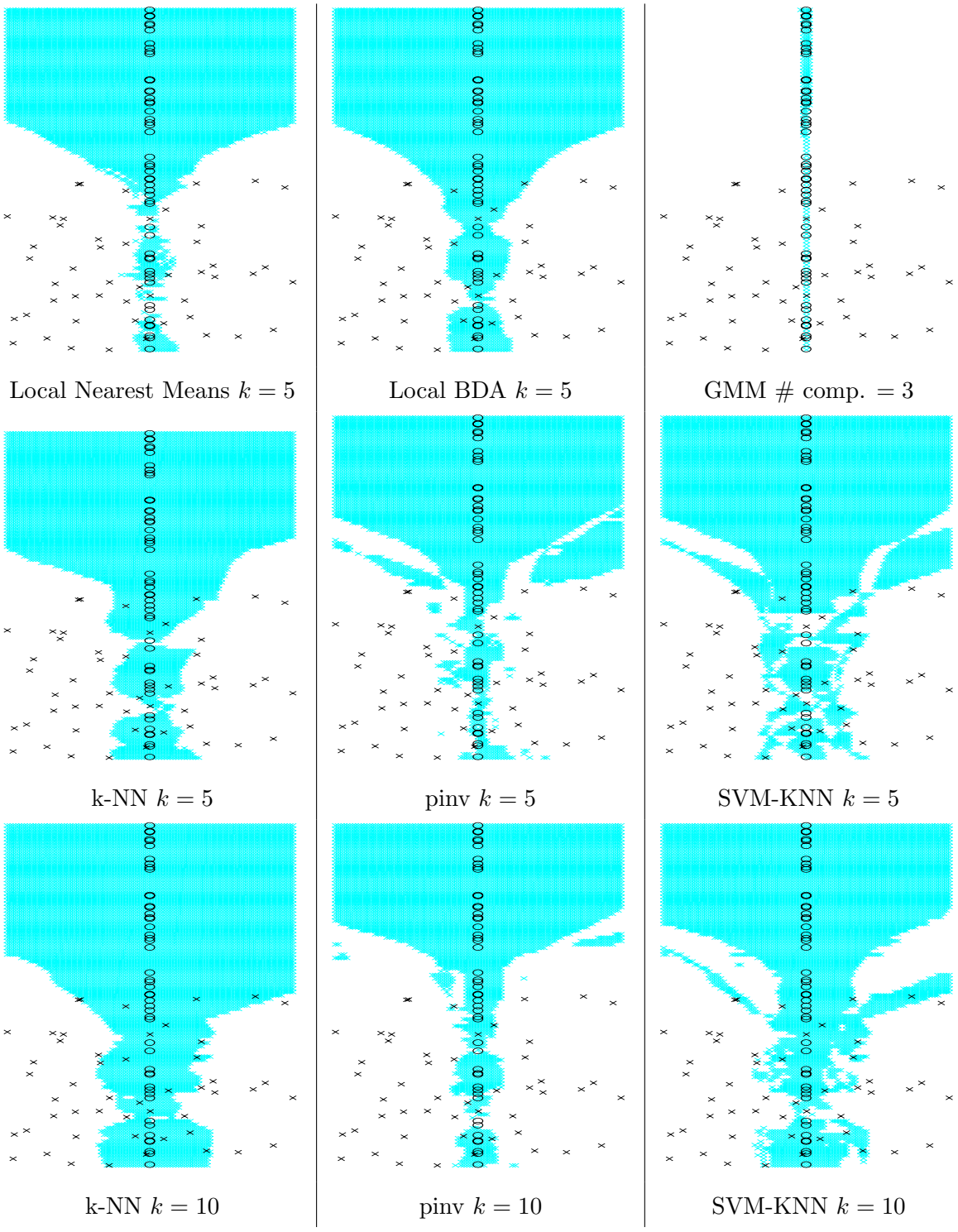

Figure 6.1: Illustrative two-dimensional examples of classification decisions for the compared classifiers. The training samples are the same for each of the examples, and are marked by circles and crosses, where the circles all lie along a line. The shaded regions mark the areas classified as class circle.

Chapter 7

FUNCTIONAL BREGMAN DIVERGENCE AND BAYESIAN ESTIMATION OF DISTRIBUTIONS

This chapter defines a functional Bregman divergence that generalizes popular distortion measures between distributions and functions, including squared error, squared bias, and relative entropy. The new definition generalizes the standard vector-based definition of Bregman divergence, and generalizes a previous pointwise Bregman divergence defined for functions. A recent result showed that the mean minimizes the expected vector Bregman divergence. The new functional definition enables the extension of this result to the continuous case to show that the mean minimizes the Bregman divergence for a set of functions or distributions. This theorem has direct application to the Bayesian estimation of distributions (as opposed to the Bayesian estimation of parameters of distributions). Estimation of the uniform distribution from independent and identically drawn samples is used as a case study of Bayesian distribution estimates, where the estimated distribution is either unrestricted or is restricted to be itself a uniform distribution.

Section 7.1 reviews the background information about Bregman divergence. In Section 7.2 the new functional definition of the Bregman divergence is given, and examples are shown for total squared difference, relative entropy, and squared bias. The relationship between the functional definition and previous Bregman definitions is established. Section 7.3 shows that the functional Bregman divergence has many of the same properties as the standard vector Bregman divergence. Section 7.4 presents the main theorem: that the expectation of a set of functions minimizes the expected Bregman divergence. Section 7.5 discusses the role of this theorem in Bayesian estimation, and as a case study compares different estimates for the uniform distribution given independent and identically drawn samples. For ease of reference for the reader, Appendix B contains relevant definitions and results from functional analysis and the calculus of variations. Proofs are in Appendix A.

The results in this chapter have been submitted for journal publication [74].

7.1 Background

Bregman divergences are a useful set of distortion functions that include squared error, relative entropy, logistic loss, Mahalanobis distance, and the Itakura-Saito function. Bregman divergences are popular in statistical estimation and information theory. Analysis using the concept of Bregman divergences has played a key role in recent advances in statistical learning [75, 76, 77, 78, 79, 39, 80, 25, 81], clustering [24, 82, 80], inverse problems [83], maximum entropy estimation [84], and the applicability of the data processing theorem [85]. Recently, it was discovered that the mean is the minimizer of the expected Bregman divergence for a set of d-dimensional points [14, 24].

7.2 Functional Bregman Divergence

Let $\left(\mathbb{R}^d, \Omega, \nu\right)$ be a measure space, where ν is a Borel measure d is a positive integer, and define a set of functions $\mathcal{A} = \left\{a \in L^p(\nu) \,\text{subject to}\, \mathbb{R}^d \to \mathbb{R},\ a \geq 0\right\}$ where $1 \leq p \leq \infty$.

Definition 7.2.1 (Functional Definition of Bregman Divergence)**.** *Let $\phi : L^p(\nu) \to \mathbb{R}$ be a strictly convex twice continuously Fréchet differentiable functional. The Bregman divergence $d_\phi : \mathcal{A} \times \mathcal{A} \to [0, \infty)$ is defined for all $f, g \in \mathcal{A}$ as*

$$d_\phi[f, g] = \phi[f] - \phi[g] - \delta\phi[g; f - g], \tag{7.1}$$

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

Here the Fréchet derivative has been used, but the definition (and results in this paper) can be easily extended using more general definitions of derivatives; an example extension is given in Section 7.2.1.

The functional Bregman divergence has many of the same properties as the standard vector Bregman divergence, including non-negativity, convexity, linearity, equivalence classes, linear separation, dual divergences, and a generalized Pythagorean inequality. These properties are detailed and established in Section 7.3.

7.2.1 Examples

Different choices of the functional ϕ lead to different Bregman divergences. Illustrative examples are given for squared error, squared bias, and relative entropy. Functionals for other Bregman divergences can be derived based on these examples, from the example functions for the discrete case given in Table 1 of [14], and from the fact that ϕ is a strictly convex functional if it has the form $\phi(g) = \int \tilde{\phi}(g(t))dt$ where $\tilde{\phi} : \mathbb{R} \to \mathbb{R}$, $\tilde{\phi}$ is strictly convex and g is in some well-defined vector space of functions [86].

Example: Total Squared Difference

Let $\phi[g] = \int g^2 d\nu$, where $\phi : L^2(\nu) \to \mathbb{R}$, and let $g, f, a \in L^2(\nu)$. Then

$$\begin{aligned} \phi[g+a] - \phi[g] &= \int (g+a)^2 d\nu - \int g^2 d\nu \\ &= 2\int gad\nu + \int a^2 d\nu. \end{aligned}$$

Because

$$\frac{\int a^2 d\nu}{\|a\|_{L^2(\nu)}} = \frac{\|a\|^2_{L^2(\nu)}}{\|a\|_{L^2(\nu)}} = \|a\|_{L^2(\nu)} \to 0$$

as $a \to 0$ in $L^2(\nu)$,

$$\delta\phi[g; a] = 2\int gad\nu$$

is a continuous linear functional in a. Then, by definition of the second Fréchet derivative,

$$\begin{aligned} \delta^2\phi[g; b, a] &= \delta\phi[g+b; a] - \delta\phi[g; a] \\ &= 2\int (g+b)ad\nu - 2\int gad\nu \\ &= 2\int bad\nu, \end{aligned}$$

thus $\delta^2\phi[g; b, a]$ is a quadratic form, where $\delta^2\phi$ is actually independent of g and strongly positive (which implies that ϕ is strictly convex):

$$\delta^2\phi[g; a, a] = 2\int a^2 d\nu = 2\|a\|^2_{L^2(\nu)}.$$

Then,

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

Example: Squared Bias

Under definition (7.1), squared bias is a Bregman divergence, which has not previously seen noted in the literature despite the importance of minimizing bias in estimation [5]. In this example the functional ϕ cannot be defined using the pointwise Bregman divergence definition given previously for functions [80, 15] (see (7.8)), if the measure ν is such that there are two disjoint measurable sets with positive measure. For example, all infinite but σ-finite Borel measures satisfy this property.

Let $\phi[g] = \left(\int g d\nu\right)^2$, where $\phi : L^1(\nu) \to \mathbb{R}$. In this case

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

and therefore

$$\delta\phi[g;a] = 2\int g d\nu \int a d\nu.$$

This follows from the fact that $2\int g d\nu \int a d\nu$ is a continuous linear functional on $L^1(\nu)$ and $\left(\int a d\nu\right)^2 \leq \|a\|^2_{L^1(\nu)}$, so that

$$0 \leq \frac{\left(\int a d\nu\right)^2}{\|a\|_{L^1(\nu)}} \leq \frac{\|a\|^2_{L^1(\nu)}}{\|a\|_{L^1(\nu)}} = \|a\|_{L^1(\nu)} \to 0$$

as $a \to 0$ in $L^1(\nu)$. Then, by the definition of the second Fréchet derivative,

$$\begin{aligned} \delta^2\phi[g;b,a] &= \delta\phi[g+b;a] - \delta\phi[g;a] \\ &= 2\int (g+b)d\nu \int a d\nu - 2\int g d\nu \int a d\nu \\ &= 2\int b d\nu \int a d\nu \end{aligned}$$

is another quadratic form, and $\delta^2\phi$ is independent of g.

Because the functions in $\mathcal{A}$ are positive, $\delta^2\phi$ is strongly positive on $\mathcal{A}$ (which again implies that ϕ is strictly convex):

$$\delta^2\phi[g;a,a] = 2\left(\int a d\nu\right)^2 = 2\|a\|^2_{L^1(\nu)} \geq 0$$

for $a \in \mathcal{A}$. The Bregman divergence is thus

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

Example: Relative Entropy of Simple Functions

Let (X, Σ, ν) be a measure space. Let $\mathcal{S}$ denote the collection of all measurable simple functions on (X, Σ, ν), that is, the set of functions which can be written as a finite linear combination of indicator or characteristic functions. If $g \in \mathcal{S}$ then it can be expressed as

$$g(x) = \sum_{i=0}^{t} \alpha_i \chi_{T_i}; \quad \alpha_0 = 0,$$

where χ_{T_i} is the indicator or characteristic function of the set T_i and $\{T_i\}_{i=0}^t$ is a collection of mutually disjoint measurable sets with the property that $X = \bigcup_{i=0}^t T_i$. One can adopt the convention that T_0 is the set on which g is zero and therefore $\alpha_i \neq 0$ if $i \neq 0$. The set $\left(\mathcal{S}, \|\cdot\|_{L^\infty(\nu)}\right)$ is a normed vector space. In this case

$$\int_X g \ln g d\nu = \sum_{i=1}^{t} \int_{T_i} \alpha_i \ln \alpha_i d\nu, \tag{7.2}$$

since $0 \ln 0 = 0$.

Note that the integral in (7.2) only exists for $g \in \mathcal{S}$ if $g \in L^1(\nu)$ and $g \geq 0$. This implies that $\nu(T_i) \leq \infty$ for all $1 \leq i \leq t$, while the measure of T_0 could be infinity. For this reason,

consider the normed vector space $(L^1(\nu)\cap\mathcal{S},\|\cdot\|_{L^\infty(\nu)})$, where $(L^1(\nu)\cap\mathcal{S})\subset\mathcal{S}\subset L^\infty(\nu)$. Let $\mathcal{W}$ be the set (not necessarily a vector space) of functions for which the integral $\int_X g\ln g\,d\nu$ is finite, that is, let

$$\mathcal{W}=\{g\in L^1(\nu)\cap\mathcal{S}\,\text{subject to}\,g\geq 0\}.$$

Define the functional ϕ on $\mathcal{W}$,

$$\phi[g]=\int_X g\ln g\,d\nu,\quad g\in\mathcal{W}. \tag{7.3}$$

The functional ϕ is not Fréchet differentiable because in general it cannot be guaranteed that $g+h$ is non-negative for all functions h in the underlying normed vector space $\left(L^1(\nu)\cap\mathcal{S},\|\cdot\|_{L^\infty(\nu)}\right)$ with norm smaller than any prescribed $\epsilon>0$. However, a generalized Gâteaux derivative can be defined if one limits the perturbing function h to a vector subspace.

