UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Bayesian kernel density estimation

by

Estian Rademeyer

Submitted in partial fulfilment of the requirements for the degree
Masters of Science Mathematical Statistics

In the Faculty of Natural & Agricultural Sciences
University of Pretoria

Pretoria
November, 2017

I, Estian Rademeyer declare that the thesis/ dissertation, which I hereby submit for the degree Masters of Science Mathematical Statistics at the University of Pretoria, is my own work and has not previously been submitted by me for a degree at this or any other tertiary institution.

Signed: _____
Estian Rademeyer

Date: _____

ABSTRACT

This dissertation investigates the performance of two-class classification credit scoring data sets with low default ratios. The standard two-class parametric Gaussian and naive Bayes (NB), as well as the non-parametric Parzen classifiers are extended, using Bayes' rule, to include either a class imbalance or a Bernoulli prior. This is done with the aim of addressing the low default probability problem. Furthermore, the performance of Parzen classification with Silverman and Minimum Leave-one-out Entropy (MLE) Gaussian kernel bandwidth estimation is also investigated. It is shown that the non-parametric Parzen classifiers yield superior classification power.

However, there is a longing for these non-parametric classifiers to posses a predictive power, such as exhibited by the odds ratio found in logistic regression (LR). The dissertation therefore dedicates a section to, amongst other things, study the paper entitled "Model-Free Objective Bayesian Prediction" (Bernardo 1999). Since this approach to Bayesian kernel density estimation is only developed for the univariate and the uncorrelated multivariate case, the section develops a theoretical multivariate approach to Bayesian kernel density estimation. This approach is theoretically capable of handling both correlated as well as uncorrelated features in data. This is done through the assumption of a multivariate Gaussian kernel function and the use of an inverse Wishart prior.

ACKNOWLEDGMENTS

The financial assistance of the National Research Foundation (NRF) towards this research is hereby acknowledged. Opinions expressed and conclusions arrived at, are those of the authors and are not necessarily to be attributed to the NRF.

Contents

List of Tables

List of Figures

NOMENCLATURE

Abbreviations

$AMISE$	Asymptotic mean integrated squared error
BLR	Bayesian logistic regression
$HSJM$	Hall Sheather Jones Marron
ISE	Integrated squared error
$LOUT$	Leave-One-Out
LR	Logistic regression
$MISE$	Mean integrated squared error
MLE	Minimum Leave-One-Out Entropy or Maximum likelihood estimator
MLL	Maximum Leave-One-Out Likelihood
MSE	Mean squared error
MVN	Multivariate normal distribution
NB	Naive Bayes
PCA	Principal component analysis
$PCA95\%$	Data set that results from applying PCA and maintaining the principal components that result in 95% of the variation being maintained.
W^{-1}	Inverse Wishart distribution
Ig	Inverse gamma distribution

Mathematical symbols

$\mathbf{X}$	Multivariate data set

$\mathbf{x}_i$	The i^{th} sample of $\mathbf{X}$ represented by a vector
$\mathbf{X}_{(l)}^{(tr)}$	The training subset of the l^{th} random partition of $\mathbf{X}$
$\mathbf{X}_{(l)}^{(te)}$	The testing subset of the l^{th} random partition of $\mathbf{X}$
$\mathbf{x}_{(\mathbf{l})j}^{(tr)}$	The j^{th} instance of the subset $\mathbf{X}_{(l)}^{(tr)}$
$K(x)$	Univariate kernel function, evaluated at x
$K(\mathbf{x})$	Multivariate kernel function, evaluated at $\mathbf{x}$
h	Scalar bandwidth value
$\mathbf{H}$	Bandwith matrix
$(\mathbf{a})^T$	Transpose of the vector $\mathbf{a}$
$\mathbf{I}$	Identity matrix
Tr	The trace
$\int_{\mathbf{H}>0}$	Integrating over all positive definite $\mathbf{H}$
$\Gamma(x)$	Gamma function of x
$\Gamma_p(\mathbf{x})$	Multivariate gamma function of x
$\psi(x)$	Digamma function of x
$\psi_p(\mathbf{x})$	Multivariate digamma function of x
$\mathbb{E}$	Expected value
$\mathbb{E}_g$	Expected value in terms of the function g
$\boldsymbol{\Psi_{ij}}$	The i,j-th scale matrix

Dissertation overview

Problem statement

A class imbalance in data result in major problems in the field of classification. The lack of data in one class results in the inaccurate training of classifiers and hence a poor fit of the model. This is especially true in the credit scoring environment. Furthermore, there exists some uncertainty whether parametric or non-parametric models perform better in this setting. Although non-parametric kernel density estimators generally perform well, they lack a much desired predictive power.

Research objectives

This dissertation has multiple research objectives. First of all, the dissertation introduces two types of priors with the aim of addressing the class imbalance problem. The dissertation goes further to investigate which of the abovementioned priors better adress the class imbalance problem for various classifiers. The dissertation also sets out to solve the question of whether parametric or non-parametric classifiers perform better in the class imbalance setting. The final objective of the dissertation is to develop a theoretical approach to multivariate Bayesian kernel density estimation, with the aim of providing kernel density estimators with the much desired predictive power.

Outline of dissertation

Chapter 1 is dedicated to providing the reader with the required background information in order to understand the setting in which the research is performed.
Chapter 2 describes the data used as well as the pre-processing methods applied to the data. The chapter also describes the methods used to evaluate the performance of the models.
Chapter 3 investigates the mathematics underlying the Naive Bayes, Gaussian and Logistic regression classifiers. This is done with the aim of giving the reader a deeper understanding of these classifiers. A very short literature review of Bayesian logistic regression is also in-

cluded in this chapter.

Chapter 4 introduces the theory surrounding non-parametric kernel density estimation. In particular, it elaborates on the Silverman's rule of thumb method of bandwidth estimation as well as the Minimum Leave-One-Out Entropy method of bandwidth estimation. The chapter goes further to introduce two types of priors, that have the aim of addressing the class imbalance problem.

Chapter 5 performs a comprehensive study, investigating the effect of the priors introduced in Chapter 4 as well as investigating the performance of the parametric versus the performance of the non-parametric classifiers.

Chapter 6 starts off by reviewing the univariate approach to Bayesian kernel density estimation as derived by Bernardo (1999). The chapter goes on to develop a theoretical multivariate approach to Bayesian kernel density estimation. It is due to theoretical nature of this chapter and the lack of a simulation study that the chapter is placed second to last in the dissertation.

In Chapter 7 concluding remarks are made and future work to be done are discussed.

Chapter 1

Introduction

Credit scoring utilises historical data and various statistical methods in order to determine the risk associated with loan applications. A scoring model, also known as a scorecard, analyse the performance of previously made loans in order to isolate the characteristics of borrowers that will result in a loan being settled in accordance with the agreed upon conditions. Typically scoring models assign a higher score to applicants considered to be a low risk and vice versa for those considered to be a high risk. The lender can set a threshold scoring value, depending on the lender's risk appetite, upon which all loans with a score exceeding the threshold is approved and all loans with a score below the threshold is denied (Mester 1997).

Linear probability models, probit models, logit models and discriminant analysis models are methods traditionally used in credit scoring models. Discriminant analysis divides loan applicants into various classes based on the risk of default. The other three methods model the probability of default of the loan applicants. Pricing theory models and neural networks are two other models also used to model credit risk. Neural networks utilise training data in order to determine the relationship that exists between the characteristics of borrowers and the probability of default. The importance of the various characteristics, in terms of the amount each characteristic contributes toward the estimated probability of default, is also determined. Mester (1997) states that Since the assumptions surrounding neural networks are more relaxed, the method is more flexible than traditional statistical methods.

A paper by Zekic-Susac et. al compare the performance of logistic regression (LR), neural networks and CART decision trees in the small business credit scoring environment. The paper by Zekic-Susac, Sarlija, and Bensic (2004) concludes that the neural network outperforms the other models. In contrast to this, papers by Castillo et al. (2003), Féraud and Clérot (2002) and Nath, Rajagopalan, and Ryker (1997) outline the poor performance of neural networks in the presence of small data sets or data sets containing irrelevant features. Clearly there exists uncertainty surrounding the use of parametric versus the use of non-parametric models when constructing credit scoring models.

The low default probability problem is a complication often encountered when modeling credit risk. The low default probability problem boils down to a class imbalance problem. There are two factors leading to a class imbalance in this context. The first being the proportion of the one class differing significantly from the other class. That is to say the number of borrowers, in the sample as well as the population, that default on their debt are considerably less than those that do not default on their debt. This shortage of data relating to defaulters often result in a distorted estimate of the probability of default. The second factor is a difference in the proportion of a class in the sample compared to the proportion of that class in the population.

There are three main categories of methods used to solve class imbalance problems. These include the feature selection approach, the algorithmic approach and finally the sampling approach. The sampling approach can be subdivided into under-sampling, in which observations from the majority class is randomly removed, and over-sampling, in which the random observations in the minority class is replicated. Under-sampling has the disadvantage that valuable information is lost, whereas over-sampling results in significantly higher computational time. The algorithmic approach includes algorithms designed to handle large class imbalances. These approaches include one-class classification as well as cost-sensitive learning. Cost-sensitive learning methods aim to maximise the loss-function of a particular data set. The problem with this method is that in most real world cases the cost of misclassification is not known. According to Longadge and Dongre (2013), since high dimensionality is often accompanied with large class imbalances, selecting the features that result in optimal performance is of vital importance.

A paper by Kennedy, Mac Namee, and Delany (2012) compared one-class and two-class classifiers. The paper only utilised the Gaussian and Parzen classifiers as one-class classifiers where the majority class was modelled. It concluded that one-class classification outperforms two-class classification when the proportion of defaulters is very low, typically when defaulters are less than 1% of the sample.

We therefore perform a comprehensive study investigating the performance of parametric versus non-parametric classifiers. Bayes' rule is used to extend the two-class Gaussian, naive Bayes (NB) and Parzen classifiers by modelling both class-conditional probability density functions (PDFs) and accounting for class imbalances through the use of class priors (the class proportions) as well as through the use of Bernoulli priors with a fixed prior parameter p. This is done so that the negative effect of imbalanced data on the performance of two class classification would be reduced. Furthermore, a rudimentary multivariate Bayesian kernel density estimate approach is developed with the aim of affording kernel density estimators a much desired predictive power.

Chapter 2

Data, data pre-processing and model evaluation

The following three data pre-processing techniques are introduced in this chapter: z-Scoring, PCA and the Mahalanobis distance. The three data sets used in this dissertation are introduced and the data pre-processing techniques are applied to them. Finally the model evaluation techniques, in the machine learning context, are discussed and the mathematical background for the metrics used in this dissertation are provided.

2.1 Data pre-processing

In machine learning the saying "garbage in, garbage out" is particularly true. One of the aims of data pre-processing is to prevent data of poor quality resulting in inaccurate results. The majority of pattern recognition algorithms require data to be pre-processed. The original features are transformed into a set of new features with the aim of reducing the complexity of the pattern recognition problem. According to Bishop (2006), another important attribute of data pre-processing is its potential to reduce computational time.

2.1.1 z-Scoring

The z-score or standard score is the number of standard deviations a data point differs from the mean of the values being observed. A positive z-score value indicates a data point above the mean of the observed values, while a negative value indicates a data point below the mean of the observed values. If the z-score value is zero the data point has the same value as the mean of the observed values.
The z-score value is calculated by subtracting the population mean and then dividing by the population standard deviation

$$z = \frac{x - \mu}{\sigma}$$

where μ is the population mean and σ is the population standard deviation. Should the population mean and standard deviation not be known, then the sample mean and standard deviation may be used. This process is also known as standardising or normalising (Murphy 2012).
Some models are sensitive to the scale of the data. By standardising the data all features have the same scale, enabling models and statistical methods that are scale sensitive to be used.

2.1.2 Principal component analysis

Principal component analysis is a method used to reduce the dimensionality of data as well as enhance the ability to interpret the data. It is of vital importance to standardise the data before performing principal component analysis, since it is sensitive to the scaling of the data. Geometrically, the method utilises an orthogonal transformation to rotate the original set of axis, used to represent the features, $\{X_1, \ldots, X_p\}$ to a new set of axis, $\{Y_1, \ldots, Y_p\}$. The transformation ensures that the directions displaying the maximum amount of variation is represented by the new set of axis (Murphy 2012). It is important to keep in mind that this is under the constraint that all of the derived axis are orthogonal to one another, i.e. $Y_1 \perp Y_2 \perp \cdots \perp Y_p$. The derived axis also ensures that the covariance structure is described in a parsimonious fashion. The derived axis are known as the principal components. Since the derived axis are orthogonal to one another, the principal components are uncorrelated.
The principal components may also be viewed as linear combinations of the original data, $\mathbf{X} = \{X_1, \ldots, X_p\}$. The linear combinations are constructed in such a way that the variances of each linear combination is maximised. Let the data, $\mathbf{X}$, have a covariance matrix, $\mathbf{\Sigma}$, with eigenvalues, $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_p \geq 0$ and corresponding eigenvectors, $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_p\}$, then the principal components are

$$\begin{aligned} Y_1 &= \mathbf{e}_1^T \mathbf{X} = e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p \\ Y_2 &= \mathbf{e}_2^T \mathbf{X} = e_{21}X_1 + e_{22}X_2 + \cdots + e_{2p}X_p \\ &\vdots \\ Y_p &= \mathbf{e}_p^T \mathbf{X} = e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p \end{aligned}$$

with respective variances and covariances

$$\begin{aligned} \operatorname{Var}(Y_i) &= \mathbf{e}_i^T \mathbf{\Sigma} \mathbf{e}_i = \lambda_i && \forall i = 1, \ldots, p \\ \operatorname{Cov}(Y_i, Y_j) &= \mathbf{e}_i^T \mathbf{\Sigma} \mathbf{e}_j = 0 && \forall i \neq j \end{aligned}$$

The principal components are therefore defined in such a fashion that they appear in descending order based on the amount of variation explained (Bishop 2006). In other words, the first

principal component explains the most variation with each successive principal component explaining a smaller proportion of variation in the data.
Often most of the variation in the data can be accounted for by only considering a small number $k < p$ of principal components. If this is the case the k principal components, that explains most of the variation, can be used to describe the data instead of the original p variables. Only a small amount of information is lost since the k principal components contain almost as much information as the original variables. The ability to use k principal components to represent the data gives rise to the problem of determining the number of principal components to retain, i.e. the value of k. The subject matter as well as the amount of sample variation explained should be considered when addressing this issue. Assuming the data is standardized, the proportion of variance explained by the i^{th} principal component is $\frac{\lambda_i}{p}$ $\forall i = 1, \ldots, p$. A good rule of thumb is to only use those components that individually explain a proportion of at least $\frac{1}{p}$ of the total variation.

Scree-plot

Another useful tool to help determine the number of principal components is a scree-plot. A scree-plot plots the eigenvalues λ_i, and thus the variation explained by each principal component, against i, the eigenvalue number. The number of principal components to be used is indicated by the position of an elbow or bend in the scree-plot. The number of components that are retained are chosen such that the eigenvalues of the remaining components are fairly small and approximately the same size (Johnson and Wichern 2014).

2.1.3 Mahalanobis distance

The Mahalanobis distance was first introduced by Mahalanobis (1936). The Mahalanobis distance can be thought of as a multivariate method to measure the number of standard deviations some point, $\mathbf{y}$, is from the mean of some distribution, D. If the point $\mathbf{y}$ and the mean of D are equal, the Mahalanobis distance would therefore be zero. By transforming the coordinate system the Mahalanobis distance corresponds to the Euclidean distance (Murphy 2012). The distance is scale-invariant, unit less and it takes the correlations of the data set into account. Considering the set of observations $\mathbf{x} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ with mean $\boldsymbol{\mu} = \{\mu_1, \ldots, \mu_n\}$ and covariance matrix $\mathbf{S}$, the Mahalanobis distance is

$$D = \sqrt{(\mathbf{x} - \boldsymbol{\mu})^T \mathbf{S}^{-1} (\mathbf{x} - \boldsymbol{\mu})}$$

By applying the Mahalanobis distance to a data set with a high dimensional feature space and considering the diagonal of the resulting matrix, the dimensionality can be reduced so that a univariate data set is obtained. This enables us to visualise the data with greater ease.

It is important to ensure that features are not highly correlated, since this may result in the covariance matrix being singular.

2.2 Data

2.2.1 Data summary

A summary of the number of observations and the number of features for the various data sets are given in Table 2.1.

Data set	**Total observations**	**Defaulting observations**	**Non-defaulting observations**	**Features**
German	1000	300	700	24
Australian	690	383	307	14
Lending club	120269	8357	111912	11

Table 2.1: Data summary

2.2.2 German credit scoring data

The German data set is a multivariate data set consisting of twenty features and a thousand instances. Professor Hans Hofmann, of the University of Hamburg, donated the data set in 1994. Originally the data set consisted of integer as well as categorical features. The University of Strathclyde modified the data set, enabling various machine learning algorithms to utilise it. In the modified data set, "german.data-numeric", all categorical variables are replaced with indicator functions. This results in the modified data set consisting of 24 features. The data set has a class imbalance of 300 instances of debtors defaulting and 700 that settle their debt. The data set is available from the machine learning repository (Lichman 2013). Table 2.2 provides a description of the various features encountered in the original data set.

Figure 2.1 illustrates the distributions of the Mahalanobis distance of the respective classes, as explained in Section 2.1.3. The figure indicates that both classes are similarly distributed; both slightly skewed to the right. It might be considered that the Mahalanobis distances of the respective classes follow various forms of a Students-t distribution. Superimposing the distributions of the classes, as done in Figure 2.2, emphasises the presence of a class imbalance in the data, the fact that both classes have similar distributional forms as well as the fact that the locations of the distributions are very similar. Since these locations are very similar, it is clear that using the Mahalanobis distance to reduce the dimensionality of the data is not a viable option for classification purposes.