Let $\mathcal{G}$ be the subspace of $\left(L^1(\nu)\cap\mathcal{S},\|\cdot\|_{L^\infty(\nu)}\right)$ defined by

$$\mathcal{G}=\{f\in L^1(\nu)\cap\mathcal{S}\,\text{subject to}\,f\,d\nu\ll g\,d\nu\}.$$

It is straightforward to show that $\mathcal{G}$ is vector space. The generalized Gâteaux derivative of ϕ at $g\in\mathcal{W}$ is defined to be the linear operator $\delta_G\phi[g;\cdot]$ if

$$\lim_{\substack{\|h\|_{L^\infty(\nu)}\to 0\\ h\in\mathcal{G}}}\frac{|\phi[g+h]-\phi[g]-\delta_G\phi[g;h]|}{\|h\|_{L^\infty(\nu)}}=0. \tag{7.4}$$

Note, that $\delta_G\phi[g;\cdot]$ is not linear in general, but it is on the vector space $\mathcal{G}$. (In general, if $\mathcal{G}$ is the entire underlying vector space then (7.4) is the Fréchet derivative, and if $\mathcal{G}$ is the span of only one element from the underlying vector space then (7.4) is the Gâteaux derivative. Here, the Gâteaux derivative has been generalized for the present case that $\mathcal{G}$ is a subspace of the underlying vector space).

It remains to be shown that given the functional (7.3), the derivative (7.4) exists and yields the relative entropy. Consider the solution

$$\delta_G\phi[g;h]=\int_X(1+\ln g)h d\nu, \tag{7.5}$$

which coupled with (7.3) does yield the relative entropy. The proof is completed by showing that (7.5) satisfies (7.4). Note that

$$\begin{aligned}\phi[g+h]-\phi[g]-\delta_G\phi[g;h] &= \int_X (h+g)\ln\frac{h+g}{g}-h d\nu \\ &= \int_E (h+g)\ln\frac{h+g}{g}-h d\nu, \end{aligned} \tag{7.6}$$

where E is the set on which g is not zero.

Since $g \in \mathcal{W}$, there are $m, M > 0$ such that $m \leq g \leq M$ on E. On one hand if $h \in \mathcal{G}$ is such that $\|h\|_{L^\infty(\nu)} \leq m$ then $g + h \geq 0$. In this case

$$(h+g)\ln\frac{h+g}{g}-h \leq (h+g)\frac{h}{g}-h = \frac{h^2}{g},$$

and therefore

$$\begin{aligned}\frac{\phi[g+h]-\phi[g]-\delta_G\phi[g;h]}{\|h\|_{L^\infty(\nu)}} &\leq \frac{1}{\|h\|_{L^\infty(\nu)}}\int_E \frac{h^2}{g}d\nu \\ &\leq \frac{1}{m}\int_E |h| d\nu \\ &\leq \frac{1}{m}\|h\|_{L^1(\nu)}.\end{aligned}$$

On the other hand

$$\int_E (h+g)\ln\frac{h+g}{g}d\nu = \int_E \frac{h+g}{g}\ln\frac{h+g}{g} g d\nu,$$

which can be written as

$$\|g\|_{L^1(\nu)}\int_E \frac{h+g}{g}\ln\frac{h+g}{g}\frac{g}{\|g\|_{L^1(\nu)}}d\nu = \|g\|_{L^1(\nu)}\int_E \lambda\left(\frac{h+g}{g}\right)d\tilde{\nu}.$$

Note, that the measure $d\tilde{\nu} = \frac{g}{\|g\|_{L^1(\nu)}}d\nu$ is a probability measure and $\lambda(x) = x\ln x$ is a convex function on $(0,\infty)$. Let's call $\|g\|_{L^1(\nu)} = M$. By Jensen's inequality

$$\begin{aligned}M\int_E \lambda\left(\frac{h+g}{g}\right)d\tilde{\nu} &\geq M\lambda\left(\int_E \frac{h+g}{g}d\tilde{\nu}\right) \\ &= M\lambda\left(\int_E \frac{h}{M}d\nu + \int_E d\tilde{\nu}\right) \\ &= M\lambda\left(\frac{1}{M}\int_E h\, d\nu + 1\right).\end{aligned}$$

Since

$$M\lambda\left(\frac{1}{M}\int_E h\,d\nu+1\right)=\left(\int_E h\,d\nu+M\right)\ln\left(\frac{1}{M}\int_E h\,d\nu+1\right).$$

As a result one can bound the integral in equation (7.6) from below by

$$\begin{aligned}\int_E (h+g)\ln\frac{h+g}{g}-h d\nu &\geq \left(\int_E h\,d\nu+M\right)\ln\left(\frac{1}{M}\int_E h\,d\nu+1\right)-\int_E h\,d\nu\\ &= \int_E h\,d\nu\ln\left(\frac{1}{M}\int_E h\,d\nu+1\right)\\ &\quad +M\ln\left(\frac{1}{M}\int_E h\,d\nu+1\right)-\int_E h\,d\nu.\end{aligned}$$

If $\int_E h\,d\nu=0$ then the integral in (7.6) is non-negative, so without loss of generality one can assume, that $\int_E h\,d\nu\neq 0$. In this case

$$\begin{aligned}&\frac{\phi[g+h]-\phi[g]-\delta_G\phi[g;h]}{\|h\|_{L^\infty(\nu)}}\\ &\geq \frac{\int_E h\,d\nu}{\|h\|_{L^\infty(\nu)}}\ln\left(\frac{1}{M}\int_E h\,d\nu+1\right)+\frac{M}{\|h\|_{L^\infty(\nu)}}\ln\left(\frac{1}{M}\int_E h\,d\nu+1\right)-\frac{\int_E h\,d\nu}{\|h\|_{L^\infty(\nu)}}\\ &\geq \frac{\int_E h\,d\nu}{\|h\|_{L^\infty(\nu)}}\ln\left(\frac{1}{M}\int_E h\,d\nu+1\right)+\left[\frac{M\ln\left(\frac{1}{M}\int_E h\,d\nu+1\right)}{\int_E h\,d\nu}-1\right]\frac{\int_E h\,d\nu}{\|h\|_{L^\infty(\nu)}}.\end{aligned}$$

Observe, that as $\int_E h\,d\nu\to 0$

$$\ln\left(\frac{1}{M}\int_E h\,d\nu+1\right)\to 0,$$

and

$$\frac{M\ln\left(\frac{1}{M}\int_E h\,d\nu+1\right)}{\int_E h\,d\nu}-1\to 0.$$

The proof ends by showing that there is a constant K which is independent of h such that

$$\left|\int_E h\,d\nu\right|\leq\|h\|_{L^1(\nu)}\leq K\|h\|_{L^\infty(\nu)}.$$

This implies that $\int_E h\,d\nu\to 0$ and $\|h\|_{L^1(\nu)}\to 0$ as $\|h\|_{L^\infty(\nu)}\to 0$ together with the fact that

$$\frac{\left|\int_E h\,d\nu\right|}{\|h\|_{L^\infty(\nu)}}\leq K,$$

which establishes (7.4). Because $h\in\mathcal{G}$, h can be expressed as

$$h=\sum_{i=0}^{v}\beta_i\chi_{V_i};\quad \beta_0=0,$$

where $\{V_i\}_{i=0}^{v}$ is a collection of mutually disjoint measurable sets with the property that $X = \bigcup_{i=0}^{v} V_i$. Also, because $h\,d\nu \ll g\,d\nu$ there is a set $N(h)$ such that $\nu(N(h)) = 0$ and

$$\bigcup_{i=1}^{v} V_i \subset \left(\bigcup_{i=1}^{t} T_i \cup N(h) \right).$$

This implies that there is a K independent of h such that

$$\sum_{i=1}^{v} \nu(V_i) \leq \sum_{i=1}^{t} \nu(T_i) = K.$$

Finally,

$$\begin{aligned} \int |h| d\nu &= \sum_{i=1}^{v} |\beta_i| \nu(V_i) \\ &\leq \|h\|_{L^\infty(\nu)} \sum_{i=1}^{v} \nu(V_i) \\ &\leq \|h\|_{L^\infty(\nu)} K. \end{aligned}$$

7.2.2 Relationship to Other Bregman Divergence Definitions

Two propositions establish the relationship of the functional Bregman divergence to other Bregman divergence definitions.

Proposition 7.2.2 (Functional Bregman Divergence Generalizes Vector Bregman Divergence)**.** *The functional definition (7.1) is a generalization of the standard vector Bregman divergence*

$$d_{\tilde{\phi}}(x, y) = \tilde{\phi}(x) - \tilde{\phi}(y) - \nabla\tilde{\phi}(y)^T(x - y), \tag{7.7}$$

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

The proof is given in Appendix A.

Jones and Byrne describe a general class of divergences between functions using a pointwise formulation [80]. Csiszár specialized the pointwise formulation to a class of divergences he termed *Bregman distances* $B_{s,\nu}$ [15], where given a σ-finite measure space (X, Ω, ν), and non-negative measurable functions $f(x)$ and $g(x)$, $B_{s,\nu}(f, g)$ equals

$$\int s(f(x)) - s(g(x)) - s'(g(x))(f(x) - g(x))d\nu(x). \tag{7.8}$$

The function $s : (0, \infty) \to \mathbb{R}$ is constrained to be differentiable and strictly convex, and the limit $\lim_{x\to 0} s(x)$ and $\lim_{x\to 0} s'(x)$ must exist, but not necessarily finite. The function s plays a role similar to the function ϕ in the functional Bregman divergence, but s acts on the range of the functions f, g, whereas ϕ acts on the pair of functions f, g.

Proposition 7.2.3 (Functional Definition Generalizes Pointwise Definition)**.** *Given a pointwise Bregman divergence as per (7.8), an equivalent functional Bregman divergence can be defined as per (7.1) if the measure ν is finite. However, given a functional Bregman divergence $d_\phi(f, g)$, there is not necessarily an equivalent pointwise Bregman divergence.*

The proof is given in the Appendix A.

7.3 Properties of the Functional Bregman Divergence

The Bregman divergence for vectors has some well-known properties, as reviewed in [24, Appendix A]. Here, it has been established that the same properties hold for the functional Bregman divergence (7.1).

1. Non-negativity

The functional Bregman divergence is non-negative. To show this, define $\tilde{\phi} : \mathbb{R} \to \mathbb{R}$ by $\tilde{\phi}(t) = \phi\left[tf + (1-t)g\right]$, $f, g \in \mathcal{A}$. From the definition of the Fréchet derivative,

$$\frac{d}{dt}\tilde{\phi} = \delta\phi[tf + (1-t)g; f - g]. \tag{7.9}$$

The function $\tilde{\phi}$ is convex because ϕ is convex by definition. Then from the mean value theorem there is some $0 \leq t_0 \leq 1$ such that

$$\tilde{\phi}(1) - \tilde{\phi}(0) = \frac{d}{dt}\tilde{\phi}(t_0) \geq \frac{d}{dt}\tilde{\phi}(0). \tag{7.10}$$

Because $\tilde{\phi}(1) = \phi[f]$, $\tilde{\phi}(0) = \phi[g]$ and (7.9), subtracting the right-hand side of (7.10) implies that

$$\phi[f] - \phi[g] - \delta\phi[g, f - g] \geq 0. \tag{7.11}$$

If $f = g$, then (7.11) holds in equality. Lastly, it is proved that equality only holds when $f = g$. Suppose (7.11) holds in equality; then

$$\tilde{\phi}(1) - \tilde{\phi}(0) = \frac{d}{dt}\tilde{\phi}(0). \tag{7.12}$$

The equation of the straight line connecting $\tilde{\phi}(0)$ to $\tilde{\phi}(1)$ is $\ell(t) = \tilde{\phi}(0) + (\tilde{\phi}(1) - \tilde{\phi}(0))t$, and the tangent line to the curve $\tilde{\phi}$ at $\tilde{\phi}(0)$ is $y(t) = \tilde{\phi}(0) + t\frac{d}{dt}\tilde{\phi}(0)$. Because $\tilde{\phi}(\tau) = \tilde{\phi}(0) + \int_0^\tau \frac{d}{dt}\tilde{\phi}(t)dt$ and $\frac{d}{dt}\tilde{\phi}(t) \geq \frac{d}{dt}\tilde{\phi}(0)$ as a direct consequence of convexity, it must be that $\tilde{\phi}(t) \geq y(t)$. Convexity also implies that $\ell(t) \geq \tilde{\phi}(t)$. However, the assumption that (7.11) holds in equality implied (7.12) which means $y(t) = \ell(t)$, and thus $\tilde{\phi}(t) = \ell(t)$, which is not strictly convex. Because ϕ is by definition strictly convex, it must be that $\phi[tf + (1-t)g] < t\phi[f] + (1-t)\phi[g]$ unless $f = g$. Thus, under the assumption of equality of (7.11), it must be that $f = g$.

2. Convexity

The Bregman divergence $d_\phi[f, g]$ is always convex with respect to f. Consider

$$\begin{aligned} \triangle d_\phi[f, g; a] &= d_\phi[f + a, g] - d_\phi[f, g] \\ &= \phi[f + a] - \phi[f] - \delta\phi[g; f - g + a] + \delta\phi[g; f - g]. \end{aligned}$$

Using linearity in the third term,

$$\begin{aligned} \triangle d_\phi[f, g; a] &= \phi[f + a] - \phi[f] - \delta\phi[g; f - g] - \delta\phi[g; a] + \delta\phi[g; f - g], \\ &= \phi[f + a] - \phi[f] - \delta\phi[g; a], \\ &\overset{(a)}{=} \delta\phi[f; a] + \frac{1}{2}\delta^2\phi[f; a, a] + \epsilon[f, a] \|a\|^2_{L(\nu)} - \delta\phi[g; a] \\ &\Rightarrow \delta^2 d_\phi[f, g; a, a] = \frac{1}{2}\delta^2\phi[f; a, a] > 0, \end{aligned}$$

where (a) and the conclusion follows from the appendix (B.2).