Considering the scree-plot in Figure 2.3, it quite difficult to determine the optimum number of principal components to retain. An "elbow" is evident at $i = 6$. However, only 40,96%

of the variation is explained should six principal components be used. Another "elbow" can be observed at $i = 21$. Retaining twenty-one principal components results in 95.78% of the variation being explained. It would therefore be more sensible to retain about twenty-one principal components.

Attribute number	**Feature**	**Data type**
1	Status of Checking Account	Qualitative 4 Levels
2	Duration	Quantitative Integer
3	Credit History	Qualitative 5 Levels
4	Purpose	Qualitative 11 Levels
5	Credit Amount	Quantitative Rounded to nearest hundred
6	Savings Account or Bonds	Qualitative 5 Levels
7	Present Employment	Qualitative 5 Levels
8	Instalment Rate	Quantitative Percentage of disposable income
9	Personal Status and Sex	Qualitative 5 Levels
10	Other Debtors	Qualitative 3 Levels
11	Present Residence since	Quantitative Integer
12	Property	Qualitative 4 Levels
13	Age	Quantitative Integer
14	Other Instalment plans	Qualitative 3 Levels
15	Housing	Qualitative 3 Levels
16	Existing Credits at this Bank	Quantitative Real value
17	Job	Qualitative 4 Levels
18	Number of dependence	Quantitative Integer
19	Telephone	Qualitative 2 Levels
20	Foreign Worker	Qualitative 2 Levels

Table 2.2: German data set

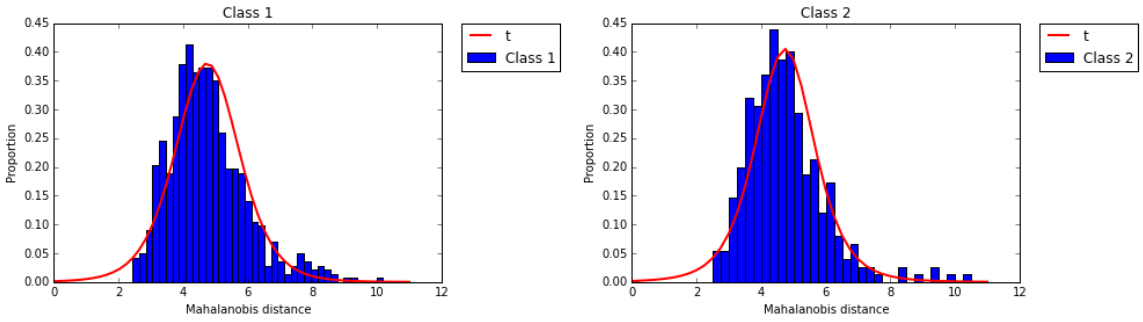

Figure 2.1: Class distribution of Mahalanobis distance of German data

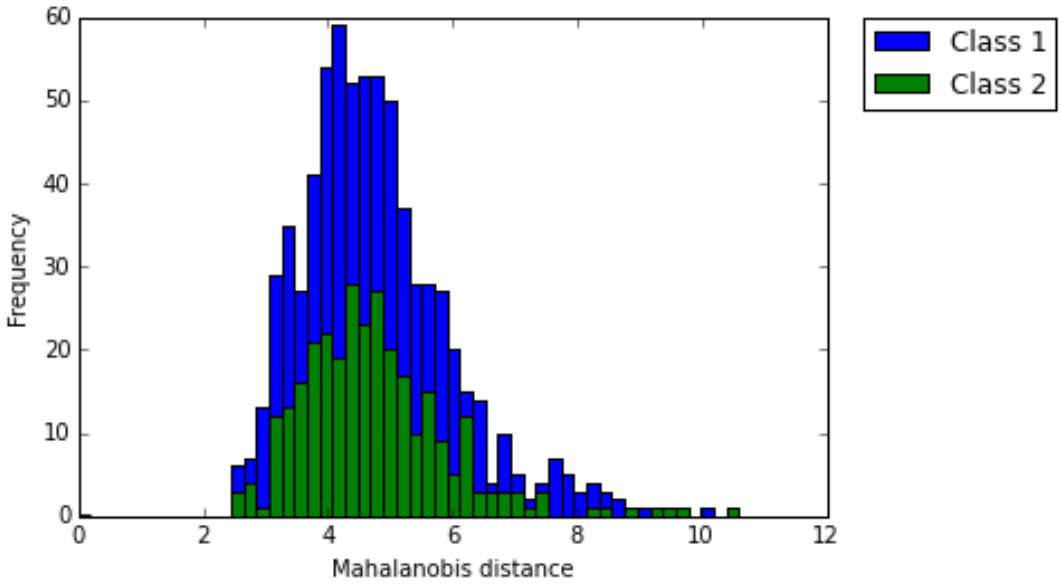

Figure 2.2: Distribution of Mahalanobis distance for German data

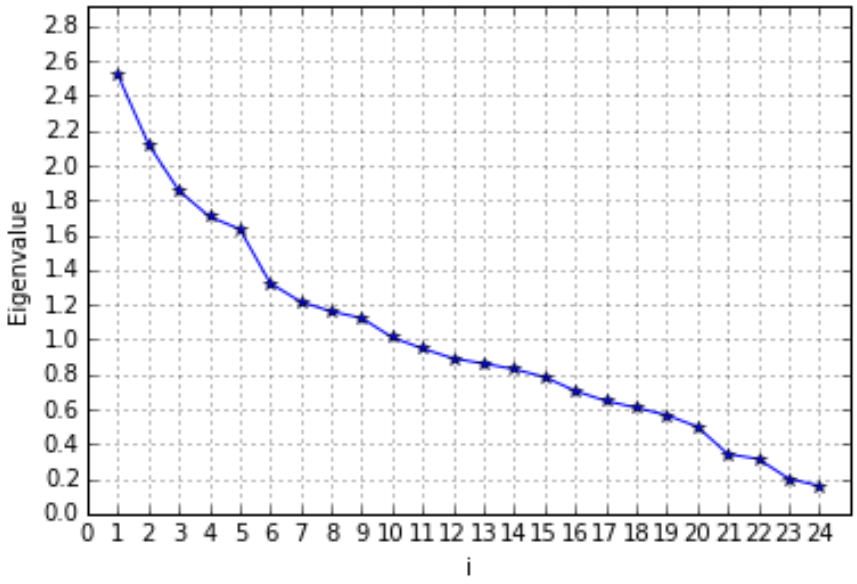

Figure 2.3: German data scree-plot

2.2.3 Australian credit approval data

The Australian credit approval data set represent data relating to credit card applications. The data set has 14 features and 690 observations. The data set has a class imbalance of 307 positive instances and 383 negative instances. Since the data set contains missing values imputation is implemented. The data is divided into the defaulting and non-defaulting classes. The average of each feature for the relative classes are calculated. These values are substituted into the missing continuous values for the respective features in the respective classes. This is known as cell mean imputation. A similar technique where the mode instead of the mean is used for the categorical variables. An additional variable is added to indicate whether a variable is measured or imputed. This method is explained by Lohr (2010). A short description of the data follow in Table 2.3. The official source of the data set is confidential. The data set containing the imputations can be downloaded from the UCI machine learning repository (Lichman 2013).

Figure 2.4 illustrates the distributions of the Mahalanobis distance of the respective classes, as explained in Section 2.1.3. The figure indicates that both classes are similarly distributed; both slightly skewed to the right. Superimposing the distributions of the classes, as done in Figure 2.5, the absence of a large class imbalance in the data is evident. The fact that both classes have similar distributional forms as well as the fact that the locations of the distributions are very similar are also highlighted. Since these locations are very similar, it is clear that using the Mahalanobis distance to reduce the dimensionality of the data is not a viable option for classification purposes.

Considering the scree-plot in Figure 2.6 an "elbow" is seen at $i = 4$, suggesting the use of four principal components. However, this will result in only 40.18% of the variance encountered in the data being explained. From the scree-plot it is difficult to determine the optimal number of principal components to retain. We will therefore use the number of components that results in a minimum of 95% of the variance being explained; particularly in this case thirteen principal components.

Attribute number	Data type	Possible values
1	Categorical	0,1
2	Continuous	
3	Continuous	
4	Categorical	1,2,3
5	Categorical	1,2,3,4,5,6,7,8,9,10,11,12,13,14
6	Categorical	1,2,3,4,5,6,7,8,9
7	Continuous	
8	Categorical	0,1
9	Categorical	0,1
10	Continuous	
11	Categorical	0, 1
12	Categorical	1,2,3
13	Continuous	
14	Continuous	
15	Class	1,2

Table 2.3: Australian credit approval data

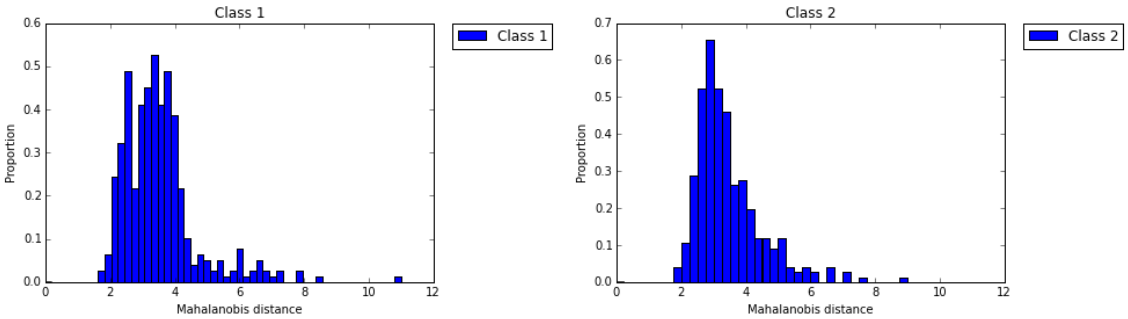

Figure 2.4: Class distribution of Mahalanobis distance of Australian data

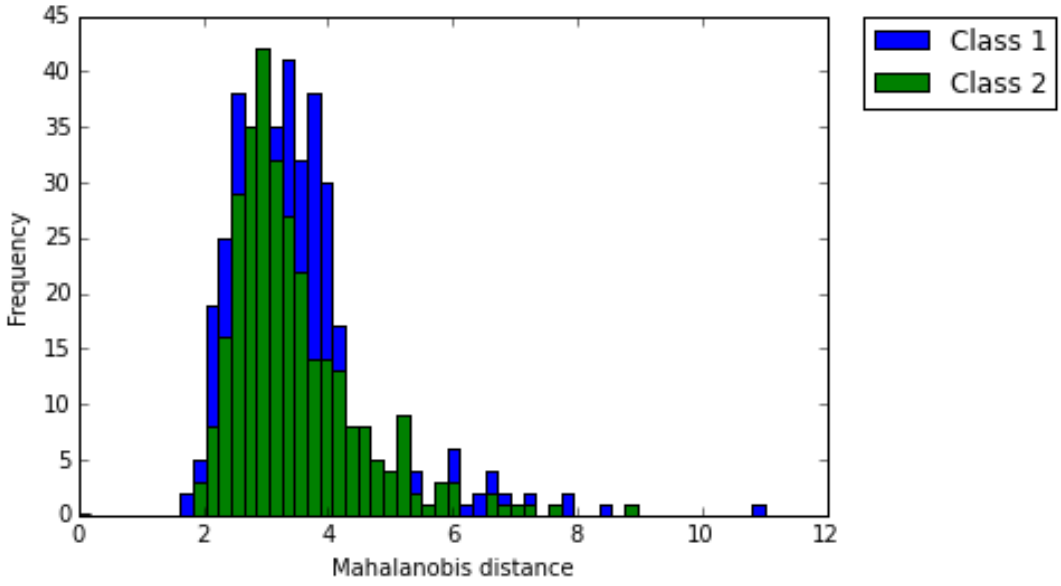

Figure 2.5: Distribution of Mahalanobis distance for Australian data

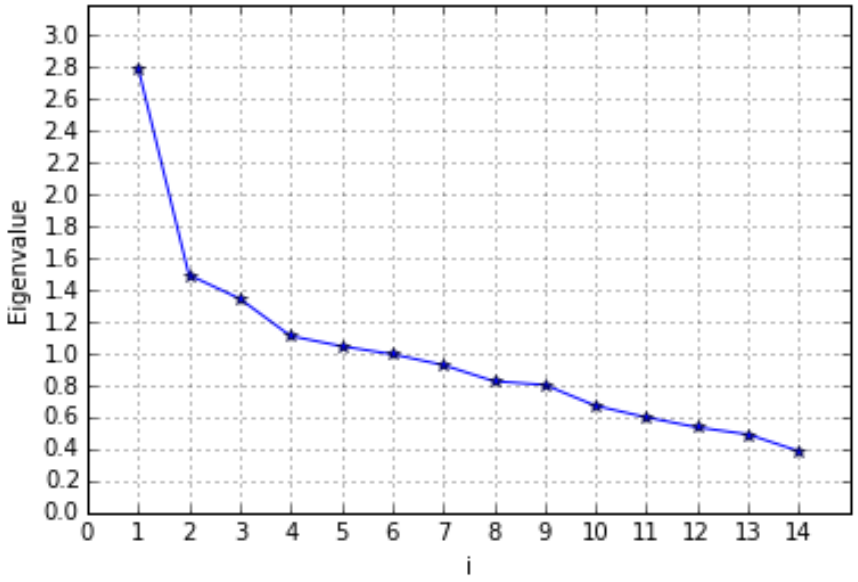

Figure 2.6: Australian data scree-plot

2.2.4 Lending club data

The Lending club data consists of a training as well as a testing data set. The training data set consists of 150000 observations and 10 features, whereas the testing data set contains 100000 observations. However, the testing set does not contain class labels. Both the training and testing sets contain observations with missing values (Lending club 2016). Since the objective of this research is not to investigate the handling of missing values and the data sets are large, all observations containing missing values are removed from the respective data set. This results in a testing set of 81000 and a training set of 120269 observations. A short description of the data is given in Table 2.4.

Figure 2.7 illustrates that the distribution of the Mahalanobis distance of both classes are skewed to the right, with a location close to two. Superimposing these distributions results in the graph given in Figure 2.8. The graph highlights the enormous class imbalance present in the data as well as the fact that the locations of both classes are very similar. The use of the Mahalanobis distance to reduce the multivariate classification problem to a univariate problem is therefore not feasible.

Considering Figure 2.9 two "elbows" are evident; the first at $i = 4$ and the second at $i = 9$. Retaining four principal components results in 57.3% of the variation in the data being explained, whereas if nine principal components are retained 99.63% of the variation is explained.

Attribute number	Description	Data type
1	Class attribute	Categorical
2	Personal credit divided by total credit limit	Percentage
3	Age	Integer
4	Number of late payments (past 30-59 days)	Integer
5	Monthley debt divided by gross income	Percentage
6	Monthley income	Real
7	Number of current loans	Integer
8	Number of late payments (past 90 days)	Integer
9	Number of mortgage and real estate loans	integer
10	Number of late payments (past 60-89 days)	integer
11	Number of dependents	integer