3. Linearity The functional Bregman divergence is linear in the sense that:

$$\begin{aligned} d_{(c_1\phi_1 + c_2\phi_2)}[f, g] &= (c_1\phi_1 + c_2\phi_2)[f] - (c_1\phi_1 + c_2\phi_2)[g] - \delta(c_1\phi_1 + c_2\phi_2)[g; f - g], \\ &= c_1 d_{\phi_1}[f, g] + c_2 d_{\phi_2}[f, g]. \end{aligned}$$

4. Equivalence Classes

Partition the set of strictly convex, differentiable functions $\{\phi\}$ on $\mathcal{A}$ into classes with respect to functional Bregman divergence, so that ϕ_1 and ϕ_2 belong to the same class if $d_{\phi_1}[f, g] = d_{\phi_2}[f, g]$ for all $f, g \in \mathcal{A}$. For brevity denote $d_{\phi_1}[f, g]$ simply by d_{ϕ_1}. Let $\phi_1 \sim \phi_2$ denote that ϕ_1 and ϕ_2 belong to the same class, then $\sim$ is an equivalence relation because it

satisfies the properties of *reflexivity* (because $d_{\phi_1} = d_{\phi_1}$), *symmetry* (because if $d_{\phi_1} = d_{\phi_2}$, then $d_{\phi_2} = d_{\phi_1}$) and *transitivity* (because if $d_{\phi_1} = d_{\phi_2}$ and $d_{\phi_2} = d_{\phi_3}$, then $d_{\phi_1} = d_{\phi_3}$).

Further, if $\phi_1 \sim \phi_2$, then they differ only by an affine transformation. To see this, note that by assumption, $\phi_1[f] - \phi_1[g] - \delta\phi_1[g; f-g] = \phi_2[f] - \phi_2[g] - \delta\phi_2[g; f-g]$, and fix g so $\phi_1[g]$ and $\phi_2[g]$ are constants. By the linearity property $\delta\phi[g; f-g] = \delta\phi[g; f] - \delta\phi[g; g]$, and because g is fixed, this equals $\delta\phi[g; f] + c_0$ where c_0 is a scalar constant. Then $\phi_2[f] = \phi_1[f] + (\delta\phi_2[g; f] - \delta\phi_1[g; f]) + c_1$, where c_1 is a constant. Thus,

$$\phi_2[f] = \phi_1[f] + Af + c_1,$$

where $A = \delta\phi_2[g; \cdot] - \delta\phi_1[g; \cdot]$ and thus $A : \mathcal{A} \to \mathbb{R}$ is a linear operator that does not depend on f.

5. Linear Separation

Fix two non-equal functions $g_1, g_2 \in \mathcal{A}$ and consider the set of all functions in $\mathcal{A}$ that are equidistant in terms of functional Bregman divergence from g_1 and g_2.

$$\begin{aligned} & d_\phi[f, g_1] = d_\phi[f, g_2] \\ \Rightarrow \quad & -\phi[g_1] - \delta\phi[g_1; f-g_1] = -\phi[g_2] - \delta\phi[g_2; f-g_2] \\ \Rightarrow \quad & -\delta\phi[g_1; f-g_1] = \phi[g_1] - \phi[g_2] - \delta\phi[g_2; f-g_2]. \end{aligned}$$

Using linearity the above relationship can be equivalently expressed

$$\begin{aligned} -\delta\phi[g_1; f] + \delta\phi[g_1; g_1] &= \phi[g_1] - \phi[g_2] - \delta\phi[g_2; f] + \delta\phi[g_2; g_2], \\ \delta\phi[g_2; f] - \delta\phi[g_1; f] &= \phi[g_1] - \phi[g_2] - \delta\phi[g_1; g_1] + \delta\phi[g_2; g_2]. \\ Lf &= c \end{aligned}$$

where L is the bounded linear functional defined by $Lf = \delta\phi[g_2; f] - \delta\phi[g_1; f]$ and c is the constant corresponding to the right hand side. In other words f has to be in the set $\{a \in \mathcal{A} : La = c\}$, where c is a constant. This set is a hyperplane.

6. Dual Divergence

Given a pair (g, ϕ) where $g \in L^p(\nu)$ and ϕ is a strictly convex twice continuously Fréchet differentiable functional, then the function-functional pair (G, ψ) is the Legendre transform

of (g, ϕ) [87], if

$$\phi[g] = -\psi[G] + \int g(x)G(x)d\nu(x), \tag{7.13}$$

$$\delta\phi[g; a] = \int G(x)a(x)d\nu(x), \tag{7.14}$$

where ψ is a strictly convex twice continuously Fréchet differentiable functional, and $G \in L^q(\nu)$, where $\frac{1}{p} + \frac{1}{q} = 1$.

Given Legendre transformation pairs $f, g \in L^p(\nu)$ and $F, G \in L^q(\nu)$,

$$d_\phi(f, g) = d_\psi(G, F).$$

The proof begins by substituting (7.13) and (7.14) into (7.1):

$$\begin{aligned} d_\phi[f, g] &= \phi[f] + \psi[G] - \int g(x)G(x)d\nu(x) - \int G(x)(f - g)(x)d\nu(x) \\ &= \phi[f] + \psi[G] - \int G(x)f(x)d\nu(x). \end{aligned} \tag{7.15}$$

Applying the Legendre transformation to (G, ψ) implies that

$$\psi[G] = -\phi[g] + \int g(x)G(x)d\nu(x) \tag{7.16}$$

$$\delta\psi[g; a] = \int g(x)a(x)d\nu(x). \tag{7.17}$$

Using (7.16) and (7.17), $d_\psi[G, F]$ can be reduced to (7.15).

7. Generalized Pythagorean Inequality

For any $f, g, h \in \mathcal{A}$,

$$d_\phi[f, h] = d_\phi[f, g] + d_\phi[g, h] + \delta\phi[g; f - g] - \delta\phi[h; f - g].$$

This can be derived as follows:

$$\begin{aligned} d_\phi[f, g] + d_\phi[g, h] &= \phi[f] - \phi[h] - \delta\phi[g; f - g] - \delta\phi[h; g - h] \\ &= \phi[f] - \phi[h] - \delta\phi[h; f - h] + \delta\phi[h; f - h] \\ &\quad -\delta\phi[g; f - g] - \delta\phi[h; g - h] \\ &= d_\phi[f, h] + \delta\phi[h; f - g] - \delta\phi[g; f - g], \end{aligned}$$

where the last line follows from the definition of the functional Bregman divergence and the linearity of the fourth and last terms.

7.4 Minimum Expected Bregman Divergence

Consider two sets of functions (or distributions), $\mathcal{M}$ and $\mathcal{A}$. Suppose there exists a probability distribution P_F over the set $\mathcal{M}$, such that $P_F(f)$ is the probability of $f \in \mathcal{M}$. For example, consider the set of Gaussian distributions, then given samples drawn independently and identically from a randomly selected Gaussian distribution N, the data implies a posterior probability $P_N(\mathcal{N})$ for each possible generating realization of a Gaussian distribution $\mathcal{N}$. The goal is to find the function $g^* \in \mathcal{A}$ that minimizes the expected Bregman divergence with respect to $P_F(f)$. The following theorem shows that if the set of possible minimizers $\mathcal{A}$ includes $E_{P_F}[F]$, then $g^* = E_{P_F}[F]$ minimizes the expectation of any Bregman divergence.

The theorem only applies to a set of functions $\mathcal{M}$ that lie on a finite-dimensional manifold M for which a differential element dM can be defined. For example, the set $\mathcal{M}$ could be parameterized by a finite number of parameters, or could be a set of functions that can be decomposed into a finite set of d basis functions $\{\psi_1, \psi_2, \ldots, \psi_d\}$ such that each f can be expressed:

$$f = \sum_{j=1}^{d} c_j \psi_j,$$

where $c_j \in \mathbb{R}$ for all j. The theorem requires slightly stronger conditions on ϕ than the definition of the Bregman divergence (7.1) requires.

Theorem 7.4.1 (Minimizer of the Expected Bregman Divergence)**.** *Let $\delta^2\phi[f; a, a]$ be strongly positive and let $\phi \in \mathcal{C}^3(L^1(\nu); \mathbb{R})$ be a three times continuously Fréchet differentiable functional on $L^1(\nu)$. Let $\mathcal{M}$ be a set of functions that lie on a finite-dimensional manifold M, and have associated differential element dM. Suppose there is a probability distribution P_F defined over the set $\mathcal{M}$. Suppose the function g^* minimizes the expected Bregman divergence between the set of functions $\mathcal{M}$ and any function $g \in \mathcal{A}$ with respect to probability distribution $P(f)$ such that*

$$g^* = \arg \inf_{g \in \mathcal{A}} E_{P_F}[d_\phi(F, g)].$$

Then, if it exists, g^* is given by

$$g^* = \int_M fP(f)dM = E_{P_F}[F]. \tag{7.18}$$

The proof is given in the Appendix A.

7.5 Bayesian Estimation

Theorem 7.4.1 can be applied to a set of distributions to find the Bayesian estimate of a distribution given a posterior or likelihood. For parametric distributions parameterized by $\theta \in \mathbb{R}^n$, a probability measure $\Lambda(\theta)$, and some risk function $R(\theta, \psi)$, $\psi \in \mathbb{R}^n$, the Bayes estimator is defined as [18]

$$\hat{\theta} = \arg \inf_{\psi \in \mathbb{R}^n} \int R(\theta, \psi) d\Lambda(\theta). \tag{7.19}$$

That is, the Bayes estimator minimizes some expected risk in terms of the parameters. It follows from recent results [14] that $\hat{\theta} = E[\Theta]$ if the risk R is a Bregman divergence, where Θ is the random variable whose realization is θ.

The principle of Bayesian estimation can be applied to the distributions themselves rather than to the parameters:

$$\hat{g} = \arg \inf_{g \in \mathcal{A}} \int_M R(f, g) P_F(f) dM, \tag{7.20}$$

where $P_F(f)$ is a probability measure on the distributions $f \in \mathcal{M}$, dM is a differential element for the finite-dimensional manifold M, and $\mathcal{A}$ is either the space of all distributions, or a subset of the space of all distributions, such as the set $\mathcal{M}$. When the set $\mathcal{A}$ includes the distribution $E_{P_F}[F]$ and the risk function R in (7.20) is a Bregman divergence, then Theorem 7.4.1 establishes that $\hat{g} = E_{P_F}[F]$.

For example, in recent work, two of the authors derived the mean class posterior distribution for each class for a Bayesian quadratic discriminant analysis classifier [39], and showed that the classification results were superior to parameter-based Bayesian quadratic discriminant analysis (using the same prior) for six simulations.

Of particular interest for estimation problems are the Bregman divergence examples given in Section 7.2.1: total squared difference (mean squared error) is a popular risk

function in regression [5]; minimizing relative entropy leads to useful theorems for large deviations and other statistical subfields [19]; and analyzing bias is a common approach to characterizing and understanding statistical learning algorithms [5].

7.5.1 Case Study: Estimating a Scaled Uniform Distribution

As an illustration, different estimators for estimating a scaled uniform distribution given independent and identically drawn samples are presented and compared. Let the set of uniform distributions over $[0, \theta]$ for $\theta \in \mathbb{R}^+$ be denoted $\mathcal{U}$. Given independent and identically distributed samples $X_1, X_2, \ldots, X_n$ drawn from an unknown uniform distribution $f \in \mathcal{U}$, the generating distribution is to be estimated. The risk function R is taken to be squared error in all cases.

Bayesian Parameter Estimate

Depending on the choice of the probability measure $\Lambda(\theta)$, the integral (7.19) may not be finite: For example, using the likelihood of θ with Lebesgue measure is not finite. A standard solution is to use a gamma prior on θ and Lebesgue measure. Let Θ be a random parameter with realization θ, let the gamma distribution have parameters t_1 and t_2, and denote the maximum of the data as $X_{\max} = \max\{X_1, X_2, \ldots, X_n\}$. Then a Bayesian estimate is formulated [18, p. 240, 285]

$$E[\Theta|\{X_1, X_2, \ldots, X_n\}, t_1, t_2] = \frac{\int_{X_{\max}}^{\infty} \theta \frac{1}{\theta^{n+t_1+1}} e^{\frac{-1}{\theta t_2}} d\theta}{\int_{X_{\max}}^{\infty} \frac{1}{\theta^{n+t_1+1}} e^{\frac{-1}{\theta t_2}} d\theta}. \tag{7.21}$$

The integrals can be expressed in terms of the chi-squared random variable χ^2_v with v degrees of freedom:

$$E[\Theta|\{X_1, X_2, \ldots, X_n\}, t_1, t_2] = \frac{1}{t_2(n + t_1 - a)} \frac{P(\chi^2_{2(n+t_1-1)} < \frac{2}{t_2 X_{\max}})}{P(\chi^2_{2(n+t_1)} < \frac{2}{t_2 X_{\max}})}. \tag{7.22}$$

Note that (7.19) presupposes that the best solution is also a uniform distribution.

Bayesian Uniform Distribution Estimate

If one restricts the minimizer of (7.20) to be a uniform distribution, then (7.20) is solved with $\mathcal{A} = \mathcal{U}$. Because the set of uniform distributions does not generally include its mean,

Theorem 7.4.1 does not apply, and thus different Bregman divergences may give different minimizers for (7.20). Let P_F be the likelihood of the data (no prior is assumed over the set $\mathcal{U}$), and use the Fisher information metric ([46, 47, 88]) for dM. Then the solution to (7.20) is the uniform distribution on $[0, 2^{1/n}X_{\max}]$. Using Lebesgue measure instead gives a similar result: $[0, 2^{1/(n+1/2)}X_{\max}]$. We were unable to find these estimates in the literature, and so their derivations are presented in Appendix A.

Unrestricted Bayesian Distribution Estimate

When the only restriction placed on the minimizer g in (7.20) is that g be a distribution, then one can apply Theorem 7.4.1 and solve directly for the expected distribution $E_{P_F}[F]$. Let P_F be the likelihood of the data (no prior is assumed over the set $\mathcal{U}$), and use the Fisher information metric for dM. Solving (7.18) given that the uniform probability of x is $f(x) = 1/a$ if $x \leq a$ and zero otherwise, and the likelihood of the n drawn points is $(1/X_{\max})^n$ if $a \geq X_{\max}$ and zero otherwise,

$$\begin{aligned} g^*(x) &= \frac{\int_{\max(x,X_{\max})}^{\infty} \left(\frac{1}{a}\right)\left(\frac{1}{a^n}\right)\left(\frac{da}{a}\right)}{\int_{X_{\max}}^{\infty} \frac{1}{a^n}\frac{da}{a}} \\ &= \frac{n\,(X_{\max})^n}{(n+1)[\max(x, X_{\max})]^{n+1}}. \end{aligned} \tag{7.23}$$

Projecting the Unrestricted Estimate onto the Set of Uniform Distributions

Consider what happens when the unrestricted solution $g^*(x)$ given in (7.23) is projected onto the set of uniform distributions with respect to squared error. That is, one solves for the uniform distribution $h(x)$ over $[0, a]$ such that:

$$\hat{a} = \arg \min_{a \in [0,\infty)} \int_0^{\infty} (h(x) - g^*(x))^2 dx. \tag{7.24}$$

The problem is straightforward to solve using standard calculus and yields the solution $\hat{a} = 2^{1/n}X_{\max}$. This is also the solution to the problem (7.20) when the minimizer is restricted to be a uniform distribution and the Fisher information metric over the uniform distributions is used (as discussed in Section 7.5.1). Thus, the projection of the unrestricted solution to (7.20) onto the set of uniform distributions is the same as the solution to (7.20) when

the minimizer is restricted to be uniform. One can conjecture that under some conditions this property will hold more generally: that the projection of the unrestricted minimizer of (7.20) onto the set $\mathcal{M}$ will be equivalent to solving (7.20) where the solution is restricted to the set $\mathcal{M}$.

7.5.2 Simulation

A simulation was done to compare the different Bayesian estimators and the maximum likelihood estimator. The simulation was run $1,000$ times: each time n data points were drawn independently and identically from the uniform over $[0, 1]$, and estimates were formed. Figure 7.1 is a log-log plot of the average squared errors between the estimated distribution and the true distribution.

For the Bayesian parameter estimator given in (7.22), estimates were calculated for three different sets of Gamma parameters, $(t_1 = 1, t_2 = 1)$, $(t_1 = 1, t_2 = 3)$, and $(t_1 = 1, t_2 = 100)$. The plotted error is the minimum of the three averaged errors for the different Gamma priors for each n. The plotted Bayesian distribution estimates used the Fisher information metric (very similar simulation results were obtained with the Lebesgue measure).

Given more than one random sample from the uniform, the unrestricted Bayesian distribution estimator (thick line) always performed better than the other estimators (as it should be by design). Of course, asymptotically as $n \to \infty$, all of the estimates will converge on the truth. For $n = 1$, the Bayesian parameter estimate performs better. One may believe this is due to the (in this case correct) bias of the prior used for the Bayesian parameter estimate. The dotted line rises upwards at $n = 155$ because the Bayesian parameter estimate was uncomputable for more than 155 data samples (we used Matlab v. 14 to evaluate (7.22), and for 155 data samples or more the numerator and denominator of (7.22) were determined to be 0, leading to an indeterminate estimate).

Three interesting conclusions are supported by the simulation results. First, the Bayesian estimates do improve significantly over the maximum likelihood estimate (dashed line). Second, although the truth is uniform, the unrestricted Bayesian distribution estimate chooses a non-uniform solution (thick line), which does significantly better than either of the Bayesian

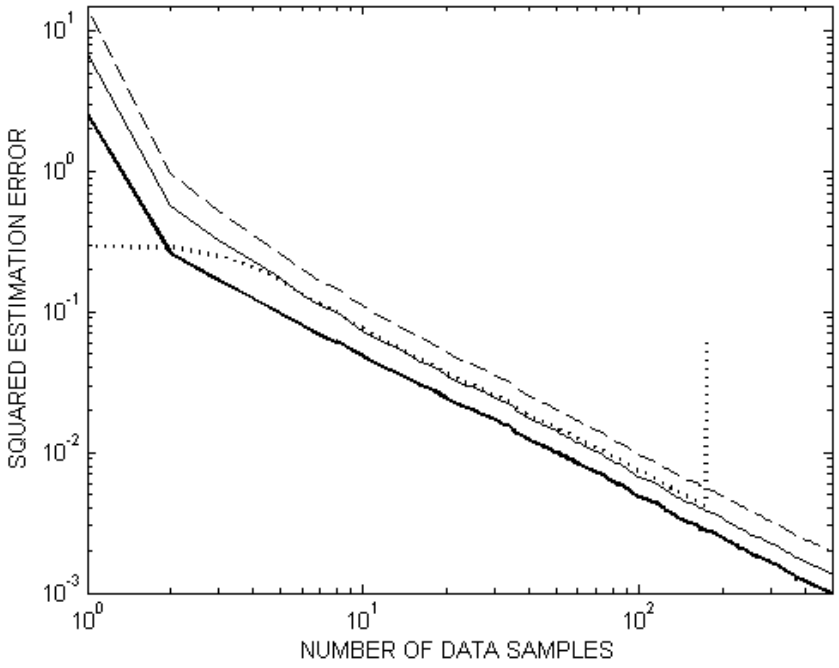

Figure 7.1: The plot shows the log of the squared error between an estimated distribution and a uniform $[0, 1]$ distribution, averaged over one thousand runs of the estimation simulation. The dashed line is the maximum likelihood estimate, the dotted line is the Bayesian parameter estimate, the thick solid line is the Bayesian distribution estimate that solves (7.20), and the thin solid line is the Bayesian distribution estimate that solves (7.20) but the minimizer is restricted to be uniform.

uniform estimates (thin line and dotted line). Third, the Bayesian parameter estimate (dotted line) and the Bayesian uniform distribution estimate (thin line) perform quite similarly. For $n < 10$ the Bayesian parameter estimate works better, but for $n > 10$, the Bayesian uniform distribution estimate is slightly better. Although these two estimates perform similarly, the Bayesian uniform distribution estimate $[0, 2^{1/n} X_{\max}]$ is a more elegant solution than the parameter estimate (7.22), and is easier to compute and to work with analytically.

7.6 Discussion

A general Bregman divergence for functions and distributions has been defined that can provide a foundation for results in statistics, information theory and signal processing. Theorem 7.4.1 is important for these fields because it ties Bregman divergences to expectation. As shown in Section 7.5, Theorem 7.4.1 can be directly applied to distributions to show that Bayesian distribution estimation simplifies to expectation when the risk function is a Bregman divergence and the minimizing distribution is unrestricted.

Bayesian distribution estimation was considered in Section 7.5 where the minimizer is restricted to a particular set of distributions, which may be particularly useful in circumstances where the expectation of a random distribution with respect to its likelihood (or the appropriate posterior) is not finite.

Acknowledgments

This work was funded in part by the Office of Naval Research, Code 321, Grant # N00014-05-1-0843. The authors thank Inderjit Dhillon, Imre Csiszár, and Galen Shorack for helpful discussions.

Chapter 8

CONCLUSION AND FEATURE WORKS

This chapter summarizes the main conclusion of the dissertation, discusses its impact and limitations, and suggest directions for future research.

8.1 Summary of Main Contributions

The first contribution of this dissertation is the definition of functional Bregman divergence which is a general Bregman divergence for functions and distributions, as was discussed in Chapter 7. We showed that functional Bregman divergence generalizes vector Bregman divergence and Jones-Byrne's and Csiszár's point-wise formulation of Bregman divergence as shown in Proposition 7.2.2 and Proposition 7.2.3. Examples are shown for total square difference, relative entropy, and square bias. We showed that square bias is a functional Bregman divergence, which was not previously seen in the literature despite the importance of minimizing bias being a common approach in estimation and statistical learning algorithms. In this example the functional ϕ can not be defined using the point-wise Bregman divergence definition. We showed that the functional Bregman divergence has many of the same properties as the standard vector Bregman divergence.

Furthermore in Section 7.5, Theorem 7.4.1, it was shown how the proposed functional definition allows us to extend Banerjee et al.'s [14] work to the continuous case. This theorem showed that when the minimizing distribution is unrestricted then the expectation of a set of functions minimizes the expected functional Bregman divergence. Theorem 7.4.1 can provide a foundation for results in statistics, information theory, signal processing, and statistical learning because it ties Bregman divergence to expectation. This theorem has direct application to the Bayesian estimation of distributions as opposed to the Bayesian estimation of parameters of distributions.

The algorithmic contribution comes from the derivation of regularized data adaptive

algorithms based on prior information and the Fisher information measure over statistical manifold of Gaussian distributions and applying them to pattern classification tasks. In Chapter 5 we showed that the proposed BDA7 classifier performs generally better than RDA, EDDA, and QB over twelve benchmark datasets and ten simulations when the number of dimensions is large compared to the number of training samples. Key aspects of the BDA7 classifier are that the seed matrix in the inverted Wishart prior defines the maximum of the prior, and that using a coarse estimate of the covariance matrix as the seed matrix pegs the prior at a relevant part of the distribution-space. We established an analytic link between Bayesian discriminant analysis and regularized discriminant analysis. In Chapter 4 we presented a functional equivalence between minimizing the expected misclassification costs and minimizing the expected Bregman divergence of class conditional distributions and also showed that distribution-based formulation of the Bayesian quadratic discriminant analysis classifier is related to the standard parameter-based formulation. We showed that the distribution-based Bayesian quadratic discriminant analysis classifier for a fixed data-dependent prior achieved the best result out of a number of non-cross-validated Gaussian model classifiers when there are relatively few data samples.

A third contribution of this dissertation is that we proposed local Bayesian quadratic discriminant analysis (local BDA) classifier as a simpler, closed-form alternative to Gaussian mixture models. In Chapter 6 we showed that on fourteen real datasets, local Bayesian QDA can achieve higher accuracy than the local nearest means classifier [65, 12], recently proposed local support vector machine classifier SVM-KNN [13], Gaussian mixture model, k-NN, and classifying by local linear regression. Furthermore, we showed that the local BDA decision boundary is locally quadratic over any region of the feature space where the neighborhood is constant.

Furthermore, we extended the discussion on estimation in Chapter 2, where we presented Bayesian estimation using Bregman divergence risk function and showed that the mean of the posterior pdf is the optimal Bayesian estimation. In Chapter 3 we showed that a generalized form of Laplace smoothing for weighted k nearest-neighbors class probability estimates can reduce the error rate by a large multiplicative factor when the misclassification costs are highly asymmetric.

8.2 Future Work

This section discusses a number of limitations of this dissertation, open questions, and suggests directions for future research.

There are some results for the standard vector Bregman divergence that have not been extended here. It has been shown that a standard vector Bregman divergence must be the risk function in order for the mean to be the minimizer of an expected risk [14, Theorem 3 and 4]. The proof of that result relies heavily on the discrete nature of the underlying vectors, and it remains an open question as to whether a similar result holds for the functional Bregman divergence. Another result shown for the vector case which is an open question in the functional case is convergence in probability [14, Theorem 2]. Also, it remains to be shown how to find a functional ϕ and derivative that yields the Itakura-Saito distance.

This dissertation and [14] showed that the mean minimizes the average Bregman divergence. Then one can ask oneself whether it can be shown that the standard deviation also minimizes some general loss functions or Bregman loss functions. This motivates the idea of how to define a Bregman divergence over the space of positive definite matrices or in general over the spaces of matrices. It is common in statistics to use quadratic loss and Stein loss for covariance estimation. Can we show that quadratic and Stein losses are also a Bregman divergence?

It is common in Bayesian estimation to interpret the prior as representing some actual prior knowledge, but in fact prior knowledge is often not available or difficult to quantify. Another viewpoint is that a prior can be used to capture coarse information from the data that may be used to stabilize the estimation [39, 81]. In practice, priors are sometimes chosen in Bayesian estimation to tame the tail of likelihood distributions so that expectations will exist when they might otherwise be infinite [18]. This mathematically convenient use of priors adds estimation bias which may be unwarranted by prior knowledge. An alternative to mathematically convenient priors is to formulate the estimation problem as a minimization of an expected Bregman divergence between the unknown distribution and the estimated distribution, and restrict the set of distributions that can be the minimizer to be a set for which there is a solution. Open questions are how to find or define a "best" restricted set of

distributions for this estimation approach, and how such restrictions affect the estimation bias and variance.

A practical algorithm BDA7 in Chapter 5 uses cross-validated data dependent priors. It cross-validates degrees of freedom $q \in \{d, 2d, \ldots 6d\}$ and seven models for seed matrix B_h of the inverted Wishart prior. Cross-validation has its obvious advantages, but one practical downside is that it slows down the computation by manyfold. For BDA7 classifier, there is slow down by a factor of 42. Maybe it is possible to make smarter choices for cross-validation parameters, especially for the degrees of freedom q. For example, are there patterns one can use to decide whether to try out higher values of q for a given choice of B_h? A recipe for finding the smart choices for q and B_h based on data dependent optimization criterion or semi-definite programming or high dimensional hypothesis testing will be beneficial to both researchers and practitioners. Modeling hyper prior over q and B_h could be another possible direction for future research.