Table 2.4: Lending club data

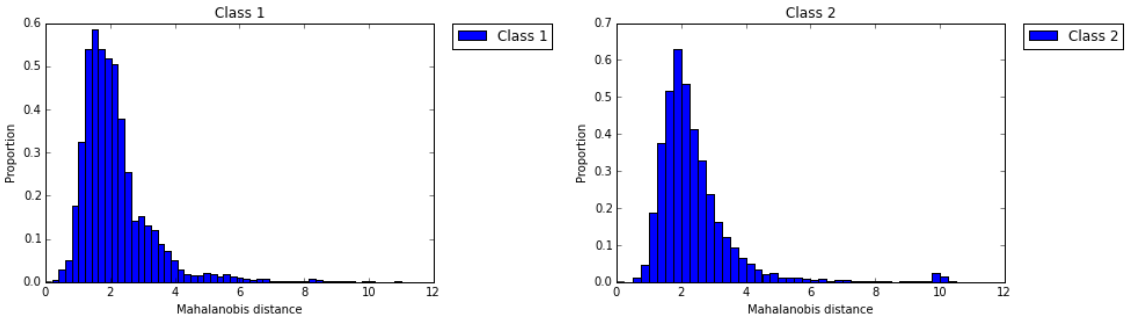

Figure 2.7: Class distribution of Mahalanobis distance of Lending club data

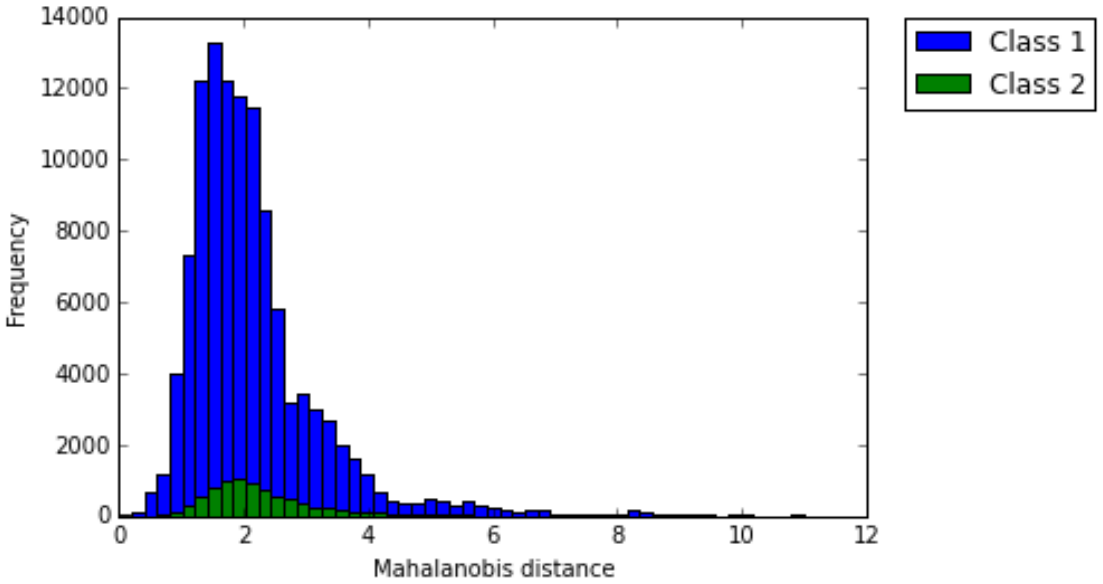

Figure 2.8: Distribution of Mahalanobis distance for Lending club data

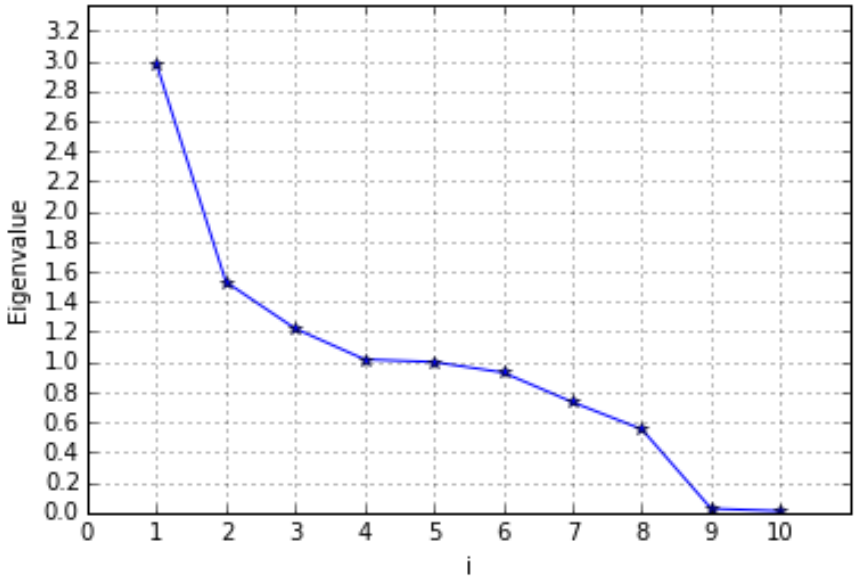

Figure 2.9: Lending club data scree-plot

2.3 Model evaluation

2.3.1 k-Fold cross-validation

Suppose a data set consists out of a training set and a validation set that are independent of one another. In the process of fitting the model, the model is optimised in such a fashion that it over fits the training data. Therefore, the model generally fits the training data considerably better than it fits the validation set. This is especially true if the data set is small. Cross-validation is thus a method used to fit a model to a data set and evaluate its performance when the data set doesn't contain a validation set.
k-Fold cross-validation divides the data set into k folds. Of these k folds, $k-1$ folds are combined to form a training set and the remaining fold serves as the validating set. The model is fitted using the training set and the performance of the fitted model is evaluated using the validation set. The process is repeated for k iterations; during each iteration another fold serves as the validation set. This results in every fold serving as the validation set exactly once, as explained by Bishop (2006) and Giudici and Figini (2009).

2.3.2 Confusion matrices

A confusion matrix is a table that is used to represent the performance of a supervised learning classification model. The table consists out of two rows and two columns; the columns representing the actual class and the rows the predicted class, or vice versa. The cells of the confusion matrix therefore summarise the number of true positives, false positives,

	C_1'	C_2'
C_1	True Positive	False Negative
C_2	False Positive	True Negative

Table 2.5: 2×2 Confusion matrix

true negatives and false negatives (Murphy 2012). An example of a confusion matrix can be viewed in Table 2.5.

2.3.3 Harmonic mean

The harmonic mean assumes the cost of incorrectly classifying instances as being positive to be equal to the cost of incorrectly classifying instances as being negative. This could be problematic if the costs of the type of classification errors differ. Should this be the case a cost matrix is required. However, if cost matrices are not available the harmonic mean is deemed an appropriate performance measure as explained by Kennedy, Mac Namee, and Delany (2012). The accuracy of a classifier for a specific threshold is measured by the harmonic mean. In order to calculate the harmonic mean a confusion matrix is constructed from which the sensitivity and specificity is calculated. The harmonic mean is calculate as

$$\text{Harmonic Mean} = \frac{2 \times \text{Sensitivity} \times \text{Specificity}}{\text{Sensitivity} + \text{Specificity}} \tag{2.1}$$

Take note that according to Giudici and Figini (2009) sensitivity is the proportion of positive instances correctly classified as such and specificity is the proportion of negative instances correctly classified as such.

2.3.4 Count-R^2

According to Gujarati and Porter (1999) the count-R^2 measure is the proportion of instances correctly classified. This measure is also known as the hit rate. It is important to compare the count-R^2 value to a benchmark, since class imbalances can result in a distorted image of performance. One such benchmark is the proportional chance criterion given by $C_{PRO} = q_1^2 + q_2^2$, where q_1^2 and q_2^2 are the respective proportions of the classes. A classifier is considered to perform sufficiently well if the count-R^2 value is a quarter greater than the proportional chance criterion. This method of evaluation is only a rough indication and becomes redundant for large class imbalances.

Chapter 3

Parametric classification

Parametric models have a fixed number of parameters, regardless of the amount of training data. Computationally speaking, parametric models are quite fast to use. However, according to Bishop (2006) and Murphy (2012) these models are very rigid in the assumptions made about the distributions of the data. This chapter provides an overview of the derivations of the naive Bayes (NB), Guassian as well as the logistic regression (LR) classifiers. The chapter goes further to provide a very short literature review of the Bayesian logistic regression (BLR) classifier.

3.1 Naive Bayes classifier

The following section elaborates on the outline for NB classifiers given in Murphy (2012). Let $\mathbf{x} = \{x_1, \ldots, x_p\}$ be a vector consisting of p features. The NB classifier assumes that given the class labels, the features are independent. Even if the independence assumption doesn't hold the model generally perform quite well due to its simplicity and small chance of overfitting. Due to the independence assumption the conditional probability density can be written in terms of the product of the density of each of the features:

$$p\left(\mathbf{x}|y=c\right) = \prod_{j=1}^{p} p\left(x_j|y=c\right) \tag{3.1}$$

As emphasised by Equation 3.1 the NB classifier depends on the probability density of each feature. For features consisting of real values, $x_j \in \mathbb{R}\ \forall j = \{1, \ldots, p\}$, the Gaussian distribution with mean μ_j and variance σ_j^2 is used. Equation 3.1 therefore becomes

$$p\left(\mathbf{x}|y=c\right) = \prod_{j=1}^{p} N\left(x_j|\mu_j, \sigma_j^2\right)$$

3.1.1 Parameter estimation

In order to fit the model it is necessary to estimate the relevant parameters. This is done using maximum likelihood. The likelihood is given by

$$\begin{aligned}
L\left(\mu,\sigma^2\right) &= p\left(x,y|\mu,\sigma^2\right) \\
&= \prod_{i=1}^{n} p\left(y_i\right) \prod_{j=1}^{p} p\left(x_{ij}|y_i,\mu_j,\sigma_j^2\right) \\
&= \prod_{i=1}^{n} \prod_{c} \pi_c^{\Delta} \prod_{j=1}^{p} p\left(x_{ij}|,\mu_{jc},\sigma_{jc}^2\right)^{\Delta} \\
&= \prod_{i=1}^{n} \prod_{c} \pi_c^{\Delta} \prod_{j=1}^{p} N\left(x_{ij}|\mu_{jc},\sigma_{jc}^2\right)^{\Delta} \\
&= \prod_{i=1}^{n} \prod_{c} \pi_c^{\Delta} \prod_{j=1}^{p} \left(\frac{1}{\sqrt{2\pi\sigma_{jc}^2}} e^{-\frac{(x_{ij}-\mu_{jc})^2}{2\sigma_{jc}^2}} \right)^{\Delta}
\end{aligned}$$

where $\pi_c = p\left(y_i = c\right)$ and

$$\Delta = I(y_i = c) = \begin{cases} 1 & \text{if } y_i = c \\ 0 & \text{otherwise} \end{cases}$$

Taking the natural logarithm of the likelihood, the log-likelihood is

$$\begin{aligned}
&\ln\left(L\left(\mu,\sigma^2\right)\right) \\
&= \ln\left(\prod_{i=1}^{n} \prod_{c} \pi_c^{\Delta} \prod_{j=1}^{p} \left(\frac{1}{\sqrt{2\pi\sigma_j^2}} e^{-\frac{(x_{ij}-\mu_j)^2}{2\sigma_j^2}} \right)^{\Delta} \right) \\
&= \sum_{i=1}^{n} \sum_{c} I\left(y_i = c\right) \ln\left(\pi_c\right) - \frac{p}{2} \sum_{i=1}^{n} \sum_{c} I\left(y_i = c\right) \ln\left(2\pi\right) \\
&- \sum_{i=1}^{n} \sum_{c} I\left(y_i = c\right) \sum_{j=1}^{p} \ln\left(\sigma_j\right) - \sum_{i=1}^{n} \sum_{c} I\left(y_i = c\right) \sum_{j=1}^{p} \frac{\left(x_{ij}-\mu_j\right)^2}{2\sigma_j^2} \\
&= \sum_{c} n_c \ln\left(\pi_c\right) - \frac{p}{2} \sum_{c} n_c \ln\left(2\pi\right) - \sum_{c} n_c \sum_{j=1}^{p} \ln\left(\sigma_j\right) - \sum_{i=1}^{n} \sum_{c} I\left(y_i = c\right) \sum_{j=1}^{p} \frac{\left(x_{ij}-\mu_j\right)^2}{2\sigma_j^2} \\
&= \sum_{c} n_c \ln\left(\pi_c\right) - \frac{np}{2} \ln\left(2\pi\right) - n \sum_{j=1}^{p} \ln\left(\sigma_j\right) - \sum_{i=1}^{n} \sum_{c} I\left(y_i = c\right) \sum_{j=1}^{p} \frac{\left(x_{ij}-\mu_j\right)^2}{2\sigma_j^2}
\end{aligned} \tag{3.2}$$

Differentiating the log-likelihood towards μ_k, for some $k \in \{1, \ldots, p\}$, and setting the derivative equal to zero the MLE of the mean is obtained

$$\begin{aligned}
\frac{\partial L\left(\mu, \sigma^2\right)}{\partial \mu_k} &= \sum_{i=1}^{n} \sum_{c} I\left(y_i = c\right) \frac{x_{ik} - \mu_k}{\sigma_k^2} = 0 \\
\sum_{i=1}^{n} \sum_{c} I\left(y_i = c\right) \mu_k &= \sum_{i=1}^{n} \sum_{c} I\left(y_i = c\right) x_{ik} \\
\hat{\mu}_k &= \frac{1}{n} \sum_{i=1}^{n} x_{ik} \\
&= \bar{x}_k
\end{aligned}$$

the sample mean of the k^{th} feature. Differentiating the log-likelihood towards σ_k^2, for some $k \in \{1, \ldots, p\}$, and setting the derivative equal to zero the MLE of the variance is obtained

$$\begin{aligned}
\frac{\partial L\left(\mu, \sigma^2\right)}{\partial \sigma_k^2} &= -\frac{n}{2\sigma_k^2} + \sum_{i=1}^{n} \sum_{c} I\left(y_i = c\right) \frac{\left(x_{ik} - \mu_k\right)^2}{2\sigma_k^4} = 0 \\
\sigma_k^2 &= \frac{1}{n} \sum_{i=1}^{n} (x_{ik} - \mu_k)^2
\end{aligned}$$

the sample variance of the k^{th} feature. In order to calculate the MLE of the prior, π_k, Lagrangian multipliers are used. The constraint $\sum_c \pi_c = 1$ is added to the log-likelihood given in Equation 3.2

$$\begin{aligned}
\ln\left(L\left(\mu, \sigma^2, \lambda\right)\right) = &\sum_{c} n_c \ln\left(\pi_c\right) - \frac{np}{2} \ln\left(2\pi\right) - n \sum_{j=1}^{p} \ln\left(\sigma_j\right) \\
&- \sum_{i=1}^{n} \sum_{c} I\left(y_i = c\right) \sum_{j=1}^{p} \frac{\left(x_{ij} - \mu_j\right)^2}{2\sigma_j^2} + \lambda \left(1 - \sum_{c} \pi_c\right)
\end{aligned} \tag{3.3}$$

Differentiating Equation 3.3 with respect to the prior, π_k for some $k \in \{1, \ldots, C\}$ and setting the derivative equal to zero, the MLE of the prior is obtained

$$\begin{aligned}
\frac{\partial L\left(\mu, \sigma^2, \lambda\right)}{\partial \pi_k} &= \frac{n_k}{\pi_k} - \lambda = 0 \\
\hat{\pi}_k &= \frac{n_k}{\lambda}
\end{aligned} \tag{3.4}$$

This estimate of the prior, π_k contains an unknown quantity, λ. Lambda is solved by using the restriction $\sum_c \pi_c = 1$

$$\begin{aligned}
\sum_{c} \hat{\pi}_c &= \sum_{c} \frac{n_c}{\lambda} \\
1 &= \frac{1}{\lambda} \sum_{c} n_c \\
\lambda &= n
\end{aligned} \tag{3.5}$$

Substituting lambda, given in Equation 3.5, into the expression for the prior of a class, given in Equation 3.4, it is seen that the estimate of the prior is simply the proportion of instances in that particular class

$$\hat{\pi}_k = \frac{n_k}{n}$$

3.2 Gaussian discriminative analysis

This section is based on the overiew of Gaussian classifiers given in Murphy (2012).
Gaussian discriminative analysis utilises multivariate normal densities to define class conditional densities. The model therefore assume the data to be normally distributed. It is important to note that, even though possible confusion due to the name, Gaussian discriminative analysis is not a discriminative, but a generative model. Let $\mathbf{x} = \{x_1, \ldots, x_p\}$ be a vector consisting of p features. The class conditional densities are

$$p\left(\mathbf{x}|y=c\right) = \frac{1}{(2\pi)^{\frac{p}{2}}\left|\Sigma_c\right|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}-\mu_\mathbf{c})^T\Sigma_c^{-1}(\mathbf{x}-\mu_\mathbf{c})} \tag{3.6}$$

Using Bayes' theorem the posterior probability of classifying $\mathbf{x}$ to the class $y = c$ is

$$p\left(y=c|\mathbf{x}\right) = \frac{\pi_c p\left(\mathbf{x}|y=c\right)}{\sum_{j=1}^{C}\pi_j p\left(\mathbf{x}|y=j\right)} \tag{3.7}$$

where $\pi_i = p(y = i)$ is the prior probability of $\mathbf{x}$ belonging to class i. A new instance is classified to the class with the largest posterior probability. Considering Equation 3.7, it is necessary to calculate estimates for the class priors, class conditional means and covariance matrices. In the special case where Σ_c is a diagonal matrix the model simplifies to a NB classification model. If the covariance matrices differ for the various classes the classification model used is quadratic discriminant analysis. However, should the covariance matrices be similar and a pooled estimate for the covariance matrix is used then the classification model used is linear discriminant analysis.

3.2.1 Parameter estimation

The parameters can be estimated using maximum likelihood estimation. The likelihood function is

$$\begin{aligned}
L\left(\boldsymbol{\mu}, \Sigma\right) &= \prod_{i=1}^{n} p\left(y_i = c\right) p\left(\mathbf{x}_i|y_i, \boldsymbol{\mu}, \Sigma\right) \\
&= \prod_{i=1}^{n}\prod_{c} \pi_c^{\Delta} p\left(\mathbf{x}_i|\boldsymbol{\mu}, \Sigma\right)^{\Delta} \\
&= \prod_{i=1}^{n}\prod_{c} \pi_c^{\Delta}\left(\frac{1}{(2\pi)^{\frac{p}{2}}\left|\Sigma_c\right|^{\frac{1}{2}}} e^{-\frac{1}{2}(\mathbf{x}_i-\boldsymbol{\mu}_\mathbf{c})^T\Sigma_c^{-1}(\mathbf{x}_i-\boldsymbol{\mu}_\mathbf{c})}\right)^{\Delta}
\end{aligned}$$

Taking the natural logarithm of the likelihood, the log-likelihood is

$$\begin{aligned}
&\ln\left(L\left(\boldsymbol{\mu},\Sigma\right)\right)\\
&=\sum_{i=1}^{n}\sum_{c} I\left(y_i=c\right)\ln\left(\pi_c\right)-\frac{p}{2}\ln\left(2\pi\right)\sum_{i=1}^{n}\sum_{c} I\left(y_i=c\right)+\frac{1}{2}\sum_{i=1}^{n}\sum_{c} I\left(y_i=c\right)\ln\left|\Sigma_c^{-1}\right|\\
&-\frac{1}{2}\sum_{i=1}^{n}\sum_{c} I\left(y_i=c\right)\left(\mathbf{x}_i-\boldsymbol{\mu_c}\right)^T\Sigma_c^{-1}\left(\mathbf{x}_i-\boldsymbol{\mu_c}\right) \qquad (3.8)\\
&=\sum_{c} n_c\ln\left(\pi_c\right)-\frac{np}{2}\ln\left(2\pi\right)+\frac{1}{2}\sum_{c} n_c\ln\left|\Sigma_c^{-1}\right|-\frac{1}{2}\sum_{i=1}^{n}\sum_{c} I\left(y_i=c\right)\left(\mathbf{x}_i-\boldsymbol{\mu_c}\right)^T\Sigma_c^{-1}\left(\mathbf{x}_i-\boldsymbol{\mu_c}\right) \qquad (3.9)
\end{aligned}$$