In Chapter 5 BDA7 does as well or better than other classifiers when there were few data points compared to the number of dimension. But, when one has lot of data, BDA7 has too much model bias to be a general purpose classifier. It is well known that Gaussian mixture model classifiers work well for a variety of problems. It would be interesting an research direction as how to effectively integrate distribution-based Bayesian ideas into a mixture model classifier.

Finally, the problems addressed by this dissertation do not cover all aspects of estimation and minimizing risk in statistical learning. There are many other aspects that would be interesting to investigate further, e.g., regularization, shrinkage, and statistical challenges with high dimension [49, 3, 89, 90], regularization by gradient descent pathfinding [91], generalized discriminant analysis based on optimization criterion [51], sparse Bayesian learning and relevance vector machine [92], and Bayes point machine [93]. While the time scale of a Ph.D. is too short to explore all these problems, I am excited at the prospect of continuing research on estimation, statistical learning and their applications in the future.

Appendix A

PROOFS

A.1 Proof of Theorem 2.4.1

Proof: Applying theorem 2.1, the desired minimizer is

$$\theta^* = \arg\min_{\phi} E_{\Theta}[d_{\psi}(\Theta, \phi)] \equiv E_{\Theta}[\Theta], \tag{A.1}$$

for any Bregman divergence R. Thus,

$$\theta_g^* = \frac{\int_{\theta} \theta_g \prod_{g=1}^{G} \theta_g^{\alpha_g} d\theta}{\int_{\theta} \prod_{g=1}^{G} \theta_g^{\alpha_g} d\theta}. \tag{A.2}$$

Using Dirichlet's integral [94, pgs. 32-34] equation,

$$\begin{aligned}
\theta_g^* &= \frac{\left(\prod_{i=1}^{G-1} \Gamma(\alpha_i + 1)\right) \Gamma(\alpha_g + 2) \left(\prod_{j=g+1}^{G} \Gamma(\alpha_i + 1)\right)}{\Gamma(G + 1 + \sum_{j=1}^{G} \alpha_j)} \frac{\Gamma(G + \sum_{j=1}^{G} \alpha_j)}{\prod_{i=1}^{G} \Gamma(\alpha_i + 1)} \\
&= \frac{(\alpha_g + 1) \prod_{j=1}^{G} \Gamma(\alpha_g + 1)}{\prod_{j=1}^{G} \Gamma(\alpha_j + 1)} \frac{\Gamma(G + \sum_{j=1}^{G} \alpha_j)}{\Gamma(G + 1 + \sum_{j=1}^{G} \alpha_j)} \\
&= \frac{\alpha_g + 1}{\sum_{g=1}^{G} \alpha_g + G}
\end{aligned}$$

□

A.2 Proof of Theorem 4.4.1

Proof: The proof employs the following identities [57, 95],

$$\int_{\mu} \exp\left[-\frac{n_h}{2} \mathrm{tr}(\Sigma^{-1}(\mu - \bar{\mathrm{X}})(\mu - \bar{\mathrm{X}})^{\mathrm{T}})\right] d\mu = \left(\frac{2\pi}{n_h}\right)^{\frac{d}{2}} |\Sigma|^{\frac{1}{2}}, \tag{A.3}$$

$$\int_{\Sigma > 0} \frac{1}{|\Sigma|^{\frac{q}{2}}} \exp[-\mathrm{tr}(\Sigma^{-1}\mathrm{B})] \mathrm{d}\Sigma = \frac{\Gamma_d(\frac{q-d-1}{2})}{|B|^{\frac{q-d-1}{2}}}, \tag{A.4}$$

$$\tag{A.5}$$

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

$$\Gamma_d(a) = \left(\Gamma\left(\frac{1}{2}\right)\right)^{\frac{d(d-1)}{d}} \prod_{i=1}^{d} \Gamma\left(a - \frac{d-i}{2}\right). \tag{A.6}$$

To expand (4.15), first we simplify (4.17). The posterior (4.18) requires calculation of the normalization constant α_h,

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

Substitute $\ell(\mathcal{N}_h, \mathcal{T})$ from (4.19) and $p(\mathcal{N}_h)$ from (4.20),

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

Integrate with respect to μ_h using identity (A.3):

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

Next, integrate with respect to Σ_h using identity (A.4):

$$\alpha_h = \frac{1}{(2\pi)^{\frac{n_h d}{2}}}\left(\frac{2\pi}{n_h}\right)^{\frac{d}{2}} \frac{\Gamma_d(\frac{n_h+q}{2})}{\left|\frac{S_h+B_h}{2}\right|^{\frac{n_h+q}{2}}}. \tag{A.7}$$

Therefore the expectation $E_{N_h}[N_h]$ is

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

Integrate with respect to μ_h and Σ_h using identities (A.3) and (A.4), and equation (4.23) to yield

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

Substitute the value of α_h from (A.7),

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

Simplify the multivariate gamma function Γ_d using (A.6):

$$E_{N_h}[N_h(X)] = \frac{1}{(2\pi)^{\frac{d}{2}}} \frac{n_h^{\frac{d}{2}} \Gamma(\frac{n_h+q+1}{2})}{(n_h+1)^{\frac{d}{2}} \Gamma(\frac{n_h+q-d+1}{2})} \frac{|\frac{S_h+B_h}{2}|^{\frac{n_h+q}{2}}}{|A_h|^{\frac{n_h+q+1}{2}}}, \tag{A.8}$$

where A_h is given by equation (4.23). Substituting (A.8) into (4.15) proves the theorem.

□

A.3 Proof of Theorem 4.4.3

The posterior requires calculation of the normalization constant β_h,

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

Substitute $\ell(\mathcal{N}_h, \mathcal{T})$ from (4.19) and $p(\mathcal{N}_h)$ from (4.25),

$$\beta_h = \int_{\Sigma_h} \int_{\mu_h} \frac{\exp[-\frac{n_h}{2}\mathrm{tr}(\Sigma_{\mathrm{h}}^{-1}(\mu_{\mathrm{h}} - \bar{\mathrm{X}}_{\mathrm{h}})(\mu_{\mathrm{h}} - \bar{\mathrm{X}}_{\mathrm{h}})^{\mathrm{T}})] \exp[-\frac{1}{2}\,\mathrm{tr}(\Sigma_{\mathrm{h}}^{-1}\mathrm{S}_{\mathrm{h}})]}{(2\pi)^{\frac{n_h d}{2}} |\Sigma_h|^{\frac{n_h}{2}}} \left[\frac{1}{|\Sigma_h|^{\frac{d+1}{2}}}\right] \frac{d\Sigma_h d\mu_h}{|\Sigma_h|^{\frac{d+2}{2}}}.$$

Integrate with respect to μ_h using identity (A.3):

$$\beta_h = \frac{1}{(2\pi)^{\frac{n_h d}{2}}} \left(\frac{2\pi}{n_h}\right)^{\frac{d}{2}} \int_{\Sigma_h} \frac{\exp[-\frac{1}{2}\mathrm{tr}(\Sigma_{\mathrm{h}}^{-1}\mathrm{S}_{\mathrm{h}})]}{|\Sigma_h|^{\frac{n_h+2d+2}{2}}} d\Sigma_h.$$

Next, integrate with respect to Σ_h using identity (A.4):

$$\beta_h = \frac{1}{(2\pi)^{\frac{n_h d}{2}}} \left(\frac{2\pi}{n_h}\right)^{\frac{d}{2}} \frac{\Gamma_d(\frac{n_h+d+1}{2})}{\left|\frac{S_h}{2}\right|^{\frac{n_h+d+1}{2}}}. \tag{A.9}$$

Therefore the expectation $E_{N_h}[N_h]$ is

$$\begin{aligned} E_{N_h}[N_h(X)] &= \int_M \mathcal{N}_h(X) f(\mathcal{N}_h) dM_h \\ &= \frac{1}{\beta_h (2\pi)^{\frac{n_h d}{2}}} \int_{\Sigma_h} \int_{\mu_h} \frac{\exp\left[-\frac{1}{2}\mathrm{tr}\left(\Sigma_{\mathrm{h}}^{-1}(\mathrm{X} - \mu_{\mathrm{h}})(\mathrm{X} - \mu_{\mathrm{h}})^{\mathrm{T}}\right)\right]}{(2\pi)^{\frac{d}{2}} |\Sigma_h|^{\frac{1}{2}}} \\ &\quad \cdot \frac{\exp\left[-\frac{1}{2}\mathrm{tr}(\Sigma_{\mathrm{h}}^{-1}\mathrm{S}_{\mathrm{h}})\right]}{|\Sigma_h|^{\frac{n_h+d+1}{2}}} \exp\left[-\frac{n_h}{2}\mathrm{tr}\Sigma_{\mathrm{h}}^{-1}(\mu_{\mathrm{h}} - \bar{\mathrm{X}}_{\mathrm{h}})(\mu_{\mathrm{h}} - \bar{\mathrm{X}}_{\mathrm{h}})^{\mathrm{T}}\right] \frac{d\mu_h d\Sigma_h}{|\Sigma_h|^{\frac{d+2}{2}}}. \end{aligned}$$

Integrate with respect to μ_h and Σ_h using identities (A.3) and (A.4), and equation (4.27) to yield

$$E_{N_h}[N_h(X)] = \frac{1}{\beta_h (2\pi)^{\frac{dn_h}{2}} (n_h+1)^{\frac{d}{2}}} \frac{\Gamma_d(\frac{n_h+d+2}{2})}{|T_h|^{\frac{n_h+d+2}{2}}}.$$

Substitute the value of β_h from (A.9),

$$E_{N_h}[N_h(X)] = \frac{1}{(2\pi)^{\frac{d}{2}}} \frac{n_h^{\frac{d}{2}} \Gamma_d(\frac{n_h+d+2}{2}) |\frac{S_h}{2}|^{\frac{n_h+d+1}{2}}}{(n_h+1)^{\frac{d}{2}} \Gamma_d(\frac{n_h+d+1}{2}) |T_h|^{\frac{n_h+d+2}{2}}}.$$

Simplify the multivariate gamma function Γ_d using (A.6):

$$E_{N_h}[N_h(X)] = \frac{1}{(2\pi)^{\frac{d}{2}}} \frac{n_h^{\frac{d}{2}} \Gamma(\frac{n_h+d+2}{2})}{(n_h+1)^{\frac{d}{2}} \Gamma(\frac{n_h+d+1}{2})} \frac{|\frac{S_h}{2}|^{\frac{n_h+d+1}{2}}}{|T_h|^{\frac{n_h+d+2}{2}}}, \tag{A.10}$$

where T_h is given by equation (4.27). Substituting (A.10) into (4.15) proves the theorem.

□

A.4 Proof of Proposition 7.2.2

A constructive proof is given such that there is a corresponding functional Bregman divergence $d_\phi[f,g]$ for a specific choice of $\phi : \mathcal{A}^\infty \to \mathbb{R}$, where $\mathcal{A}^1$ is the function space $\mathcal{A}$ with $p = 1$, and where $\nu = \sum_{i=1}^n \delta_{c_i}$ and $f, g \in \mathcal{A}^\infty$. Here, δ_x is the Dirac measure (such that all mass is concentrated at x) and $\{c_1, c_2, \ldots, c_n\}$ is a collection of n distinct points in $\mathbb{R}^d$. For any $x \in \mathbb{R}^n$, define $\phi[f] = \tilde{\phi}(x_1, x_2, \ldots, x_n)$, where $f(c_1) = x_1, f(c_2) = x_2, \ldots, f(c_n) = x_n$. Then the difference is

$$\begin{aligned}
\Delta\phi[f;a] &= \phi[f+a] - \phi[f] \\
&= \tilde{\phi}\left((f+a)(c_1), \ldots, (f+a)(c_n)\right) - \tilde{\phi}\left(x_1, \ldots, x_n\right) \\
&= \tilde{\phi}\left(x_1 + a(c_1), \ldots, x_n + a(c_n)\right) - \tilde{\phi}\left(x_1, \ldots, x_n\right).
\end{aligned}$$

Let a_i be short-hand for $a(c_i)$, and use the Taylor expansion for functions of several variables to yield

$$\Delta\phi[f;a] = \nabla\tilde{\phi}(x_1, \ldots, x_n)^T (a_1, \ldots, a_n) + \epsilon[f,a]\|a\|_{L^1}.$$

Therefore,

$$\delta\phi[f;a] = \nabla\tilde{\phi}(x_1, \ldots, x_n)^T (a_1, \ldots, a_n) = \nabla\tilde{\phi}(x)^T a,$$

where $x = (x_1, x_2, \ldots, x_n)$ and $a = (a_1, \ldots, a_n)$. Thus from (3), the functional Bregman divergence definition (7.1) for ϕ is equivalent to the standard vector Bregman divergence:

$$\begin{aligned}
d_{\tilde{\phi}}[f,g] &= \phi[f] - \phi[g] - \delta\phi[g; f-g] \\
&= \tilde{\phi}(x) - \tilde{\phi}(y) - \nabla\tilde{\phi}(y)^T (x-y).
\end{aligned} \tag{A.11}$$

□

A.5 Proof of Proposition 7.2.3

A constructive proof of the first part of the proposition is given by showing that given a $B_{s,\nu}$, there is an equivalent functional divergence d_ϕ. The second part of the proposition is shown by example: the squared bias functional Bregman divergence given in Section 7.2.1 is an example of a functional Bregman divergence that cannot be defined as a pointwise Bregman divergence.

Note that the integral to calculate $B_{s,\nu}$ does not always turn out to be finite. To ensure finite $B_{s,\nu}$, explicitly constrain $\lim_{x\to 0} s'(x)$ and $\lim_{x\to 0} s(x)$ to be finite. From the assumption that s is strictly convex, s must be continuous on $(0,\infty)$. Recall from the assumptions that the measure ν is finite, and that the function s is differentiable on $(0,\infty)$.