Differentiating the log-likelihood towards $\boldsymbol{\mu_k}$ for some $k\in\{1,\ldots,C\}$ and setting the derivative equal to zero, the MLE of the mean is obtained

$$\begin{aligned}
\frac{\partial\ln\left(L\left(\boldsymbol{\mu},\Sigma\right)\right)}{\partial\boldsymbol{\mu}_k}&=\sum_{i=1}^{n} I\left(y_i=k\right)\Sigma_k^{-1}\left(\mathbf{x}_i-\boldsymbol{\mu_k}\right)=0\\
\sum_{i=1}^{n} I\left(y_i=k\right)\boldsymbol{\mu_k}&=\sum_{i=1}^{n} I\left(y_i=k\right)\mathbf{x}_i\\
n_k\boldsymbol{\mu_k}&=\sum_{i=1}^{n} I\left(y_i=k\right)\mathbf{x}_i\\
\hat{\boldsymbol{\mu}}_k&=\frac{\sum_{i=1}^{n} I\left(y_i=k\right)\mathbf{x}_i}{n_k}\\
\hat{\boldsymbol{\mu}}_k&=\bar{\mathbf{x}}_k
\end{aligned}$$

where $\bar{\mathbf{x}}_k$ is the average vector of the k^{th} class. In order to calculate the MLE of the covariance matrix the log-likelihood first have to be written in the form

$$\begin{aligned}
&\ln\left(L\left(\boldsymbol{\mu},\Sigma\right)\right)\\
&=\sum_{c} n_c\ln\left(\pi_c\right)-\frac{np}{2}\ln\left(2\pi\right)+\frac{1}{2}\sum_{c} n_c\ln\left|\Sigma_c^{-1}\right|-\frac{1}{2}\sum_{i=1}^{n}\sum_{c} I\left(y_i=c\right)\left(\mathbf{x}_i-\boldsymbol{\mu_c}\right)^T\Sigma_c^{-1}\left(\mathbf{x}_i-\boldsymbol{\mu_c}\right)\\
&=\sum_{c} n_c\ln\left(\pi_c\right)-\frac{np}{2}\ln\left(2\pi\right)+\frac{1}{2}\sum_{c} n_c\ln\left|\Sigma_c^{-1}\right|\\
&-\frac{1}{2}\sum_{i=1}^{n}\sum_{c} I\left(y_i=c\right)\mathrm{Tr}\left[\left(\mathbf{x}_i-\boldsymbol{\mu_c}\right)\left(\mathbf{x}_i-\boldsymbol{\mu_c}\right)^T\Sigma_c^{-1}\right]
\end{aligned}$$

Differentiating the log-likelihood towards Σ_k^{-1} for some $k \in \{1, \ldots, C\}$ and setting the derivative equal to zero, the MLE of the covariance matrix is obtained

$$\begin{aligned}
\frac{\partial \ln\left(L\left(\boldsymbol{\mu}, \Sigma\right)\right)}{\partial \Sigma_k^{-1}} &= \frac{n_k}{2}\Sigma_k^T - \frac{1}{2}\sum_{i=1}^{n} I\left(y_i = k\right)\left(\mathbf{x}_i - \boldsymbol{\mu_k}\right)\left(\mathbf{x}_i - \boldsymbol{\mu_k}\right)^T = 0 \\
\frac{n_k}{2}\Sigma_k^T &= \frac{1}{2}\sum_{i=1}^{n} I\left(y_i = k\right)\left(\mathbf{x}_i - \boldsymbol{\mu_k}\right)\left(\mathbf{x}_i - \boldsymbol{\mu_k}\right)^T \\
\hat{\Sigma}_k &= \frac{\sum_{i=1}^{n} I\left(y_i = k\right)\left(\mathbf{x}_i - \boldsymbol{\mu_k}\right)\left(\mathbf{x}_i - \boldsymbol{\mu_k}\right)^T}{n_k} \\
\hat{\Sigma}_k &= \mathbf{S_k}
\end{aligned}$$

where $\mathbf{S_k}$ is the sample covariance matrix of the k^{th} group. In order to determine the MLE of the class priors Lagrangian multipliers are used. The constraint $\sum_c \pi_c = 1$ is added to the log-likelihood given in Equation 3.9

$$\begin{aligned}
L\left(\boldsymbol{\mu}, \Sigma, \lambda\right) = &\sum_c n_c \ln\left(\pi_c\right) - \frac{np}{2}\ln\left(2\pi\right) + \frac{1}{2}\sum_c n_c \ln\left|\Sigma_c^{-1}\right| \\
&- \frac{1}{2}\sum_{i=1}^{n}\sum_c I\left(y_i = c\right)\left(\mathbf{x}_i - \boldsymbol{\mu_c}\right)^T \Sigma_c^{-1}\left(\mathbf{x}_i - \boldsymbol{\mu_c}\right) + \lambda\left(1 - \sum_c \pi_c\right)
\end{aligned} \tag{3.10}$$

Differnetiating the log-likelihood in Equation 3.10 towards π_k for some $k \in \{1, \ldots, C\}$ and setting the derivative equal to zero, the MLE of the class prior is obtained

$$\begin{aligned}
\frac{\partial \ln\left(L\left(\boldsymbol{\mu}, \Sigma, \lambda\right)\right)}{\partial \pi_k} &= \frac{n_k}{\pi_k} - \lambda = 0 \\
\hat{\pi}_k &= \frac{n_k}{\lambda}
\end{aligned} \tag{3.11}$$

Using the constraint $\sum_c \pi_c = 1$ lambda can be solved. Taking the summation over the classes on both sides of Equation 3.11

$$\begin{aligned}
\sum_c \hat{\pi}_c &= \sum_c \frac{n_c}{\lambda} \\
1 &= \frac{n}{\lambda} \\
\lambda &= n
\end{aligned} \tag{3.12}$$

Substituting Equation 3.12 into Equation 3.11, the MLE of the prior for the various classes are obtained

$$\hat{\pi}_k = \frac{n_k}{n}$$

3.2.2 Estimated posterior probabilities

Once the parameters are estimated it is required to estimate the posterior probabilities in order to do classification. Combining the MLE's of the parameters with Equations 3.7 and

Equation 3.6 the posterior probabilities can be estimated

$$\begin{aligned}
\hat{p}\left(y=c|\mathbf{x}\right) &= \frac{\hat{\pi}_c \hat{p}\left(\mathbf{x}|y=c\right)}{\sum_{j=1}^{C} \hat{\pi}_j \hat{p}\left(\mathbf{x}|y=j\right)} \\
&= \frac{\hat{\pi}_c (2\pi)^{-\frac{p}{2}} \left|\hat{\Sigma}_c\right|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\hat{\mu}_c)^T \hat{\Sigma}_c^{-1}(\mathbf{x}-\hat{\mu}_c)}}{\sum_{j=1}^{C} \hat{\pi}_j (2\pi)^{-\frac{p}{2}} \left|\hat{\Sigma}_j\right|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\hat{\mu}_j)^T \hat{\Sigma}_j^{-1}(\mathbf{x}-\hat{\mu}_j)}} \\
&= \frac{\frac{n_c}{n} (2\pi)^{-\frac{p}{2}} \left|\mathbf{S}_c\right|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_c)^T \mathbf{S}_c^{-1}(\mathbf{x}-\bar{\mathbf{x}}_c)}}{\sum_{j=1}^{C} \frac{n_j}{n} (2\pi)^{-\frac{p}{2}} \left|\mathbf{S}_j\right|^{-\frac{1}{2}} e^{-\frac{1}{2}(\mathbf{x}-\bar{\mathbf{x}}_j)^T \mathbf{S}_j^{-1}(\mathbf{x}-\bar{\mathbf{x}}_j)}}
\end{aligned}$$

3.3 Logistic regression

An overview of logistic regression (LR) is available in the textbook by Hastie, Tibshirani, and Friedman (2008).

Let $\mathbf{x} = \{x_1, \ldots, x_p\}$ be a vector consisting of p features and let C be the class so that $C \in \{1, \ldots, k\}$. The i^{th} odds ratio is defined as

$$odds_i = \frac{P\left(C=i|X=x\right)}{P\left(C=k|X=x\right)}$$

The LR model is defined in terms of the odds ratio

$$\begin{aligned}
\ln\left(odds_1\right) &= \ln\left(\frac{P\left(C=1|X=x\right)}{P\left(C=k|X=x\right)}\right) = \alpha_1 + \beta_1^T x \\
\ln\left(odds_2\right) &= \ln\left(\frac{P\left(C=2|X=x\right)}{P\left(C=k|X=x\right)}\right) = \alpha_2 + \beta_2^T x \\
&\vdots \\
\ln\left(odds_{k-1}\right) &= \ln\left(\frac{P\left(C=k-1|X=x\right)}{P\left(C=k|X=x\right)}\right) = \alpha_{k-1} + \beta_{k-1}^T x
\end{aligned}$$

It is clear that the model consists of $k-1$ logit transformations. Since the probabilities of all the classes sum to one, the probability of the k^{th} class can be rewritten as

$$\begin{aligned}
P\left(C=k|X=x\right) &= 1 - \sum_{i=1}^{k-1} P\left(C=i|X=x\right) \\
P\left(C=k|X=x\right) &= 1 - P\left(C=k|X=x\right) \sum_{i=1}^{k-1} e^{\alpha_i + \beta_i^T x} \\
\left(1 + \sum_{i=1}^{k-1} e^{\alpha_i + \beta_i^T x}\right) P\left(C=k|X=x\right) &= 1 \\
P\left(C=k|X=x\right) &= \frac{1}{1 + \sum_{i=1}^{k-1} e^{\alpha_i + \beta_i^T x}}
\end{aligned}$$

In the case of two-class data, as is the case in credit scoring, the model simplifies to

$$\ln\left(\frac{P\left(C=1|X=x\right)}{P\left(C=2|X=x\right)}\right) = \alpha_1 + \beta_1^T x$$

Let $p(x,\theta) = P(C=1|X=x)$, so that $1-p(x,\theta) = P(C=2|X=x)$. Let y_i be the response variable, so that when $c_i = 1$ then $y_i = 1$ and when $c_i = 2$ then $y_i = 0$. The conditional log-likelihood is expressed as

$$\begin{aligned}
L(X,\theta) &= \sum_{i=1}^{n} \ln\left(P(C=c_i|X=x_i)\right) \\
&= \sum_{i=1}^{n} \left[\ln\left(p(x_i,\theta)\right) + \ln\left(1-p(x_i,\theta)\right)\right] \\
&= \sum_{i=1}^{n} \left[y_i \ln\left(p(x_i,\theta)\right) + (1-y_i)\ln\left(1-p(x_i,\theta)\right)\right] \\
&= \sum_{i=1}^{n} \left[y_i \ln\left(p(x_i,\theta)\right) + \ln\left(1-p(x_i,\theta)\right) - y_i \ln\left(1-p(x_i,\theta)\right)\right] \\
&= \sum_{i=1}^{n} \left[y_i \ln\left(\frac{p(x_i,\theta)}{1-p(x_i,\theta)}\right) + \ln\left(1-p(x_i,\theta)\right)\right] \\
&= \sum_{i=1}^{n} \left[y_i \ln\left(\frac{\frac{e^{\theta^T x_i}}{1+e^{\theta^T x_i}}}{1-\frac{e^{\theta^T x_i}}{1+e^{\theta^T x_i}}}\right) + \ln\left(1-\frac{e^{\theta^T x_i}}{1+e^{\theta^T x_i}}\right)\right] \\
&= \sum_{i=1}^{n} \left[y_i \ln\left(e^{\theta^T x_i}\right) + \ln\left(\frac{1}{1+e^{\theta^T x_i}}\right)\right] \\
&= \sum_{i=1}^{n} \left[y_i\left(\theta^T x_i\right) - \ln\left(1+e^{\theta^T x_i}\right)\right]
\end{aligned}$$

The maximum of the log-likelihood is obtained by differentiating with respect to θ and equating the derivatives to zero

$$\frac{\partial L(X,\theta)}{\partial \theta} = \sum_{i=1}^{n}\left[y_i x_i - \frac{x_i e^{\theta^T x_i}}{1+e^{\theta^T x_i}}\right] = \sum_{i=1}^{n} x_i\left(y_i - p(x_i,\theta)\right) = 0 \tag{3.13}$$

The derivative of the log-likelihood, given in Equation 3.13, results in $p+1$ non-linear equations in terms of θ. These equations have to be solved numerically. In order to use the Newton-Raphson algorithm the second derivative of the log-likelihood has to be calculated

$$\begin{aligned}
\frac{\partial^2 L(X,\theta)}{\partial\theta\partial\theta^T} &= \sum_{i=1}^{n} \frac{-x_i x_i^T e^{\theta^T x_i}\left(1+e^{\theta^T x_i}\right) - x_i x_i^T e^{2\theta^T x_i}}{\left(1+e^{\theta^T x_i}\right)^2} \\
&= \sum_{i=1}^{n} \frac{-x_i x_i^T e^{\theta^T x_i}}{\left(1+e^{\theta^T x_i}\right)^2} \\
&= -\sum_{i=1}^{n} x_i x_i^T p(x_i,\theta)\left(1-p(x_i,\theta)\right)
\end{aligned}$$

The iterative Newton-Raphson algorithm used to determine the solution for θ is given by

$$\theta^{new} = \theta^{old} - \left(\frac{\partial^2 L(X,\theta)}{\partial\theta\partial\theta^T}\right)^{-1}\left(\frac{\partial L(X,\theta)}{\partial\theta}\right)$$

3.4 Bayesian logistic regression

A paper by Holmes and Knorr-Held (2003) suggests an approach to Bayesian logistic regression (BLR) models that is efficient to use in conjunction with a block Gibbs sampler. However, this approach requires the sampling from a Kolmogorov-Smirnov distribution. Since the focus of this dissertation is not BLR, the Laplace approximation method will be utilised as explained in (Murphy 2012). This classification technique is included purely for comparison purposes.

Chapter 4

Non-parametric classification

Considering parametric models there is a desire to overcome the restrictions enforced by the rigid assumptions made regarding the distributions of the data. This can be done through the use of non-parametric models. Non-parametric models are considerably more flexible in the distributional assmuptions made. However, one of the most important characteristics of a non-parametric model is that the number of parameters increase, as the amount of training data increase. This results in non-parametric models being computationally intractable for data sets consisting of a large amount of observations (Murphy 2012).

4.1 Kernel density estimation overview

In 1985 Rosenblat developed the Kernel Density Estimation method. The aim of this new method was to address the discontinuities encountered when performing Nave Density Estimation. Non-parametric kernel density estimation uses a kernel function to estimate the density function of the data. A kernel function is fitted over each data point, with the data point forming the centre of the kernel function. Finally the contributions of the kernel functions are summed for each data point, resulting in the estimated density function. That is to say; considering the univariate case, let $\mathbf{x} = \{x_1, \ldots, x_n\}$ be a random sample from an unknown distribution, $p(.)$, then the density estimated using non-parametric kernel density estimation is

$$\begin{aligned} p(x|h) &\approx \frac{1}{n}\sum_{i=1}^{n} K_h(x - x_i) \\ &= \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x - x_i}{h}\right) \end{aligned} \tag{4.1}$$

where h is the bandwidth and $K(.)$ the kernel function.
Considering the multivariate case, let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ be a random sample, consisting of n instances each with p features, from an unknown distribution, $p(.)$. The multivariate density,

estimated using non-parametric kernel density estimation, is

$$\begin{aligned} p\left(\mathbf{x}|\mathbf{H}\right) &\approx \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}}\left(\mathbf{x}-\mathbf{x}_i|\mathbf{H}\right) \\ &= \frac{1}{n}\left|\mathbf{H}\right|^{-\frac{1}{2}}\sum_{i=1}^{n} K\left(\mathbf{H}^{-\frac{1}{2}}\left(\mathbf{x}-\mathbf{x}_i\right)\right) \end{aligned} \tag{4.2}$$

where $K\left(.\right)$ represents the multivariate kernel function and $\mathbf{H}$ is a symmetric, positive definite, $p \times p$ matrix.

4.1.1 Kernel function

The kernel function is a predetermined function that integrates to one, is non-negative and it has a mean of zero. That is to say, for the kernel function $K\left(.\right)$, in the univariate case the following holds $\forall t \in \mathbb{R}$

$$K\left(t\right) > 0, \qquad \int_{\mathbb{R}} K\left(t\right) dt = 1, \qquad \mathbb{E}\left[K\left(t\right)\right] = 0$$

and for the multivariate case the following holds $\forall \mathbf{t} \in \mathbb{R}^p$

$$K_{\mathbf{H}}\left(\mathbf{t}\right) > 0, \qquad \int_{\mathbb{R}^p} K_{\mathbf{H}}\left(\mathbf{t}\right) d\mathbf{t} = 1, \qquad \mathbb{E}\left[K_{\mathbf{H}}\left(\mathbf{t}\right)\right] = \mathbf{0}$$

The uniform density, Gaussian density, triangular, biweight, triweight and Epanechnikov functions are just a few functions that are commonly used as kernel functions. It is important to take into account that the chosen kernel function transfers it's smoothness properties to the estimated density function. The choice of kernel function should therefore be based on the desired mathematical properties. Taking this into account, one of the major advantages of using the Gaussian density as kernel function is it's desired properties of continuity and differentiability (Van der Walt and Barnard 2013). For the univariate case, using a Gaussian density as kernel function the estimated density given in Equation 4.1 becomes

$$p\left(\mathbf{x}|h\right) \approx \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\sqrt{2\pi}h} e^{-\left(\frac{x-x_i}{h}\right)^2}$$

Considering the multivariate case, using a multivariate Gaussian density as kernel function, the estimate density in Equation 4.2 becomes

$$p\left(\mathbf{x}|\mathbf{H}\right) \approx \frac{1}{n}\sum_{i=1}^{n} \frac{1}{\left(2\pi\right)^{\frac{p}{2}}\left|\mathbf{H}\right|^{\frac{1}{2}}} e^{-\frac{1}{2}\left(\mathbf{x}-\mathbf{x}_i\right)^T\mathbf{H}^{-1}\left(\mathbf{x}-\mathbf{x}_i\right)}$$

4.1.2 Bandwidth estimation

The bandwidth determines the smoothness of the resulting estimated density. The smaller the value of the bandwidth the smoother the resulting density estimate. The bandwidths are

estimated by optimising objective functions that serves as measures for goodness-of-fit. The three measures for goodness-of-fit that are commonly used include the mean squared error (MSE), mean integrated squared error (MISE) and the asymptotic mean integrated squared error (AMISE). There exist various methods to estimate the bandwidth. The three main methods include plug-in, cross-validation and rule-of-thumb bandwidth estimation.