Given a $B_{s,\nu}$, define the continuously differentiable function

$$\tilde{s}(x) = \begin{cases} s(x) & x \geq 0 \\ -s(-x) + 2s(0) & x < 0. \end{cases}$$

and specify $\phi : L^\infty(\nu) \to \mathbb{R}$ as

$$\phi[f] = \int_X \tilde{s}(f(x))d\nu.$$

Note that if $f \geq 0$,

$$\phi[f] = \int_X s(f(x))d\nu.$$

Because $\tilde{s}$ is continuous on $\mathbb{R}$, $\tilde{s}(f) \in L^\infty$ whenever $f \in L^\infty$, so the integrals always make sense.

It remains to show that $\delta\phi[f;\cdot]$ completes the equivalence when $f \geq 0$. For $h \in L^\infty$,

$$\begin{aligned} \phi[f+h] - \phi[f] &= \int_X \tilde{s}(f(x)+h(x))d\nu - \int_X s(f(x))d\nu \\ &= \int_X \tilde{s}(f(x)+h(x)) - s(f(x))d\nu \\ &= \int_X \tilde{s}'(f(x))h(x) + \epsilon\left(f(x),h(x)\right)h(x)d\nu \\ &= \int_X s'(f(x))h(x) + \epsilon\left(f(x),h(x)\right)h(x)d\nu, \end{aligned}$$

where

$$\begin{aligned}\tilde{s}(f(x)+h(x)) &= \tilde{s}(f(x)) + \tilde{s}'(f(x))h(x) \\ &\quad + \epsilon(f(x),h(x))h(x) \\ &= s(f(x)) + s'(f(x))h(x) \\ &\quad + \epsilon(f(x),h(x))h(x),\end{aligned}$$

because $f \geq 0$. On the other hand, if $h(x) = 0$ then $\epsilon(f(x), h(x)) = 0$ and if $h(x) \neq 0$ then

$$|\epsilon(f(x),h(x))| \leq |\frac{\tilde{s}(f(x)+h(x)) - \tilde{s}(f(x))}{h(x)}| + |s'(f(x))|.$$

Suppose $\{h_n\} \subset L^\infty(\nu)$ such that $h_n \to 0$. Then there is a measurable set E such that its complement is of measure 0 and $h_n \to 0$ uniformly on E. There is some $N > 0$ such that for any $n > N$, $|h_n(x)| \leq \epsilon$ for all $x \in E$. Without loss of generality assume that there is some $M > 0$ such that for all $x \in E$ $|f(x)| \leq M$. Since $\tilde{s}$ is continuously differentiable there is a $K > 0$ such that $\max\{\tilde{s}'(t) \,\text{subject to}\, t \in [-M-\epsilon, M+\epsilon]\} \leq K$ and by the mean value theorem

$$|\frac{\tilde{s}(f(x)+h(x)) - \tilde{s}(f(x))}{h(x)}| \leq K,$$

for almost all $x \in X$. Then

$$|\epsilon(f(x),h(x))| \leq 2K,$$

except on a set of measure 0. The fact that $h(x) \to 0$ almost everywhere implies that $|\epsilon(f(x), h(x))| \to 0$ almost everywhere, and by Lebesgue's dominated convergence theorem the corresponding integral goes to 0. As a result the Fréchet derivative of ϕ is

$$\delta\phi[f;h] = \int_X s'(f(x))h(x)d\nu. \tag{A.12}$$

Thus the functional Bregman divergence is equivalent to the given pointwise $B_{s,\nu}$.

Additionally it has been noted that the assumptions that $f \in L^\infty$ and that the measure ν is finite are necessary for this proof. Counterexamples can be constructed if $f \in L^p$ or $\nu(X) = \infty$ such that the Fréchet derivative of ϕ does not obey (A.12). This concludes the first part of the proof.

To show that the squared bias given in Section 7.2.1 is an example of a functional Bregman divergence that cannot be defined as a pointwise Bregman divergence one must prove that the converse statement leads to a contradiction.

Suppose (X, Σ, ν) and (X, Σ, μ) are measure spaces where ν is a non-zero σ-finite measure and that there is a differentiable function $f : (0, \infty) \to \mathbb{R}$ such that

$$\left(\int \xi d\nu \right)^2 = \int f(\xi) d\mu, \tag{A.13}$$

where $\xi \in \mathcal{A}^1$, the set of functions $\mathcal{A}$ with $p = 1$. Let $f(0) = \lim_{x \to 0} f(x)$, which can be finite or infinite, and let α be any real number. Then

$$\begin{aligned} \int f(\alpha\xi) d\mu = \left(\int \alpha\xi d\nu \right)^2 &= \alpha^2 \left(\int \xi d\nu \right)^2 \\ &= \alpha^2 \int f(\xi) d\mu. \end{aligned}$$

Because ν is σ-finite, there is a measurable set E such that $0 < |\nu(E)| < \infty$. Let $X \backslash E$ denote the complement of E in X. Then

$$\begin{aligned} \alpha^2 \nu^2(E) &= \alpha^2 \left(\int I_E d\nu \right)^2 \\ &= \alpha^2 \int f(I_E) d\mu \\ &= \alpha^2 \int_{X \backslash E} f(0) d\mu + \alpha^2 \int_E f(1) d\mu \\ &= \alpha^2 f(0) \mu(X \backslash E) + \alpha^2 f(1) \mu(E). \end{aligned}$$

Also,

$$\alpha^2 \nu^2(E) = \left(\int \alpha I_E d\nu \right)^2.$$

However,

$$\begin{aligned} \left(\int \alpha I_E d\nu \right)^2 &= \int f(\alpha I_E) d\mu \\ &= \int_{X \backslash E} f(\alpha I_E) d\mu + \int_E f(\alpha I_E) d\mu \\ &= f(0) \mu(X \backslash E) + f(\alpha) \mu(E); \end{aligned}$$

so one can conclude that

$$\alpha^2 f(0)\mu(X\backslash E) + \alpha^2 f(1)\mu(E) \quad = \quad f(0)\mu(X\backslash E) + f(\alpha)\mu(E). \tag{A.14}$$

Apply equation (A.13) for $\xi = 0$ to yield

$$0 = \left(\int 0 d\nu\right)^2 = \int f(0) d\mu = f(0)\mu(X).$$

Since $|\nu(E)| > 0$, $\mu(X) \neq 0$, so it must be that $f(0) = 0$, and (A.14) becomes

$$\alpha^2\nu^2(E) = \alpha^2 f(1)\mu(E) = f(\alpha)\mu(E) \quad \forall\alpha \in \mathbb{R}.$$

The first equation implies that $\mu(E) \neq 0$. The second equation determines the function f completely:

$$f(\alpha) = f(1)\alpha^2.$$

Then (A.13) becomes

$$\left(\int \xi d\nu\right)^2 = \int f(1)\xi^2 d\mu.$$

Consider any two disjoint measurable sets, E_1 and E_2, with finite nonzero measure. Define $\xi_1 = I_{E_1}$ and $\xi_2 = I_{E_2}$. Then $\xi = \xi_1 + \xi_2$ and $\xi_1\xi_2 = I_{E_1}I_{E_2} = 0$. Equation (A.13) becomes

$$\int \xi_1 d\nu \int \xi_2 d\nu = f(1) \int \xi_1\xi_2 d\mu. \tag{A.15}$$

This implies the following contradiction:

$$\int \xi_1 d\nu \int \xi_2 d\nu = \nu(E_1)\nu(E_2) \neq 0, \tag{A.16}$$

but

$$f(1) \int \xi_1\xi_2 d\mu = 0. \tag{A.17}$$

□

A.6 Proof of Theorem 7.4.1

Let

$$\begin{aligned} J[g] \quad &= \quad E_{P_F}[d_\phi(F, g)] = \int_M d_\phi[f, g]P(f)dM \\ &= \quad \int_M (\phi[f] - \phi[g] - \delta\phi[g; f - g])P(f)dM, \end{aligned} \tag{A.18}$$

where (A.18) follows by substituting the definition of Bregman divergence (7.1). Consider the increment

$$\Delta J[g;a] = J[g+a] - J[g] \tag{A.19}$$

$$= -\int_M \left(\phi[g+a] - \phi[g]\right) P(f)dM - \int_M \left(\delta\phi[g+a; f-g-a] \right.$$
$$\left. -\delta\phi[g; f-g]\right) P(f)dM, \tag{A.20}$$

where (A.20) follows from substituting (A.18) into (A.19). Using the definition of the differential of a functional (see Appendix A, (B.1)), the first integrand in (A.20) can be written as

$$\phi[g+a] - \phi[g] \quad = \quad \delta\phi[g;a] + \epsilon[g,a] \left\|a\right\|_{L^1(\nu)}. \tag{A.21}$$

Take the second integrand of (A.20), and subtract and add $\delta\phi[g; f-g-a]$,

$$\begin{aligned}
&\delta\phi[g+a; f-g-a] - \delta\phi[g; f-g] \\
&= \delta\phi[g+a; f-g-a] - \delta\phi[g; f-g-a] + \delta\phi[g; f-g-a] - \delta\phi[g; f-g] \\
&\overset{(a)}{=} \delta^2\phi[g; f-g-a, a] + \epsilon[g,a] \left\|a\right\|_{L^1(\nu)} + \delta\phi[g; f-g] - \delta\phi[g; a] - \delta\phi[g; f-g] \\
&\overset{(b)}{=} \delta^2\phi[g; f-g, a] - \delta^2\phi[g; a, a] + \epsilon[g,a] \left\|a\right\|_{L^1(\nu)} - \delta\phi[g; a],
\end{aligned} \tag{A.22}$$

where (a) follows from (B.3) and the linearity of the third term, and (b) follows from the linearity of the first term. Substitute (A.21) and (A.22) into (A.20),

$$\triangle J[g;a] \quad = \quad -\int_M \Big(\delta^2\phi[g; f-g, a] - \delta^2\phi[g; a, a] + \epsilon[g,a] \left\|a\right\|_{L^1(\nu)}\Big) P(f)dM.$$

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

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

where,

$$\delta J[g;a] \quad = \quad -\int_M \delta^2\phi[g; f-g, a] P(f)dM. \tag{A.23}$$

For fixed a, $\delta^2\phi[g;\cdot,a]$ is a bounded linear functional in the second argument, so the integration and the functional can be interchanged in (A.23), which becomes

$$\delta J[g;a] = -\delta^2\phi\left[g; \int_M (f-g)\,P(f)dM, a\right].$$

Using the functional optimality conditions (stated in Appendix B), $J[g]$ has an extremum for $g = \hat{g}$ if

$$\delta^2\phi\left[\hat{g}; \int_M (f-\hat{g})\,P(f)dM, a\right] = 0. \tag{A.24}$$

Set $a = \int_M (f-\hat{g})\,P(f)dM$ in (A.24) and use the assumption that the quadratic functional $\delta^2\phi[g;a,a]$ is strongly positive, which implies that the above functional can be zero if and only if $a = 0$, that is,

$$\begin{aligned} 0 &= \int_M (f-\hat{g})P(f)dM, &\text{(A.25)}\\ \hat{g} &= E_{P_f}[F], &\text{(A.26)} \end{aligned}$$

where the last line holds if the expectation exists (i.e. if the measure is well-defined and the expectation is finite). 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 see that (A.26) is in fact a minimum of $J[g]$, from the functional optimality conditions it is enough to show that $\delta^2 J[\hat{g};a,a]$ is strongly positive. To show this, for $b \in \mathcal{A}$, consider

$$\begin{aligned} &\delta J[g+b;a] - \delta J[g;a]\\ \overset{(c)}{=}\ &-\int_M \big(\delta^2\phi[g+b; f-g-b, a] - \delta^2\phi[g; f-g, a]\big)P(f)dM\\ \overset{(d)}{=}\ &-\int_M \big(\delta^2\phi[g+b; f-g-b, a] - \delta^2\phi[g; f-g-b, a] + \delta^2\phi[g; f-g-b, a]\\ &-\delta^2\phi[g; f-g, a]\big)P(f)dM\\ \overset{(e)}{=}\ &-\int_M \big(\delta^3\phi[g; f-g-b, a, b] + \epsilon[g,a,b]\,\|b\|_{L^1(\nu)} + \delta^2\phi[g; f-g, a] - \delta^2\phi[g; b, a]\\ &-\delta^2\phi[g; f-g, a]\big)P(f)dM\\ \overset{(f)}{=}\ &-\int_M \big(\delta^3\phi[g; f-g, a, b] - \delta^3\phi[g; b, a, b] + \epsilon[g,a,b]\,\|b\|_{L^1(\nu)}\\ &-\delta^2\phi[g; b, a]\big)P(f)dM, \end{aligned} \tag{A.27}$$

where (c) follows from using integral (A.23); (d) from subtracting and adding $\delta^2\phi[g; f-g-b, a]$; (e) from the fact that the variation of the second variation of ϕ is the third variation of

ϕ; and (f) from the linearity of the first term and cancellation of the third and fifth terms. Note that in (A.27) for fixed a, the term $\delta^3\phi[g; b, a, b]$ is of order $\|b\|^2_{L^1(\nu)}$, while the first and the last terms are of order $\|b\|_{L^1(\nu)}$. Therefore,

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

where

$$\delta^2 J[g;a,b] = -\int_M \delta^3\phi[g; f-g, a, b]P(f)dM + \int_M \delta^2\phi[g;a,b]P(f)dM. \quad \text{(A.28)}$$

Substitute $b = a$, $g = \hat{g}$ and interchange integration and the continuous functional $\delta^3\phi$ in the first integral of (A.28), then

$$\begin{aligned}
\delta^2 J[\hat{g};a,a] &= -\delta^3\phi\left[\hat{g}; \int_M (f-\hat{g})P(f)dM, a, a\right] + \int_M \delta^2\phi[\hat{g};a,a]P(f)dM \\
&= \int_M \delta^2\phi[\hat{g};a,a]P(f)dM \qquad \text{(A.29)} \\
&\geq \int_M k\,\|a\|^2_{L^1(\nu)}\,P(f)dM \\
&= k\,\|a\|^2_{L^1(\nu)} > 0, \qquad \text{(A.30)}
\end{aligned}$$

where (A.29) follows from (A.25), and (A.30) follows from the strong positivity of $\delta^2\phi[\hat{g};a,a]$. Therefore, from (A.30) and the functional optimality conditions, $\hat{g}$ is the minimum.