Plug-in methods estimate the bandwidth by optimising the AMISE with respect to the bandwidth. It attempts to express this solution in a closed-form expression. Methods that attempt to optimise the AMISE with respect to the bandwidth requires the density function, f, that we are attempting to estimate using KDE. Since the density function is unknown, plug-in methods require an initial estimate of the unknown density, f. If the initial estimate is determined using KDE the initial estimate also requires a bandwidth to be estimated. The Hall Sheather Jones Marron (HSJM) estimator can be used to determine all the unknowns required to calculate the bandwidth of the initial estimate of the unknown density, f.

Cross-validation methods can estimate the bandwidth through either least squares or maximum likelihood. When least squares are used in conjunction with cross-validation the integrated squared error (ISE) is optimised. The ISE requires the difference between the estimated and true densities. This difference is approximated using leave-one-out cross-validation. If maximum likelihood is used in conjunction with cross-validation the likelihood function is maximised with respect to the bandwidth. It is important to use leave-one-out cross-validation in order to prevent situations where the bandwidth is zero, resulting in a maximum likelihood estimate that is infinite.

Just like the plug-in estimators, rule-of-thumb estimators optimise the AMISE with respect to the bandwidth. As mentioned before, the optimization of the AMISE requires the unknown densities, f. However, instead of using an initial estimate for the unknown densities, as done by plug-in estimators, rule-of-thumb estimators make certain distributional assumptions regarding these unknown densities (Van der Walt 2014).

4.2 Silverman's rule of thumb

Silverman's rule-of-thumb optimise the AMISE with respect to the kernel bandwidth, thus (as the name suggests) classifying it as a rule-of-thumb estimator. As mentioned in Section 4.1.2, rule-of-thumb estimators make certain distributional assumptions regarding the unknown distribution encountered in the formulation of the AMISE. Silverman suggested assuming the unknown distribution to be Gaussian. Considering the univariate case, let $\hat{\sigma}^2$ be the sample variance and n the number of instances in the sample. Then the optimal bandwidth, assuming the use of a Gaussian kernel function, is

$$h = 1.06\hat{\sigma}n^{-\frac{1}{5}}$$

The reader is referred to Appendix A for a formal proof of Silverman's univariate rule of thumb. By estimating the bandwidth independently for each dimension the estimated bandwidth is extended to the multivariate case. Let $\mathbf{H}_{ii}$ be the bandwidth of the i^{th} feature, $\hat{\sigma}_i^2$ the variance of the i^{th} feature and p the number of features. Then the multivariate Silverman rule-of-thumb kernel bandwidth estimator is

$$\mathbf{H}_{ii}^{\frac{1}{2}} = \left(\frac{4}{p+2}\right)^{\frac{1}{p+4}} n^{\frac{-1}{p+4}} \hat{\sigma}_i$$

Van der Walt and Barnard (2013) compared the performance of traditional bandwidth estimators in high-dimensional feature spaces. They concluded that the Silverman rule-of-thumb estimator performed competitively well for all the high-dimensional data sets investigated. They go further to suggest it's use as an initialising estimator for iterative bandwidth estimation methods.

4.3 Minimum leave-one-out entropy

Unlike some of the conventional bandwidth estimators, the MLE bandwidth estimator is feasible for density estimation in higher dimensions. This can be attributed it's ability to adjust the bandwidth accordingly for each individual data point. The MLE estimator also has the advantage of being able to estimate larger bandwidths, often required by outliers. The MLE and the MLL (maximum leave-one-out likelihood) have similar equations. The only difference being that the numerator as well as the denominator of the MLE equation are normalised. The normalization results in data points encountered in regions with low density having an increased effect on the estimated bandwidth. Similarly, the normalization results in data points encountered in regions with high density having a decreased effect on the estimated bandwidth. As a result the MLE and MLL estimators have similar performance, with the MLE slightly outperforming the MLL estimator for the investigated data sets (Van der Walt 2014).
The MLE estimator requires an initial bandwidth matrix. The MLE estimator can be considered a plug-in estimator in cases where instead of following an iterative scheme, the estimator is initialised and updated once. Van der Walt and Barnard (2017) suggest the use of Silverman's rule-of-thumb estimator as the initialising bandwidth matrix.

4.3.1 Leave-one-out likelihood estimation

As mentioned before, when maximising the log-likelihood with respect to the bandwidth, the likelihood for a point will tend to infinity as the bandwidth for that data point tends to zero. This is overcome by using a leave-one-out estimate. The data point for which the density is evaluated is left out of the summation. Thus, the likelihood for the specific data point is

calculated by using the remaining $n-1$ data points to estimate the density. The expression for the LOUT approach to estimate the density for some point $\mathbf{x}_i$ is

$$p_{\mathbf{H}(-i)}(\mathbf{x}_i) = \frac{1}{n-1}\sum_{j\neq i} K_{\mathbf{H}_j}(\mathbf{x}_i - \mathbf{x}_j|\mathbf{H}_j) \tag{4.3}$$

The log-likelihood can now be calculated by taking the logarithm of the product of Equation 4.3

$$\begin{aligned} l_{\mathbf{H}(-i)}(\mathbf{X}) &= \sum_{i=1}^{n} \ln\left[p_{\mathbf{H}(-i)}(\mathbf{x}_i)\right] \\ &= \sum_{i=1}^{n} \ln\left[\frac{1}{n-1}\sum_{j\neq i} K_{\mathbf{H}_j}(\mathbf{x}_i - \mathbf{x}_j|\mathbf{H}_j)\right] \end{aligned} \tag{4.4}$$

The situation where the likelihood function tends to infinity as the bandwidth tends to 0 is thus circumvented by optimising Equation 4.4 with respect to the bandwidth.

4.3.2 Diagonal bandwidth matrix

The MLE bandwidth estimator is derived using the LOUT estimator. The MLE estimator is obtained by optimising the log-likelihood of the LOUT expression in Equation 4.4, with respect to bandwidth $\mathbf{H}_k$ of some point $\mathbf{x}_k$. It can be shown that the partial derivative of the log-likelihood of the LOUT expression is

$$\frac{d}{d\mathbf{H}_k} l_{MLE}(\mathbf{X}) = \frac{1}{n-1}\sum_{i=1}^{n} \frac{\frac{\partial}{\partial \mathbf{H}_k}\left[K_{\mathbf{H}_k}(\mathbf{x}_i - \mathbf{x}_k)\right]}{p_{\mathbf{H}(-i)}(\mathbf{x}_i)} \tag{4.5}$$

Equation 4.5 clearly illustrates the dependence of the derivative on the kernel function. The MLE has been derived for an univariate Gaussian kernel, a multivariate Gaussian kernel with full covariance matrix and a multivariate Gaussian kernel with a diagonal covariance matrix (Van der Walt 2014). However, we will only focus on the multivariate Gaussian kernel with diagonal covariance matrix.

Consider the challenge of estimating the MLE diagonal bandwidth matrix. Let $K_{h_{jd}}(.)$ represent the univariate Gaussian kernel function, with bandwidth h_{jd}, that is centered on the d^{th} feature of the data point $\mathbf{x}_j$. Furthermore let $\mathbf{H}_j$ be a $p \times p$ diagonal matrix with entries $\mathbf{H}_{j(d,d)} = h_{jd}^2$. The multivariate kernel function can be expressed as the product of the univariate kernel functions

$$K_{\mathbf{H}_j}(\mathbf{x}_i - \mathbf{x}_j|\mathbf{H}_j) = \prod_{d=1}^{p} K_{h_{jd}}\left(\frac{x_{id} - x_{jd}}{h_{jd}}\right) \tag{4.6}$$

Substituting the product of univariate kernels in Equation 4.6 into Equation 4.5 and setting the resulting equation equal to zero, an estimate for the bandwidth in the l^{th} dimension for

some data point $\mathbf{x}_k$ can be obtained. The l^{th} entry of the resulting diagonal MLE bandwidth matrix, for some point $\mathbf{x}_k$ is

$$\mathbf{H}_{k(l,l)} = \frac{\sum_{i=1}^{n} \frac{K_{\mathbf{H}_k}(\mathbf{x}_i - \mathbf{x}_k|\mathbf{H}_k)\left(x_{il} - x_{jl}\right)^2}{p_{\mathbf{H}(-i)}(\mathbf{x}_i)}}{\sum_{i=1}^{n} \frac{K_{\mathbf{H}_j}(\mathbf{x}_i - \mathbf{x}_k|\mathbf{H}_k)}{p_{\mathbf{H}(-i)}(\mathbf{x}_i)}}$$

It is important to realise that the use of a diagonal bandwidth matrix doesn't impair the kernel density estimator's ability to model correlation. It is still possible for the kernel density estimator to capture covariance to some degree. The assumption of independence between variables are therefore not the same as assuming a diagonal covariance matrix.
The diagonal MLE estimator has a time complexity of order $O\left(n^3 p\right)$. For each of the n diagonal bandwidth matrices there are p parameters that require estimation. Therefore, in total np parameters have to be estimated.

4.4 Class priors

Two-class classification methods take into account the distributions of both classes, whereas one-class classification methods only account for a single class's distribution. In the credit scoring environment this can be translated as two-class classifiers accounting for the distributions relating to both borrowers that settle their debt, as well as those that default. On the other hand, one-class classifiers only take into account the distribution of debtors that settle their debt. Intuitively it is expected that two-class classifiers outperform one-class classifiers, since two-class classifiers utilise all available information. One-class classifiers are traditionally associated with outlier and anomaly detection. The performance of one-class classifiers is compared to that of two-class classifiers in the credit scoring environment by Kennedy, Mac Namee, and Delany (2012). The paper fits the one-class classifying models on the majority class. Their results indicate that two-class classifiers are outperformed by one-class classifiers for very large class imbalances, especially when the minority class make out less than 1% of the investigated sample.
In order to overcome the adverse effect of the class imbalance, on the two-class classifiers, we investigate the use of priors in conjunction with Bayes' rule. Bayes' theorem, as it is known today, originated from a paper published by Bayes in 1963. This paper, entitled "An Essay Towards Solving a Problem in the Doctrine of Chances", describes the theorem of elementary probability theory, giving rise to Bayes' rule. The notion that the posterior distribution is proportional to the likelihood was made by Fienberg (2006). Bayes' rule suggests that the posterior distribution is proportional to the likelihood multiplied by the prior distribution. That is to say

$$p\left(\alpha|\mathbf{X}\right) \propto p\left(\mathbf{X}|\alpha\right) p\left(\alpha\right)$$

4.4.1 Frequentist priors

The "frequensist priors" can be thought of as weights that sum to one. These weights are calculated as the proportion of instances in the respective classes. That is to say, in the case of credit scoring, the priors are the proportion of defaulters and non-defaulters. Suppose n is the number of instances in the data set and n_{C_i} for $i = 1, 2$ is the number of instances in the class C_i, then priors are

$$\pi_{C_1} = \frac{n_{C_1}}{n}$$
$$\pi_{C_2} = \frac{n_{C_2}}{n}$$

This implies that as the class imbalance is altered the values of the priors change. The prior aims to reduce the impact of the class imbalance by assigning a larger weight to the likelihood scores of the majority class and assigning a smaller weight to the likelihood scores of the minority class. This method can be thought of as penalising the minority class; thereby placing larger emphasis on the class that contains more data.

4.4.2 Bayesian priors

One of the shortcomings of the frequentist prior is the fact that the prior is restricted to the class imbalance. This can be overcome by placing a discrete prior distribution on the respective classes. Since we are considering a two-class problem, a Bernoulli prior with parameter $p \in (0, 1)$ is placed on the respective classes. The posterior distribution is obtained by multiplying the Bernoulli prior of each class with the likelihood function of each class as calculated with the respective two-class classifiers. In mathematical terms that is: $P(X) = P(C_i)\, P(X|C_i)$, $\forall i \in \{1, 2\}$. The Bernoulli distribution, and thus the prior, is given by

$$P(C) = p^c (1 - p)^{1-c} \tag{4.7}$$

with $c \in \{0, 1\}$. From Equation 4.7 it is clear that by setting the prior parameter p equal to the proportion of the class imbalance encountered in the data, the prior reduce to the frequentist prior encountered in Section 4.4.1. It also follows that setting $p = 0.5$ is equivalent to performing two-class classification without the use of any priors, setting $p = 0$ classifies all instances to the class $C_1 = 0$ and setting $p = 1$ classifies all instances to the class $C_2 = 1$. The choice of the prior parameter p is subsequently of great importance. It is therefore recommended to do a grid search; choosing the prior parameter p to be the value that maximise the respective model performance evaluation measures.

It is of vital importance to realise that the use of a Bernoulli prior will not result in Bayesian properties. This is due to the fact that the Bernoulli prior is placed on the class label, which is not a feature in the data set, nor a parameter of the model, but rather a dependent variable.

Therefore, the Bernoulli prior will result in frequentist properties with the added flexibility of choosing the hyperparameter that best suits the class imbalance encountered in the data.

Chapter 5

Application

In Chapter 4 the class priors were introduced. It is of vital importance to understand the role these priors play on the performance of various parametric as well as non-parametric classifiers. This is particularly true for large class imbalances. Furthermore, it is of interest to investigate how parametric classifiers compare to non-parametric classifiers, in terms of performance, for various class imbalances. This chapter can be considered to be a comprehensive study set out to investigate the abovementioned points of interest.

5.1 Experimental design

Experiments are designed with the purpose of addressing the following questions:

- What is the optimal Bernoulli prior parameter?
- Which prior selection performs best - Bernoulli, frequentist or no prior?
- Which class of classifiers performs the best - parametric or non-parametric?

The details are discussed in this subsection.

5.1.1 Bernoulli priors

In the case of the Bernoulli prior a grid search is done in order to determine the optimal prior parameter, p, for each of the two-class classifiers. The performance of the respective classifiers are evaluated in terms of the harmonic mean as well as the count-R^2/ hit rate value for the various prior parameter values. The performance with respect to the various priors is evaluated at various class imbalances. The class imbalance is altered by reducing the number of observations in the defaulting class. The results for each classifier are tabulated. It is also graphically represented with the class imbalances or default ratios represented on the x-axis of the graphs. Note that even though the class imbalance is varied the prior parameter

remains fixed regardless of the class imbalance.
This is done for the German, Australian and Lending club data sets. For each of the data sets the data is z-scored and PCA 95% is performed on the data. This results in a total of six data sets on which the experiment is performed.

5.1.2 Comparison of priors

The performance of the optimal prior parameter, as determined in Section 5.1.1, is compared to the performance of the respective classifiers with no prior and a frequentist prior. The performance is evaluated over a range of class imbalances. The performance is measured in terms of the harmonic mean as well as the count-R^2/ hit rate. This is done for the German, Australian and Lending club data sets. For each of the data sets the data is z-scored and PCA 95% is performed on the data. This results in a total of six data sets on which the experiment is performed. Since the German and Australian data sets do not have accompanying testing sets 10-fold cross-validation is applied to the data. The results are tabulated in such a fashion that the performance of the z-scored data can be compared to that of the PCA 95% data. The prior resulting in optimal performance for a specific default ratio is indicated with a green cell, the worst performing prior with a red cell and the intermediate performing prior with a yellow cell. By presenting the results in this fashion the table forms a type of heatmap. The overall best performing prior (i.e. between the green cells of the PCA95% and the z-scored results) is indicated by a bold font.

5.1.3 Parametric versus non-parametric classifiers

The optimal Bernoulli prior, with regard to each individual default ratio, is applied to the Gaussian, NB, Silverman and MLE classifiers. The performance of this configuration for these classifiers are not only compared to one another, but also to the BLR as well as the LR classifiers. This is done over an increasing class imbalance. This is performed on the PCA 95% as well as the z-scored data for the German, Australian and Lending club data sets. 10-Fold cross-validation is performed on the German as well as the Australian data sets, whereas 5-fold cross-validation is performed on the Lending club data set. Due to the large size of the Lending club data set, the MLE classifier is computationally intractable and therefore not applied to the data set.

5.2 Results

5.2.1 Bernoulli priors

German data

When considering the Gaussian classifier it is seen that regardless whether the data is z-scored or whether PCA 95% is performed on the data, the classifier has a general increasing performance. The performance, in terms of the hit rate and the harmonic mean, is increasing regardless of prior used. The increase in hit ratio as the class imbalance increase is expected, as in the worst case the classifier can simply classify to the majority class and thereby present an increasing hit rate. See figures 5.1, 5.2, 5.3 and 5.4.
From figures 5.2 and 5.4 it is clear that the hit rate of the classifier increase as the prior decrease. It follows that a prior of $p = 0.1$ should be used to obtain the optimal hit rate for the Gaussian classifier applied to both the z-scored as well as the PCA95% data.
The prior $p = 0.1$ results in inconsistent performance in terms of the harmonic mean. This is true for the z-scored as well as the PCA95% data. This can be interpreted as too much weight being assigned to the minority class. The shortage of data in the minority class results in the erratic behaviour. Identifying the prior that optimise the performance in terms of the harmonic mean isn't as simplistic as in the case of the hit rate. There are numerous priors that perform well across the various class imbalances, with no one classifier outperforming the rest for all considered default ratios. Considering the PCA 95% data, the priors $p = 0.3$, 0.4, 0.5 and 0.6 all perform well. The prior $p = 0.6$ is outperformed by the before mentioned priors for the majority of class imbalances and only becomes competitive at larger class imbalances. On the other hand the priors $p = 0.4$ and 0.5 outperform all priors for the majority of class imbalances. This is illustrated in Figure 5.3. Figure 5.1 highlights the fact that the priors $p = 0.3$, 0.4 and 0.5 perform well, in terms of the harmonic mean, across the various default ratios for the z-scored data. However, the prior $p = 0.3$ is outperformed by the priors $p = 0.8$, 0.7, 0.6, 0.5 and 0.4 at the greatest tested class imbalance.

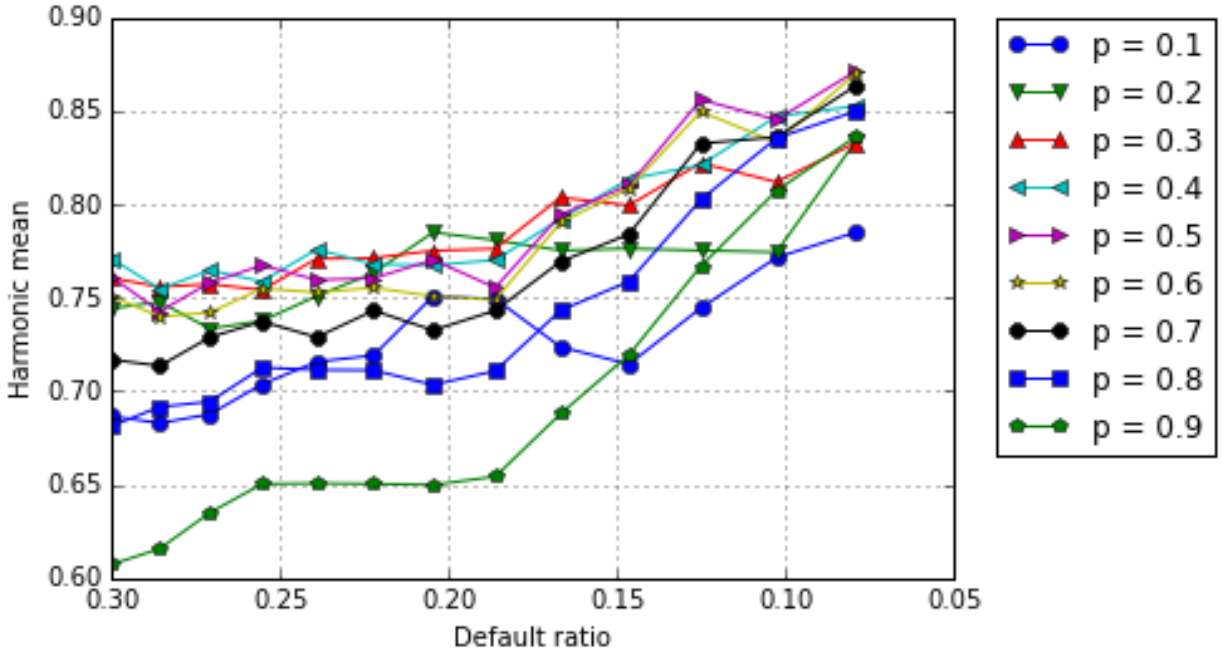

Figure 5.1: Performance of Gaussian classifier with Bernoulli priors: German z-scored data

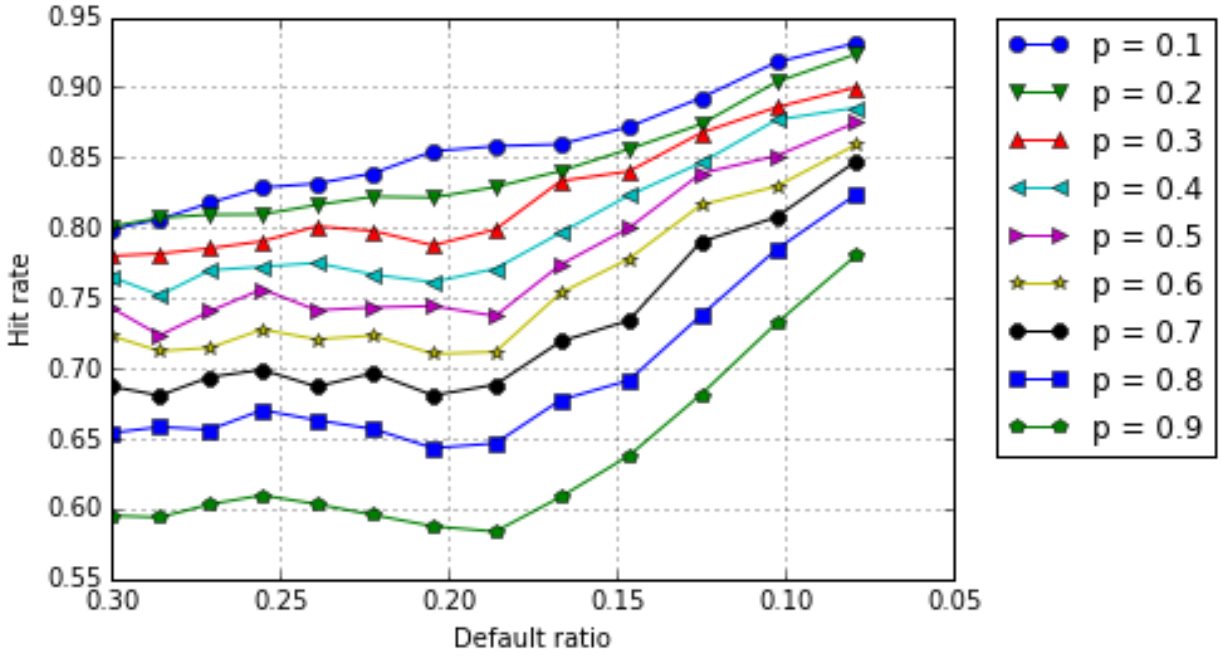

Figure 5.2: Performance of Gaussian classifier with Bernoulli priors: German z-scored data

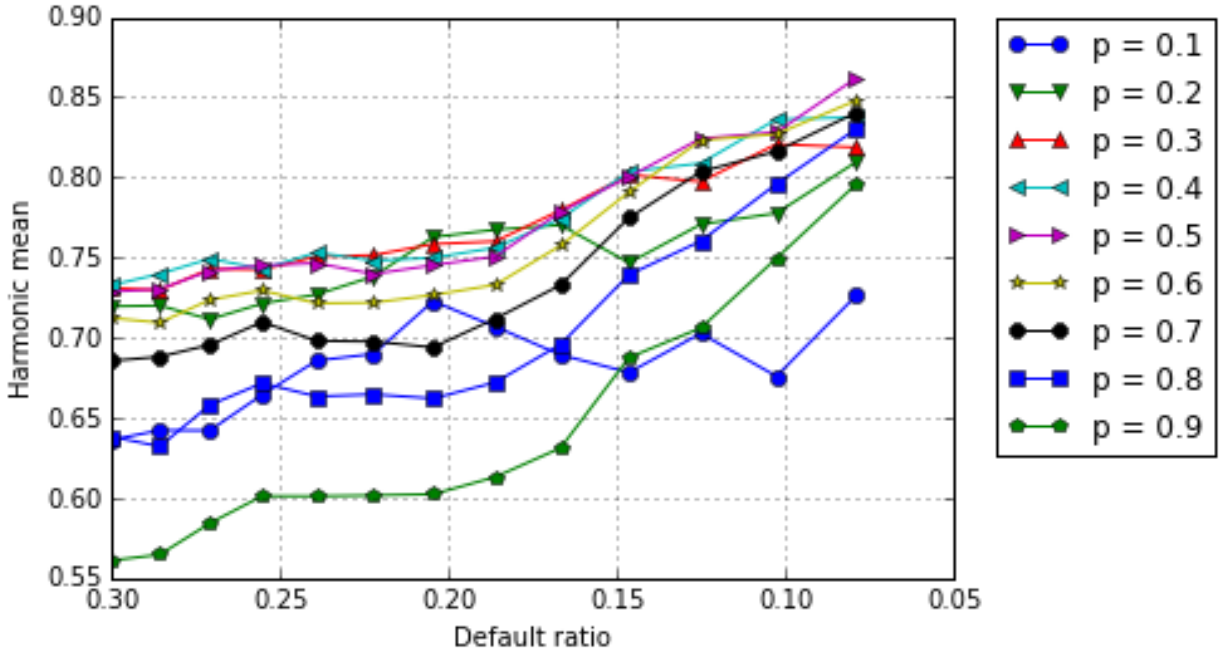

Figure 5.3: Performance of Gaussian classifier with Bernoulli priors: German PCA 95% data

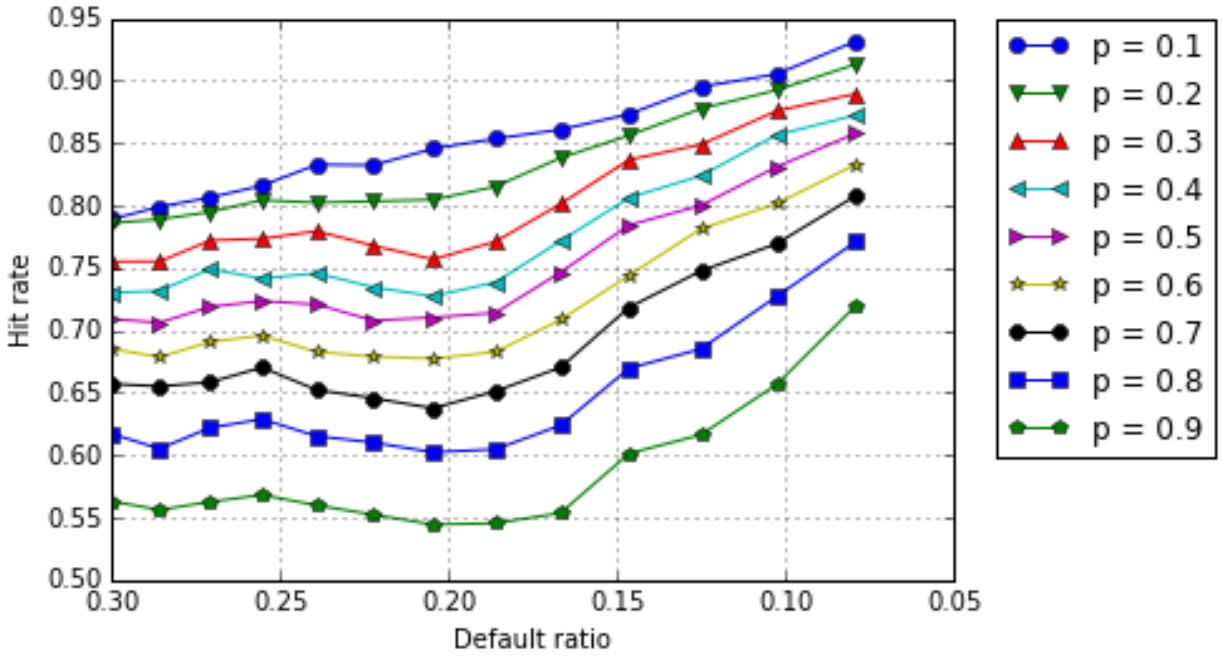

Figure 5.4: Performance of Gaussian classifier with Bernoulli priors: German PCA 95% data

The hit rate of the NB classifier applied to the z-scored data exhibit an increasing trend for the higher performing priors. As the ranking (in terms of performance) of the prior

decrease, the hit rate seems constant up to a default ratio just above 25% after which the hit rate decrease slightly before eventually increasing. The hit rate of the classifier generally increase as the prior decrease, for both the z-scored data as well as the PCA 95% data. An exception to this is the prior $p = 0.2$ applied to the PCA 95% data, which outperforms the prior $p = 0.1$ for the majority of class imbalances. In the case of the PCA95% data the prior $p = 0.9$ exhibit a decreasing trend, the priors $p = 0.8$, 0.7 and 0.6 remain fairly constant and finally the priors $p = 0.5$ to $p = 0.1$ exhibit an increasing trend. The prior $p = 0.1$ has the greatest slope. This is illustrated in Figures 5.6 and 5.8.

It is interesting to note that the prior placed on the NB classifier and the prior's compliment result in similar performance in terms of harmonic mean for the PCA95% data. The harmonic mean of the classifier increase as the priors approach $p = 0.5$. That is to say the priors $p = 0.1$ and $p = 0.9$ result in the lowest harmonic mean. The best performing priors remain fairly constant in performance, whereas the priors performing poorly have an increasing trend. See Figure 5.7. Considering the z-scored data, erratic behaviour in the prior $p = 0.1$ for the NB classifier is observed. Most of the priors have a general upward trend, whereas the prior $p = 0.1$ initially increase significantly after which it significantly decrease. This may be an indication of too much weight being placed on the minority class. The priors $p = 0.3$ to $p = 0.6$ all perform well across the various default ratios. It is clear that the choice of prior will depend on the default ratio. With the aim of addressing large class imbalances the use of no prior or even $p = 0.6$ might be considered. This is illustrated in Figure 5.5.

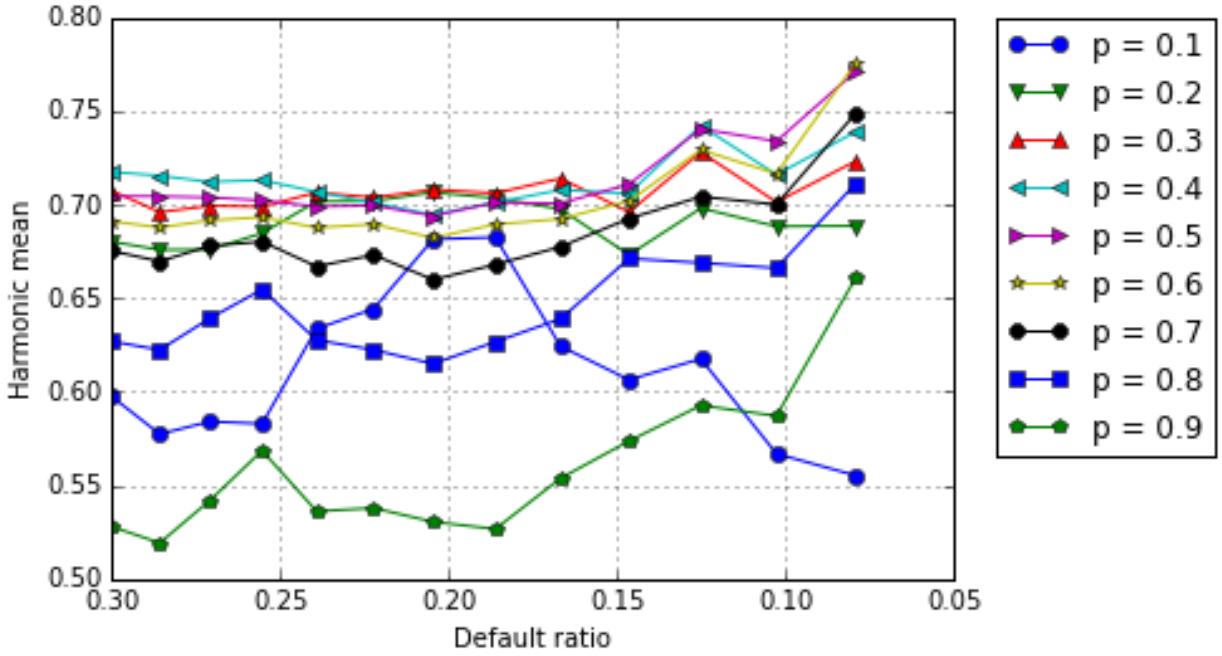

Figure 5.5: Performance of NB classifier with Bernoulli priors: German z-scored data

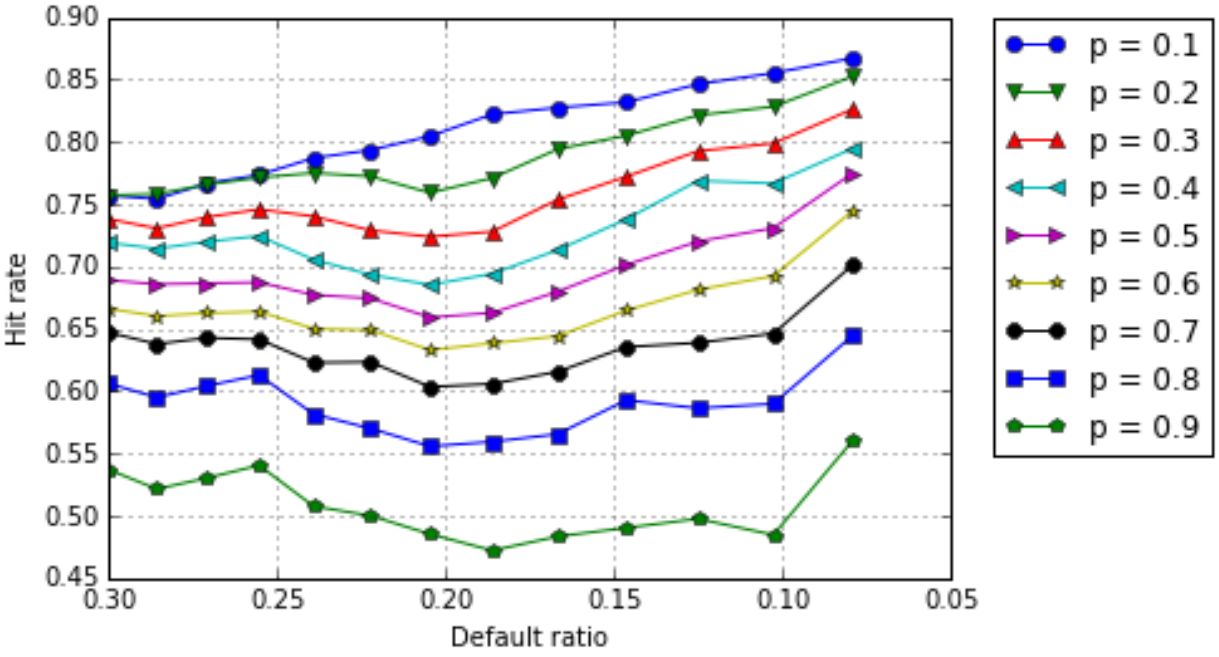

Figure 5.6: Performance of NB classifier with Bernoulli priors: German z-scored data