□

A.7 Derivation of Bayesian Distribution-based Uniform Estimate Restricted to a Uniform Minimizer

Let $f(x) = 1/a$ for all $0 \leq x \leq a$ and $g(x) = 1/b$ for all $0 \leq x \leq b$. Assume at first that $b > a$; then the total squared difference between f and g is

$$\begin{aligned}
\int_x (f(x)-g(x))^2 dx &= a\left(\frac{1}{a}-\frac{1}{b}\right)^2 + (b-a)\left(\frac{1}{b}\right)^2 \\
&= \frac{b-a}{ab} \\
&= \frac{|b-a|}{ab},
\end{aligned}$$

where the last line does not require the assumption that $b > a$.

In this case, the integral (7.20) is over the one-dimensional manifold of uniform distributions $\mathcal{U}$; a Riemannian metric can be formed by using the differential arc element to convert Lebesgue measure on the set $\mathcal{U}$ to a measure on the set of parameters a such that (7.20) is re-formulated in terms of the parameters for ease of calculation:

$$b^* = \arg\min_{b\in\mathbb{R}^+} \int_{a=X_{\max}}^{\infty} \frac{|b-a|}{ab} \frac{1}{a^n} \left\| \frac{df}{da} \right\|_2 da, \tag{A.31}$$

where a^n is the likelihood of the n data points being drawn from a uniform distribution $[0, a]$, and the estimated distribution is uniform on $[0, b^*]$. The differential arc element $\left\| \frac{df}{da} \right\|_2$ can be calculated by expanding df/da in terms of the Haar orthonormal basis $\{\frac{1}{\sqrt{a}}, \phi_{jk}(x)\}$, which forms a complete orthonormal basis for the interval $0 \le x \le a$, and then the required norm is equivalent to the norm of the basis coefficients of the orthonormal expansion:

$$\left\| \frac{df}{da} \right\|_2 = \frac{1}{a^{3/2}}. \tag{A.32}$$

For estimation problems, the measure determined by the Fisher information metric may be more appropriate than Lebesgue measure [46, 47, 88]. Then

$$dM = |I(a)|^{\frac{1}{2}} da, \tag{A.33}$$

where I is the Fisher information matrix. For the one-dimensional manifold M formed by the set of scaled uniform distributions $\mathcal{U}$, the Fisher information matrix is

$$\begin{aligned} I(a) &= E_X \left[\left(\frac{d \log \frac{1}{a}}{da} \right)^2 \right] \\ &= \int_0^a \frac{1}{a^2} \frac{1}{a} dx = \frac{1}{a^2}, \end{aligned}$$

so that the differential element is $dM = \frac{da}{a}$.

Solves (7.20) using the Lebesgue measure (A.32); the solution with the Fisher differential element follows the same logic. Then (A.31) is equivalent to

$$\begin{aligned} \arg\min_b J(b) &= \int_{a=X_{\max}}^{\infty} \frac{|b-a|}{ab} \frac{1}{a^{n+3/2}} da \\ &= \int_{a=X_{\max}}^{b} \frac{b-a}{ab} \frac{da}{a^{n+3/2}} + \int_b^{\infty} \frac{a-b}{ab} \frac{da}{a^{n+3/2}} \\ &= \frac{2}{(n+1/2)(n+3/2)b^{n+3/2}} - \frac{1}{b(n+\frac{1}{2})X_{\max}^{n+\frac{1}{2}}} + \frac{1}{(n+3/2)X_{\max}^{n+3/2}}. \end{aligned}$$

The minimum is found by setting the first derivative to zero,

$$\begin{aligned} J'(\hat{b}) &= \frac{2}{(n+1/2)(n+3/2)} \frac{(n+3/2)}{\hat{b}^{n+5/2}} + \frac{1}{\hat{b}^2(n+1/2)X_{\max}^{n+1/2}} = 0 \\ \Rightarrow \hat{b} &= 2^{\frac{1}{n+1/2}} X_{\max}. \end{aligned}$$

To establish that $\hat{b}$ is in fact a minimum, note that

$$J''(\hat{b}) = \frac{1}{\hat{b}X_{\max}^{n+1/2}} = \frac{1}{2^{\frac{3}{n+1/6}} X_{\max}^{n+7/2}} > 0.$$

Thus, the restricted Bayesian estimate is the uniform distribution over $[0, 2^{\frac{1}{n+1/2}} X_{\max}]$.

□

Appendix B

RELEVANT DEFINITIONS AND RESULTS FROM FUNCTIONAL ANALYSIS

For the aid of the reader, this appendix explains the basic definitions and results from functional analysis used in this paper. This material can be found in standard books on the calculus of variations, including the book by Gelfand and Fomin [87].

Let $\left(\mathbb{R}^d, \Omega, \nu\right)$ be a measure space, where ν is a Borel measure d is a positive integer, and define a set of functions $\mathcal{A} = \left\{a \in L^p(\nu) \,\text{subject to}\, \mathbb{R}^d \to \mathbb{R},\ a \geq 0\right\}$ where $1 \leq p \leq \infty$. The subset $\mathcal{A}$ is a convex subset of $L^p(\nu)$ because for $a_1,\ a_2 \in \mathcal{A}$ and $0 \leq \omega \leq 1$, $\omega a_1 + (1-\omega)a_2 \in \mathcal{A}$.

Definition of continuous linear functionals

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

1. $\psi[\omega a_1 + a_2] = \omega\psi[a_1] + \psi[a_2]$ for any $a_1, a_2 \in L^p(\nu)$ and any real number ω (linearity);

2. There is a constant C such that $|\psi[a]| \leq C\|a\|$ for all $a \in L^p(\nu)$.

Functional Derivatives

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

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

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

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

by

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

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

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

5. Using (B.2) we have for ϕ

$$\begin{aligned} \phi[f+a] &= \phi[f] + \delta\phi[f;a] + \frac{1}{2}\delta^2\phi[f;a,a] \\ &\quad + o(\|a\|^2), \\ \phi[f] &= \phi[f+a] - \delta\phi[f+a;a] + \\ &\quad \frac{1}{2}\delta^2\phi[f+a;a,a] + o(\|a\|^2), \end{aligned}$$

where $o(\|a\|^2)$ stands for a function which goes to zero as $\|a\|$ goes to zero even if it is divided by $\|a\|^2$. Adding the above two equations yields

$$\begin{aligned} 0 &= \delta\phi[f;a] - \delta\phi[f+a;a] + \frac{1}{2}\delta^2\phi[f;a,a] \\ &\quad + \frac{1}{2}\delta^2\phi[f+a;a,a] + o(\|a\|^2), \end{aligned}$$

which is equivalent to

$$\delta\phi[f+a;a]-\delta\phi[f;a]=\delta^2\phi[f;a,a]+o(\|a\|^2), \tag{B.3}$$

because

$$|\delta^2\phi[f+a;a,a]-\delta^2\phi[f;a,a]| \quad \leq \quad \|\delta^2\phi[f+a;\cdot,\cdot]-\delta^2\phi[f;\cdot,\cdot]\|\|a\|^2$$

and we assumed $\phi \in \mathcal{C}^2$, so $\delta^2\phi[f+a;a,a]-\delta^2\phi[f;a,a]$ is of order $o(\|a\|^2)$. This shows that the variation of the first variation of ϕ is the second variation of ϕ. A procedure like the above can be used to prove that analogous statements hold for higher variations if they exist.

Functional Optimality Conditions A necessary condition for a functional J to have an extremum (minimum) at $f=\hat{f}$ is that

$$\delta J[f;a]=0, \quad \text{and} \quad \delta^2 J[f;a,a]\geq 0,$$

for $f=\hat{f}$ and for all admissible functions $a \in \mathcal{A}$. 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 $f=\hat{f}$.

BIBLIOGRAPHY

[1] T. Mitchell, "The discipline of machine learning," *School of Computer Science, Carnegie Mellon University*, pp. 1–9, 2006.

[2] D. L. Donoho, "High-dimensional data analysis: the curse and blessing of dimensionality," *Aide-Memoire of the lecture in AMS conference: Maths challenges of 21st century*, 2000.

[3] J. Fan and R. Li, "Statistical challenges with high dimensionality: feature selection in knowledge discovery," *Proceedings of the International Congress of Mathematicians, Madrid*, vol. 3, pp. 595–622, 2006.

[4] J. H. Friedman, "Regularized discriminant analysis," *Journal of the American Statistical Association*, vol. 84, no. 405, pp. 165–175, 1989.

[5] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York: Springer, 2001.

[6] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern Classification, 2nd Edition*. New York: John Wiley and Sons, 2001.

[7] P. Vincent and Y. Bengio, "k-local hyperplane and convex distance nearest neighbor algorithms," *Advances in Neural Information Processing System*, pp. 985–992, 2001.

[8] J. C. Platt, N. Cristianini, and J. S. Taylor, "Large margin DAGs for multiclass classification," *Advances in Neural Information Processing System*, pp. 547–553, 2000.

[9] R. E. Fan, P. H. Chen, and C. J. Lin, "Working set selection using second order information for training support vector machines," *Journal of Machine Learning Research*, vol. 6, pp. 1889–1918, 2005.

[10] P. J. Brown, T. Fearn, and M. S. Haque, "Discrimination with many variables," *Journal of the American Statistical Association*, vol. 94, no. 448, pp. 1320–1329, 1999.

[11] H. Bensmail and G. Celeux, "Regularized Gaussian discriminant analysis through eigenvalue decomposition," *Journal of the American Statistical Association*, vol. 91, pp. 1743–1748, 1996.

[12] Y. Mitani and Y. Hamamoto, "A local mean-based nonparametric classifier," *Pattern Recognition Letter*, vol. 27, pp. 1151–1159, 2006.

[13] H. Zhang, A. C. Berg, M. Maire, and J. Malik, "Svm-knn: discriminative nearest neighbor classification for visual category recoginition," *Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition*, vol. 2, pp. 2126–2136, 2006.

[14] A. Banerjee, X. Guo, and H. Wang, "On the optimality of conditional expectation as a Bregman predictor," *IEEE Trans. on Information Theory*, vol. 51, no. 7, pp. 2664–2669, 2005.

[15] I. Csiszár, "Generalized projections for non-negative functions," *Acta Mathematica Hungarica*, vol. 68, pp. 161–185, 1995.

[16] M. R. Gupta, L. Cazzanti, and S. Srivastava, "Minimum expcted risk estimates for nonparametric neighborhood classifiers," *Proc. of the IEEE Workshop on Statistical Processing*, pp. 631–636 pages, 2005.

[17] S. M. Kay, *Fundamentals of Statistical Signal Processing: Estimation Theory.* New Jersey: Prentice Hall, 1993.

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

[19] T. Cover and J. Thomas, *Elements of Information Theory.* United States of America: John Wiley and Sons, 1991.

[20] M. Palus, "On entropy rates of dynamical systems and Gaussian processes," *Physics Letter A*, vol. 227, pp. 301–308, 1997.

[21] I. Csiszár, "Why least squares and maximum entropy? an axiomatic approah to inference for linear inverse problems," *The Annals of Statistics*, vol. 19, no. 4, pp. 2032–2066, 1991.

[22] D. Kazakos and P. P. Kazakos, "Spectral distance measures between Gaussian processes," *IEEE Transaction on Automatic Control*, vol. 25, no. 5, pp. 950–959, 1980.

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

[24] A. Banerjee, S. Merugu, I. S. Dhillon, and J. Ghosh, "Clustering with Bregman divergences," *Journal of Machine Learning Research*, vol. 6, pp. 1705–1749, 2005.

[25] M. R. Gupta, S. Srivastava, and L. Cazzanti, "Optimal estimation for nearest neighbor classifiers," *In review by the IEEE Trans. of Information Theory*, pp. 1–4 pages, 2006.

[26] C. J. Stone, "Consistent nonparametric regression," *The Annals of Statistics*, vol. 5, no. 4, pp. 595–645, 1977.

[27] D. J. C. MacKay, *Information Theory, Inference, and Learning Algorithms*. Cambridge: Cambridge University Press, 2003.

[28] L. Devroye, L. Gyorfi, and G. Lugosi, *A Probabilistic Theory of Pattern Recognition*. New York: Springer-Verlag, 1996.

[29] R. Rosenfeld, "Two decades of statistical language modeling: Where do we go from here?" *Proceedings of the IEEE*, vol. 88, no. 8, 2000.

[30] J. M. Bernardo and A. F. M. Smith, *Bayesian Theory*. Chichester: Wiley Series in Probability and Statistics, 2000.

[31] E. T. Jaynes and G. T. Bretthorst, *Probability Theory: The Logic of Science*. Cambridge: Cambridge University Press, 2003.

[32] F. Provost and P. Domingos, "Tree induction for probability-based rankings," *Machine Learning*, vol. 52, no. 3, pp. 199–216, 2003.

[33] E. Bauer and R. Kohavi, "An empirical comparison of voting classification algorithms: Bagging, boosting, and variants," *Machine Learning*, vol. 36, pp. 105–139, 1999.