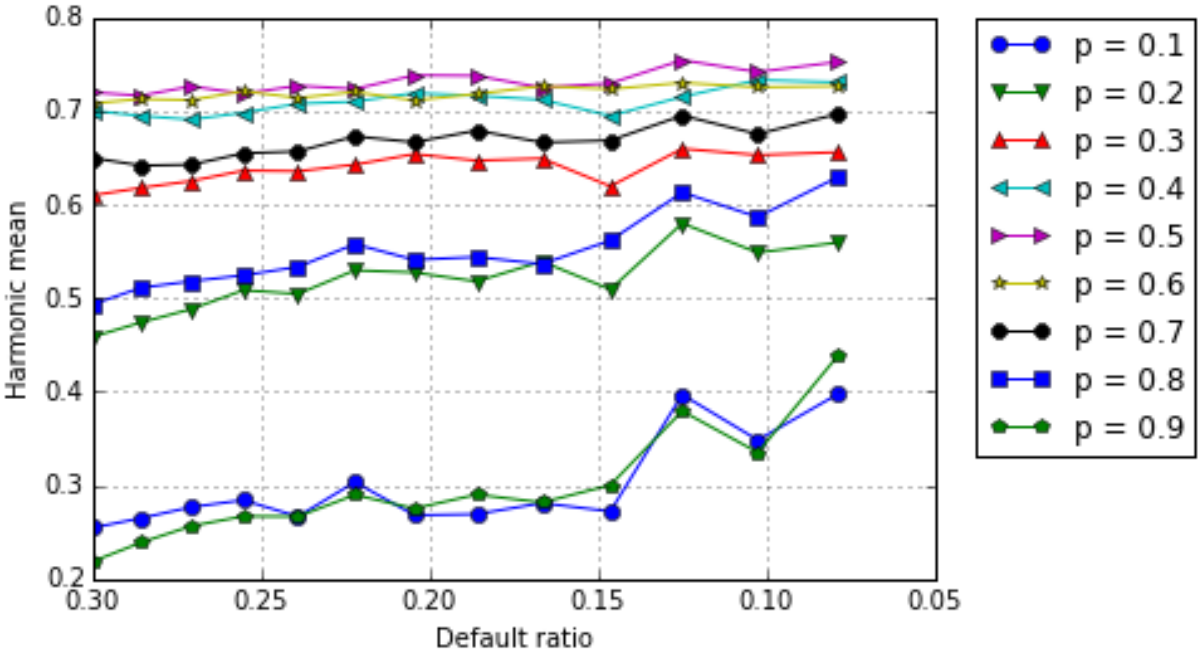

Figure 5.7: Performance of NB classifier with Bernoulli priors: German PCA 95% data

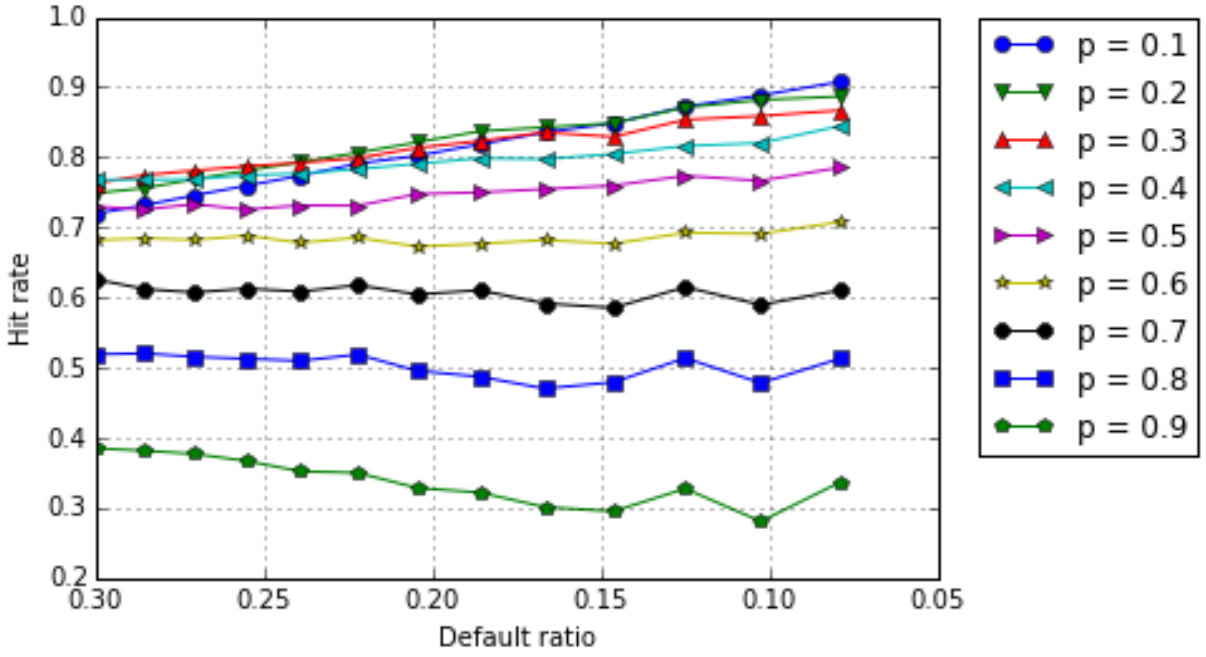

Figure 5.8: Performance of NB classifier with Bernoulli priors: German PCA 95% data

The hit rate of all considered priors for the Silverman classifier have a general upward trend regardless of whether PCA95% or z-scoring is applied. Considering the z-scored data it is observed that although the performance of the priors $p = 0.7$ and 0.8 are not optimal for greater default ratio's, the rate at which the hit rate increase is of such a nature that for the greatest class imbalance these priors outperform all the other priors. The prior $p = 0.4$ performs consistently well across the majority of default ratios. This is observed in Figure 5.9. Focusing on the hit rate of the Silverman classifier for the PCA95% data, it is seen that the priors $p = 0.3$ to $p = 0.5$ perform well over the majority of default ratios. The difference in performance between these priors are minuet. At the largest tested class imbalance the priors $p = 0.7$ and $p = 0.6$ also perform extremely well. See Figure 5.12.

For both the z-scored as well as the PCA95% data a general trend is observed in terms of the harmonic mean of the Silverman classifier. There are a few priors that remain fairly constant regardless of the default ratio. However, for both data sets there are priors that result in a fairly constant harmonic mean up and till various default ratios, after which the harmonic mean decreases. The rate at which the harmonic mean decreases, increase as the class imbalance increase. This is evident from figures 5.9 and 5.11. For the z-scored data the priors that results in decreased performance are $p = 0.1$ to $p = 0.7$, leaving the performance of the priors $p = 0.8$ and 0.9 remaining fairly constant regardless of class imbalance. The ranking in terms of performance of the priors that decrease, decrease as the value of the prior decrease. That is to say $p = 0.7$ performs the best and $p = 0.1$ performs the worst. Considering the PCA95% data the priors that result in decreasing performance after a certain

default ratio are $p = 0.1$ to $p = 0.5$, leaving the performance of the priors $p = 0.6$ to $p = 0.9$ remaining fairly constant regardless of class imbalance.
Figures 5.9 and 5.11 highlight the role the optimal Bernoulli priors play in combating the adverse effect of class imbalance.

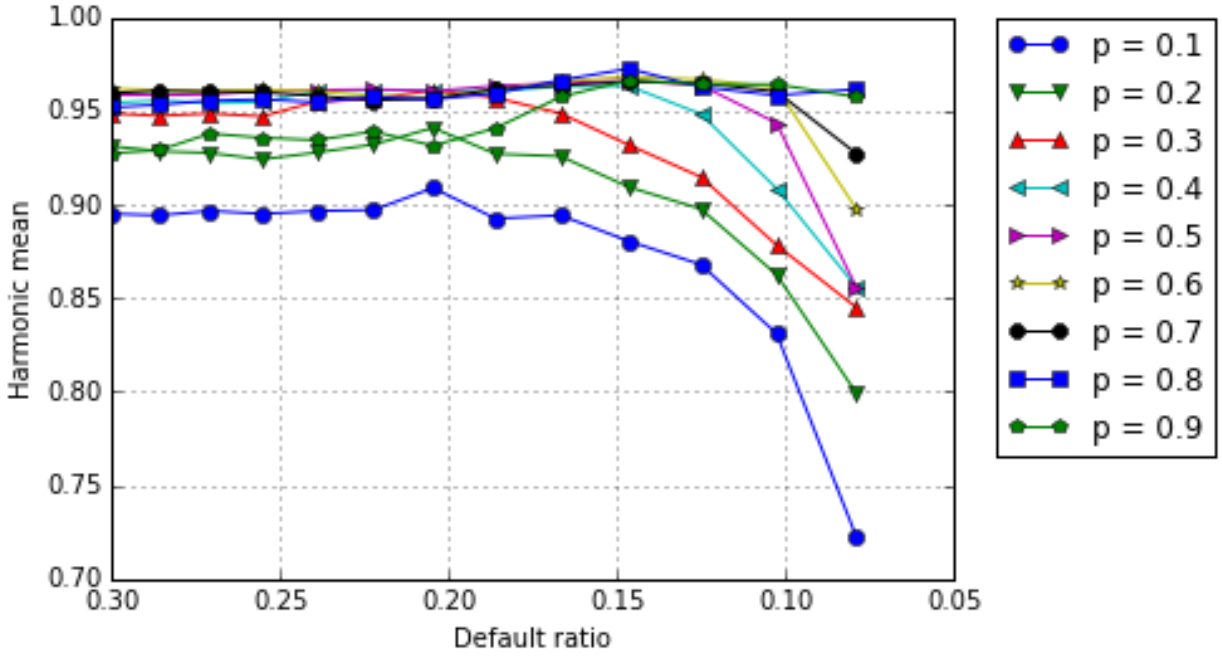

Figure 5.9: Performance of Silverman classifier with Bernoulli priors: German z-scored data

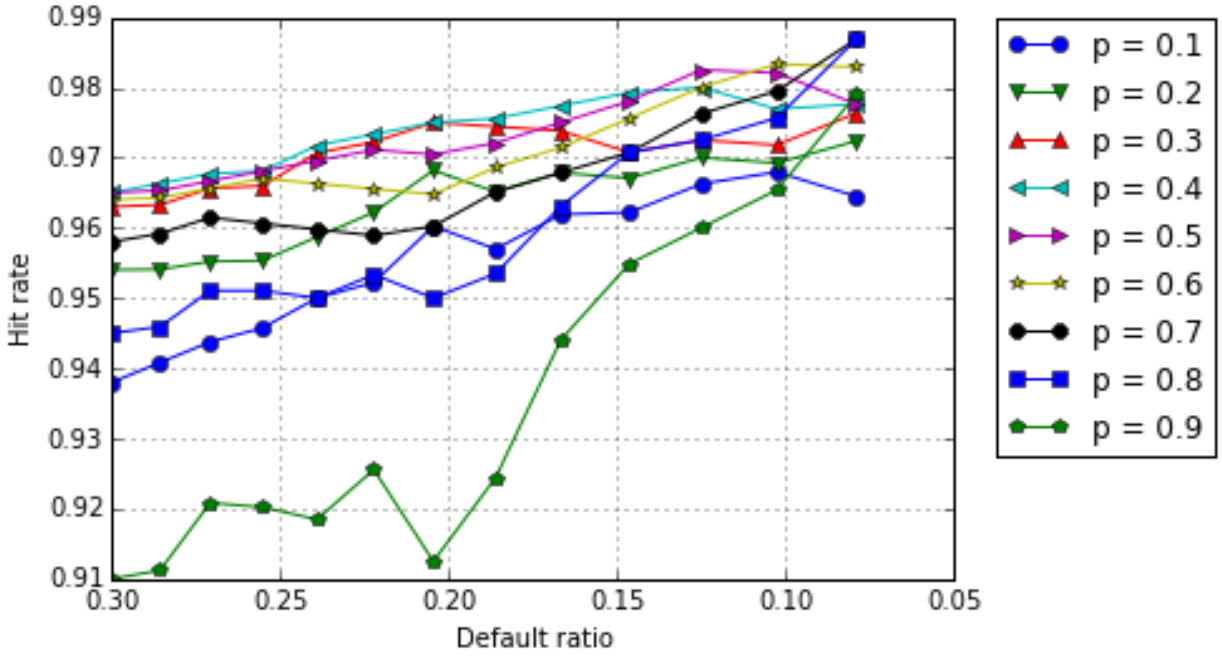

Figure 5.10: Performance of Silverman classifier with Bernoulli priors: German z-scored data

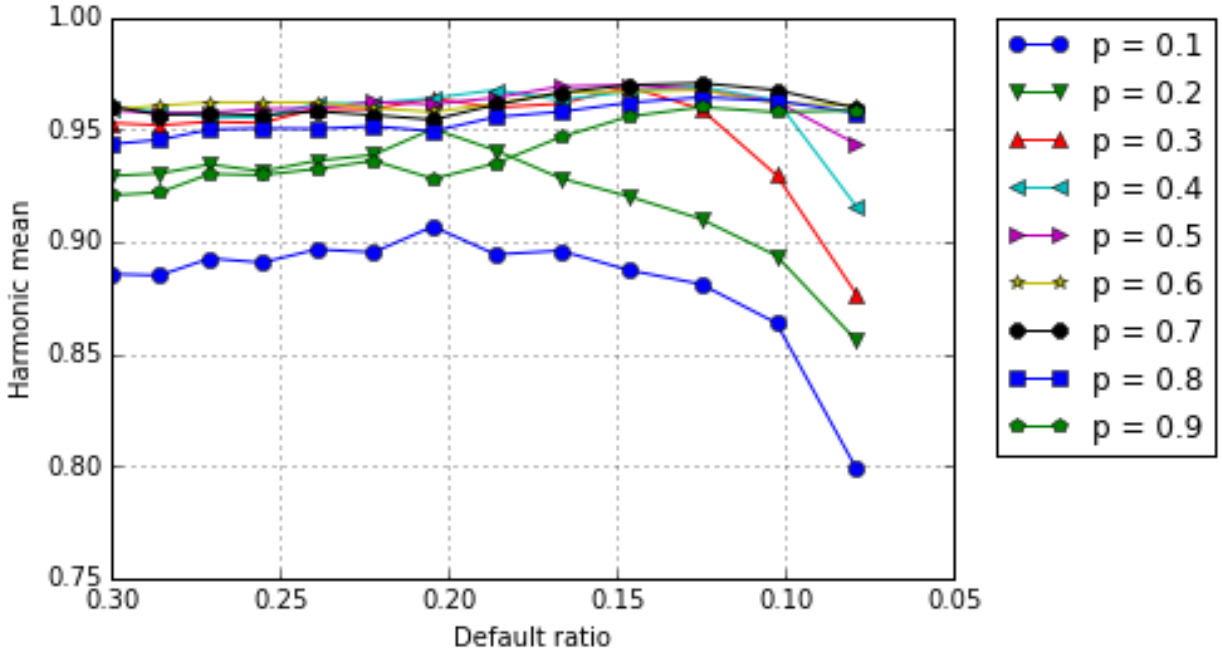

Figure 5.11: Performance of Silverman classifier with Bernoulli priors: German PCA 95% data

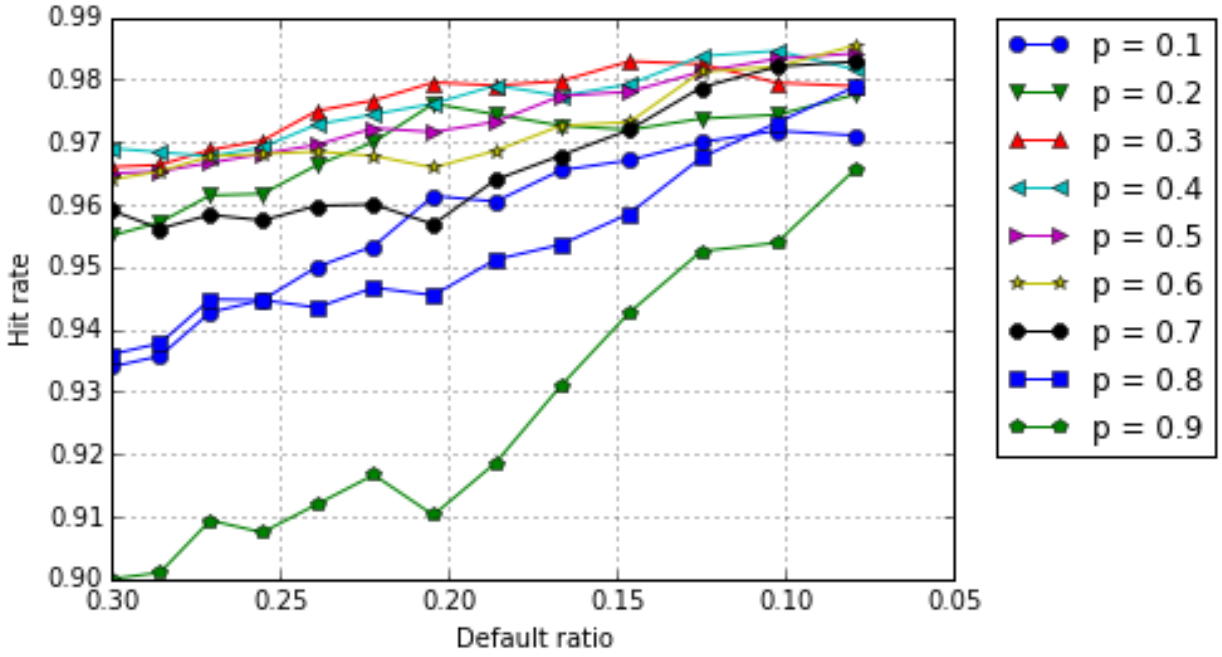

Figure 5.12: Performance of Silverman classifier with Bernoulli priors: German PCA 95% data

Consider the z-scored data. Applying the priors $p = 0.4$ to $p = 0.6$ all result in a harmonic mean above 0.95 regardless of class imbalance. The priors all have an initial upward trend, with the priors $p = 0.2$ to $p = 0.5$ as well as $p = 0.7$ showing a decrease in performance at the largest evaluated default ratio. This decrease in performance is worrisome since it might

indicate a general negative slope for larger class imbalances and hence poor performance for class imbalances that are even larger than those evaluated. The priors $p = 0.3$ and $p = 0.4$ outperform the MLE classifier when no prior is used, for the larger class imbalance. However, the prior $p = 0.4$ is the only prior that competes with the classifier when no prior is used, regardless of class imbalance. This is reflected in Figure 5.13.