[34] J. P. Bradford, C. Kunz, R. Kohavi, C. Brunk, and C. E. Brodley, "Pruning decision trees with misclassification costs," *Proceedings of the Tenth European Conference on Machine Learning*, pp. 131–136, 1998.

[35] T. Niblett, "Constructing decision trees in noisy domains," *Proceedings of the Second European Working Session on Learning*, pp. 67–78, 1987.

[36] B. Cestnik, "Estimating probabilities: A crucial task in machine learning," *Proceedings of the European Conference on Artificial Intelligence*, pp. 147–149, 1990.

[37] P. Clark and R. Boswell, "Rule induction with CN2: Some recent improvements," *Proceedings of the Fifth European Working Session on Learning*, pp. 151–163, 1991.

[38] P. Domingos, "Unifying instance-based and rule-based induction," *Machine Learning*, vol. 24, pp. 141–168, 1996.

[39] S. Srivastava and M. R. Gupta, "Distribution-based Bayesian minimum expected risk for discriminant analysis," *Proc. of the IEEE Intl. Symposium on Information Theory*, 2006.

[40] S. Srivastava, M. R. Gupta, and B. A. Frigyik, "Bayesian quadratic discriminant analysis," *To appear in the Journal of Machine Learning Research*, pp. 1–30 pages, 2007.

[41] S. Geisser, "Posterior odds for multivariate normal distributions," *Journal of the Royal Society Series B Methodological*, vol. 26, pp. 69–76, 1964.

[42] B. Ripley, *Pattern recognition and neural nets*. Cambridge: Cambridge University Press, 2001.

[43] S. Geisser, *Predictive Inference: An Introduction*. New York: Chapman and Hall, 1993.

[44] D. G. Keehn, "A note on learning for Gaussian properties," *IEEE Trans. on Information Theory*, vol. 11, pp. 126–132, 1965.

[45] 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*, vol. 13, no. 3, pp. 252–264, 1991.

[46] R. E. Kass, "The geometry of asymptotic inference," *Statistical Science*, vol. 4, no. 3, pp. 188–234, 1989.

[47] S. Amari and H. Nagaoka, *Methods of Information Geometry*. New York: Oxford University Press, 2000.

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

[49] P. J. Bickel and B. Li, "Regularization in statistics," *Test*, vol. 15, no. 2, pp. 271–344, 2006.

[50] D. L. Swets and J. Weng, "Using discriminant eigenfeatures for image retrieval," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 18, no. 8, pp. 831–836, 1996.

[51] J. Ye, "Characterization of a family of algorithms for generalized discriminant analysis of undersample problems," *Journal of Machine Learning Research*, vol. 6, pp. 483–502, 2005.

[52] P. S. Dwyer, "Some applications of matrix derivatives in multivariate analysis," *Journal of the American Statistical Association*, vol. 333, pp. 607–625, 1967.

[53] B. Effron and C. Morris, "Multivariate empirical Bayes and estimation of covariance matrices," *The Annals of Statistics*, vol. 4, pp. 22–32, 1976.

[54] L. R. Haff, "Empirical Bayes estimation of the multivariate normal covariance matrix," *The Annals of Statistics*, vol. 8, no. 3, pp. 586–597, 1980.

[55] L. Wasserman, "Asymptotic inference of mixture models using data-dependent prior," *Journal of the Royal Statistical Society Series B*, vol. 62, no. 1, pp. 159–180, 2000.

[56] T. W. Anderson, *An Introduction to Multivariate Statistical Analysis.* Hoboken, NJ: Wiley-Interscience, 2003.

[57] G. E. P. Box and G. C. Tiao, *Bayesian inference in statistical analysis.* Reading, Massachusetts: Addison-Wesley, 1973.

[58] M. R. Gupta and S. Srivastava, "Local Bayesian quadratic discriminant analysis," *In review by Neural Information Processing Systems Conference*, pp. 1–8 pages, 2007.

[59] W. Lam, C. Keung, and D. Liu, "Discovering useful concept prototypes for classification based on filtering and abstraction," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, pp. 1075–1090, 2002.

[60] M. R. Gupta, R. M. Gray, and R. A. Olshen, "Nonparametric supervised learning with linear interpolation and maximum entropy," *IEEE Trans. on Pattern Analysis and Machine Learning*, vol. 28, pp. 766–781, 2006.

[61] L. Bottou and V. Vapnik, "Local learning algorithms," *Neural Computation*, vol. 4, pp. 888–900, 1992.

[62] R. Goldstone and A. Kersten, *Comprehensive Handbook of Psychology.* New Jersey: Wiley, 2003.

[63] S. Thorpe, D. Fize, and C. Marlot, "Speed of processing in the human visual system," *Nature*, vol. 381, pp. 520–522, 1996.

[64] D. Aha, *Lazy Learning.* New York: Springer, 1997.

[65] Y. Mitani and Y. Hamamoto, "Classifier design based on the use of nearest neighbor samples," *Proc. of the Intl. Conf. on Pattern Recognition*, pp. 769–772, 2000.

[66] L. Cazzanti and M. R. Gupta, "Local similarity discriminant analysis," *Proc. of the Intl. Conf. on Machine Learning*, 2007.

[67] J. Huang, A. Lin, B. Narasimhan, T. Quertermous, C. A. Hsiung, L. T. Ho, J. S. Grove, M. Oliver, K. Ranade, N. J. Risch, and R. A. Olshen, "Tree-structure supervised learning and the genetics of hypertension," *Proc. National Academy of Sciences*, vol. 101, pp. 10 529–10 534, 2004.

[68] T. Hastie and R. Tibshirani, "Discriminative adaptive nearest neighbour classification," *IEEE Trans. on Pattern Analysis and Machine Learning*, vol. 18, no. 6, pp. 607–615, 1996.

[69] C. Domeniconi, J. Peng, and D. Gunopulos, "Locally adaptive metric nearest neighbor classification," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 9, pp. 1281–1285, 2002.

[70] R. Paredes and E. Vidal, "learning weighted metrics to minimize nearest-neighbor classification error," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 1100–1110, 2006.

[71] Y. Hamamoto, S. Uchimura, and S. Tomita, "A bootstrap technique for nearest neighbor classifier design," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 73–79, 1997.

[72] C. C. Chang and C. J. Lin, "Libsvm," *A library for support vector machines*, 2001.

[73] P. Hall, J. S. Marron, and A. Neeman, "Geometric representation of high dimensional, low sample size data," *Journal Royal Statistical Society B*, vol. 67, pp. 427–444, 2005.

[74] B. A. Frigyik, S. Srivastava, and M. R. Gupta, "Functional Bregman divergence and bayesian estimation of distributions," *In review by the IEEE Trans. of Information Theory*, pp. 1–13 pages, 2006.

[75] B. Taskar, S. Lacoste-Julien, and M. I. Jordan, "Structured prediction, dual extragradient and Bregman projections," *Journal of Machine Learning Research*, vol. 7, pp. 1627–1653, 2006.

[76] N. Murata, T. Takenouchi, T. Kanamori, and S. Eguchi, "Information geometry of U-Boost and Bregman divergence," *Neural Computation*, vol. 16, pp. 1437–1481, 2004.

[77] M. Collins, R. E. Schapire, and Y. Singer, "Logistic regression, AdaBoost and Bregman distances," *Machine Learning*, vol. 48, pp. 253–285, 2002.

[78] J. Kivinen and M. Warmuth, "Relative loss bounds for multidimensional regression problems," *Machine Learning*, vol. 45, no. 3, pp. 301–329, 2001.

[79] J. Lafferty, "Additive models, boosting, and inference for generalized divergences," *Proc. of Conf. on Learning Theory (COLT)*, 1999.

[80] L. K. Jones and C. L. Byrne, "General entropy criteria for inverse problems, with applications to data compression, pattern classification, and cluster analysis," *IEEE Trans. on Information Theory*, vol. 36, pp. 23–30, 1990.

[81] S. Srivastava, M. R. Gupta, and B. A. Frigyik, "Bayesian discriminannt analysis," *In review by the Journal of Machine Learning Research*, 2006.

[82] R. Nock and F. Nielsen, "On weighting clustering," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 28, no. 8, pp. 1223–1235, 2006.

[83] G. LeBesenerais, J. Bercher, and G. Demoment, "A new look at entropy for solving linear inverse problems," *IEEE Trans. on Information Theory*, vol. 45, pp. 1565–1577, 1999.

[84] Y. Altun and A. Smola, "Unifying divergence minimization and statistical inference via convex duality," *Proc. of Conf. on Learning Theory (COLT)*, 2006.

[85] M. C. Pardo and I. Vajda, "About distances of discrete distributions satisfying the data processing theorem of information theory," *IEEE Trans. on Information Theory*, vol. 43, no. 4, pp. 1288–1293, 1997.

[86] T. Rockafellar, "Integrals which are convex functionals," *Pacific Journal of Mathematics*, vol. 24, no. 3, pp. 525–539, 1968.

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

[88] G. Lebanon, "Axiomatic geometry of conditional models," *IEEE Transaction. on Information Theory*, vol. 51, no. 4, pp. 1283–1294, 2005.

[89] O. Ledoit and M. Wolf, "Some hypothesis tests for covariance matrix when the dimension is large compared to the sample size," *The Annals of Statistics*, vol. 30, no. 4, pp. 1081–1102, 2002.

[90] M. J. Daniels and R. E. Kass, "Shrinkage estimators for covariance matrices," *Biometrics*, vol. 57, pp. 1173–1184, 2001.

[91] J. H. Friedman and B. E. Popescu, "Gradient directed regularization," *Technical Report, Stanford*, pp. 1–30, 2004.

[92] M. E. Tippings, "Sparse Bayesian learning and the relevance vector machine," *Journal of Machine Learning Research*, vol. 1, pp. 211–244, 2001.

[93] R. Herbrich, T. Graepel, and C. Campbell, "Bayes point machine," *Journal of Machine Learning Research*, vol. 1, pp. 245–279, 2001.

[94] G. E. Andrews, R. Askey, and R. Roy, *Special Functions*. New York: Cambridge University Press, 2000.

[95] A. Gupta and D. Nagar, *Matrix Variate Distributions*. Florida: Chapman and Hall/CRC, 2000.

VITA

Santosh Srivastava

Education

- University of Washington — Seattle, WA
  Ph.D., Department of Applied Mathematics, 2007

- University of Washington — Seattle, WA
  M.S., Department of Applied Mathematics, 2004

- Indian Institute of Technology, Bombay — Mumbai, India
  M.Sc., Department of Mathematics, 1998

- Magadh University — Patna, India
  B.Sc., Department of Physics, 1994

Academic Position

- Research Assistant, Information Design Lab — 2004-2007
  Department of Electrical Engineering, University of Washington

- Project Engineer, Electrical Engineering Department — 1999-2000
  Indian Institute of Technology, Bombay

- Research Assistant, Department of Mechanical Engineering — 1998-1999
  Indian Institute of Technology, Bombay

Honors

- MCM Full Scholarship — 1995-1997
  Indian Institute of Technology, Bombay

- Boeing Scholarship 2000
  University of Washington

Publications

- Functional Bregman Divergence and Bayesian Estimation of Distributions, B. A. Frigyik, S. Srivastava, M. R. Gupta. In review by IEEE Trans. On Information Theory (13 pages).

- Local Bayesian Quadratic Discriminant Analysis, M. R. Gupta, S. Srivastava, In review by Neural Information Processing Systems Conference, (8 pages), 2007.

- Bayesian Quadratic Discriminant Analysis, S. Srivastava, M. R. Gupta and B. A. Frigyik. Journal of Machine Learning Research, 8(Jun):1277–1305, 2007.

- Optimal Estimation for Nearest-Neighbor Classifiers, M. R. Gupta, S. Srivastava, L. Cazzanti. In review by IEEE Trans. On Information Theory (4 pages).

- Distribution-based Bayesian Minimum Expected Risk for Discriminant Analysis, S. Srivastava and M. R. Gupta. Proc of the IEEE Intl. Symposium on Information Theory, Seattle, 2006 (6 pages).

- Minimum Expected Risk Probability Estimates for Nonparametric Neighborhood Classifiers, M. R. Gupta and L. Cazzanti and S. Srivastava. Proc. of the IEEE Workshop on Statistical Signal Processing, pages 631-636, 2005.

Project And Work Experience

- My Ph.D. thesis research is on Bayesian minimum expected risk estimation of distributions for statistical learning, including near-neighbor learning, and discriminant analysis, using statistical manifold tools and information geometry. Advisor: Prof. Maya Gupta, Electrical Engineering Dept.

- Master Degree Presentation on "Einstein-Podolsky-Rosen (EPR) Pairs, Teleportation, and Quantum Computation" Feb 2004. My committee members were Prof. Mark Kot, Prof. Robert O. Malley from Applied Mathematics and Prof. Maya Gupta from Electrical Engineering Dept.

- Project Engineer on "Under-Actuated Mechanical Systems" project funded by Indian Space Research Organization (ISRO) in Systems and Controls Group, Electrical Engineering Department, Indian Institute of Technology, Bombay.

- Research Assistant under the supervision of Prof. Sandipan Ghosh Moulic, Mechanical Engineering Department, Indian Institute of Technology, Bombay for a project sponsored by Indian Space Research Organization (ISRO) "Finding Numerical Solution of Axis-symmetric Sloshing Motion in Low Gravity Condition".

- Master of Science project "Riemann Solution of Nonlinear Hyperbolic System of Conservation Laws" under the supervision of Prof. V.D. Sharma, Head of the Mathematics Department, Indian Institute of Technology, Bombay.

- Reviewer for the Proc. of the IEEE Intl. Symposium on Information Theory, 2006.

- Reviewer for JMLR (Journal of Machine Learning Research).

- Reviewer for IEEE Trans. on Pattern Analysis and Machine Intelligence.

Invited Talks

- Bayesian quadratic discriminant analysis, Information Theory and Applications Workshop, UCSD, 2007.