Considering the hit rate of the MLE classifier applied to the z-scored data, it is satisfying to observe the priors $p = 0.3$ and $p = 0.4$ outperforming the use of no prior. The fact that these priors optimise the harmonic mean as well as the hit rate implies that they not only result in the optimal number of instances being correctly classified, but also that they result in the optimal proportion of defaulting and non-defaulting instances being correctly classified as such. The prior $p = 0.2$ also result in a competitive hit rate. The harmonic mean of the classifier with regard to this prior suggest that the proportion of defaulters it correctly classifies is suboptimal. The priors all result in a general upward trend in terms of the hit rate. It is interesting to note the slope with which the hit rate of the classifier increase for the prior $p = 0.1$. This suggests the prior is placing too much emphasis on the majority class and thereby possibly allowing the classifier to classify the majority of instances to the majority class. See Figure 5.14.

The hit rate of the MLE classifier applied to the PCA95% data exhibit a general increasing trend regardless of the prior used. The difference in hit rate between the best performing prior and the worst performing prior is at most about 0.03. The priors $p = 0.2$ to $p = 0.4$ all perform well, outperforming the MLE classifier with no prior. The prior $p = 0.9$ results in the lowest hit rate. See Figure 5.16.

The harmonic mean of the MLE classifier with priors $p = 0.1$ to $p = 0.6$ applied to the PCA95% data have a decreasing trend. However, the priors $p = 0.5$ and $p = 0.6$ increase slightly for the greatest evaluated class imbalance. Taking into account the above mentioned fact that the hit rate for these priors are increasing, suggests that the priors result in the largest proportion of instances correctly classified being in the non-defaulting class. This implies that too much weight is assigned to the majority class. The priors $p = 0.8$ and $p = 0.9$ both perform well; outperforming the use of no prior. The prior $p = 0.9$ outperforms all other evaluated priors for all default ratios less than 0.2391, whereas the prior $p = 0.8$ outperforms all other evaluated priors for all default ratios greater than 0.2391. The performance of the prior $p = 0.9$ remains fairly constant for larger class imbalances, whereas the prior $p = 0.8$ decrease for the smallest evaluated default ratio.

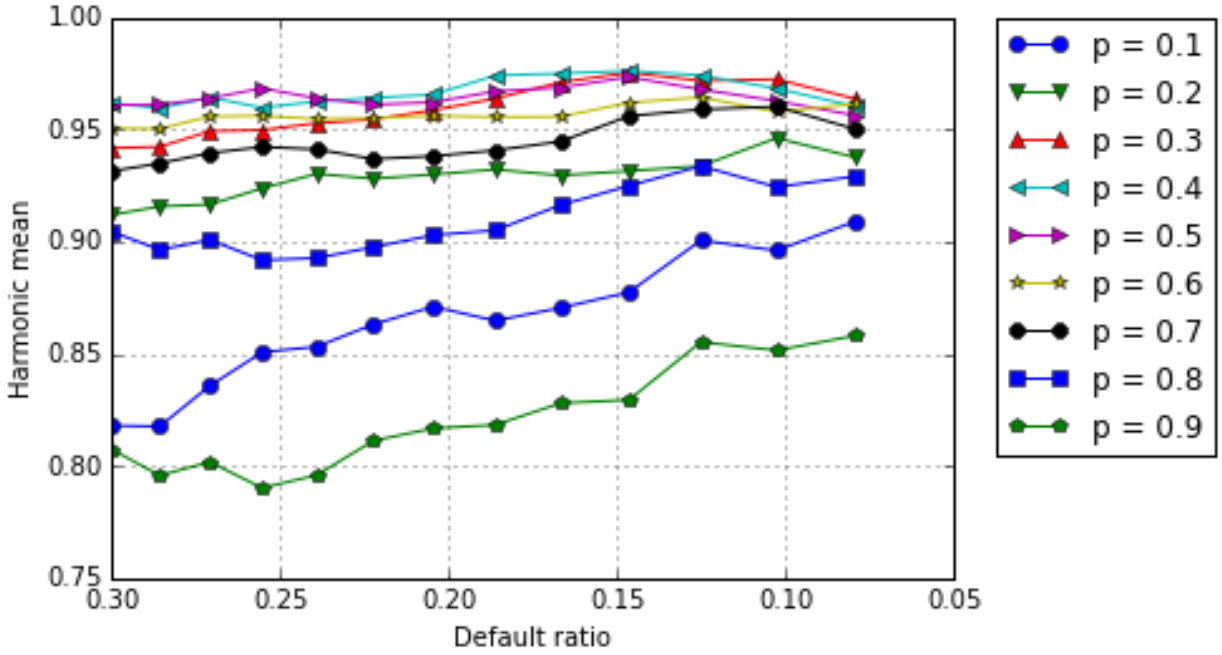

Figure 5.13: Performance of MLE classifier with Bernoulli priors: German z-scored data

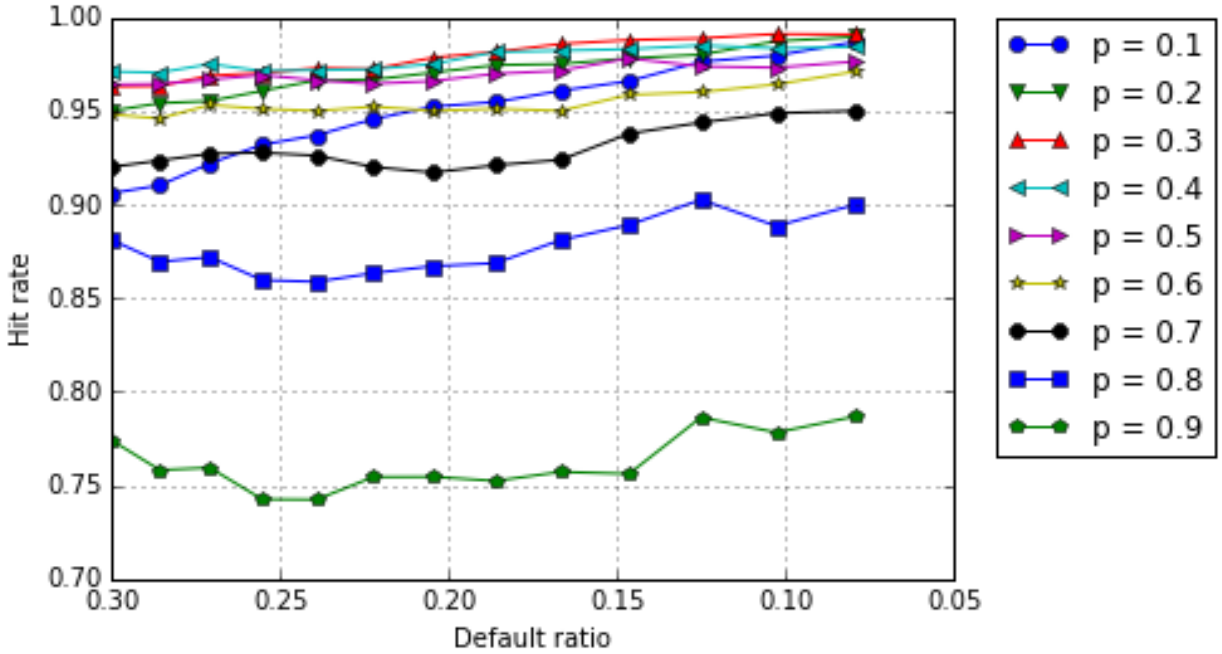

Figure 5.14: Performance of MLE classifier with Bernoulli priors: German z-scored data

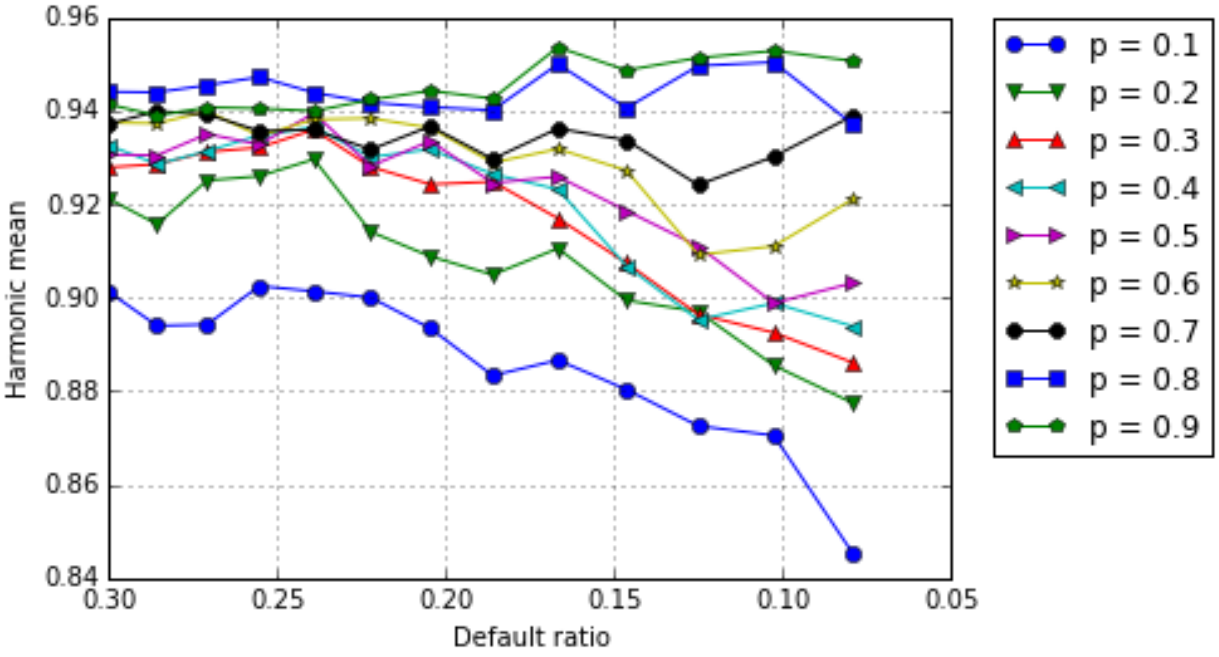

Figure 5.15: Performance of MLE classifier with Bernoulli priors: German PCA 95% data

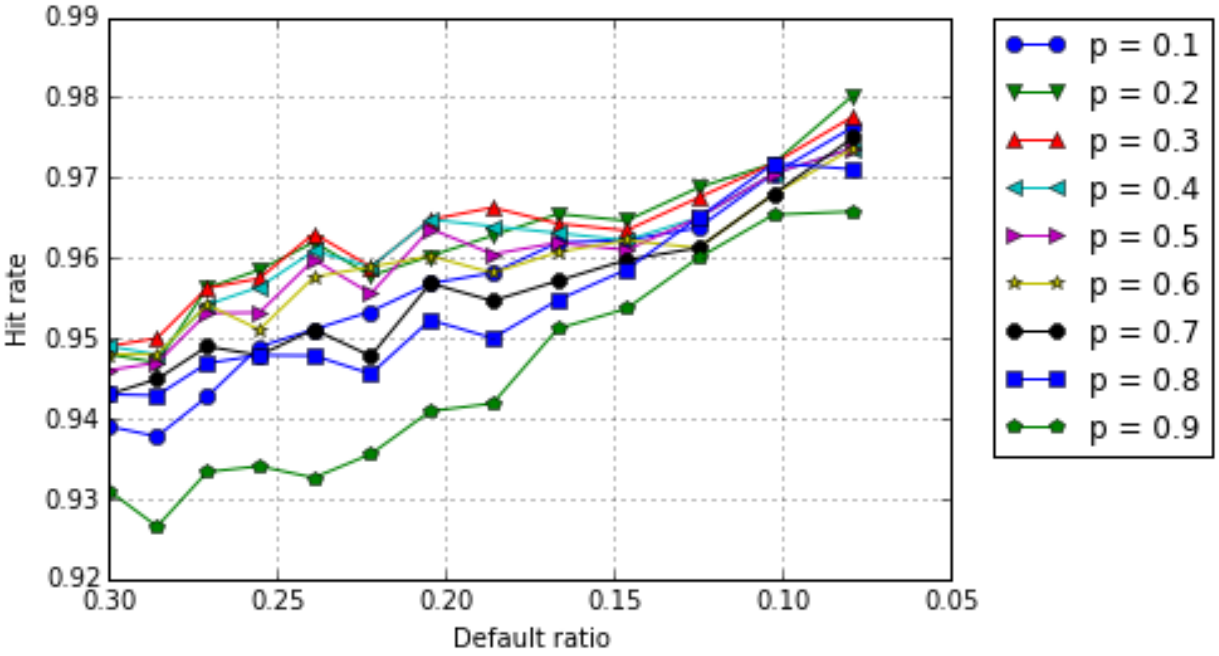

Figure 5.16: Performance of MLE classifier with Bernoulli priors: German PCA 95% data

Australian data

Considering figures 5.17, 5.18, 5.19 and 5.20 it is seen that the performance of the Gaussian classifier decrease, for the majority of default ratios, as the value of the prior parameter p decrease. An exception to this may be the prior $p = 0.2$ which has a higher harmonic mean

than the prior $p = 0.1$ at higher default ratios. Note that this is only the true for the PCA 95% data. It follows that for the z-scored data a prior parameter of $p = 0.1$ is recommended to optimise the harmonic mean as well as the hit rate. However, for the PCA 95% data a prior parameter of $p = 0.1$, or $p = 0.2$ for greater class imbalances, may be considered appropriate in order to optimise the harmonic mean. The hit rate of the PCA95% data is optimised using a prior of $p = 0.1$.

The hit rate of the Gaussian classifier decrease up and till a default ratio of about 0.3, after which it has general increasing trend. This is true for the majority of priors, with the only exceptions being the top performing priors. The top performing priors evaluated using the PCA 95% data tend to remain relatively constant, before increasing slightly. The harmonic mean trend of the Gaussian classifier tend to have a general upward trend, regardless of the prior used and whether the z-scored or PCA95% data is considered. This is surprising as it indicates the classifier performs better when there are less data available in the defaulting class.

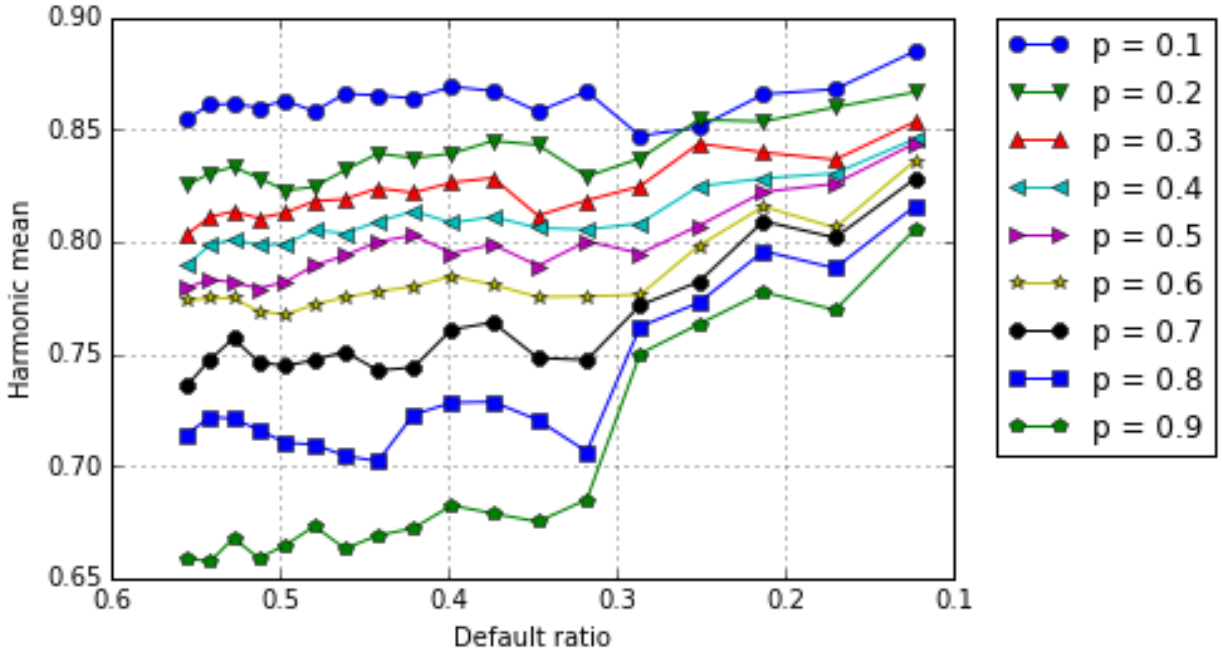

Figure 5.17: Performance of Gaussian classifier with Bernoulli priors: Australian z-scored data

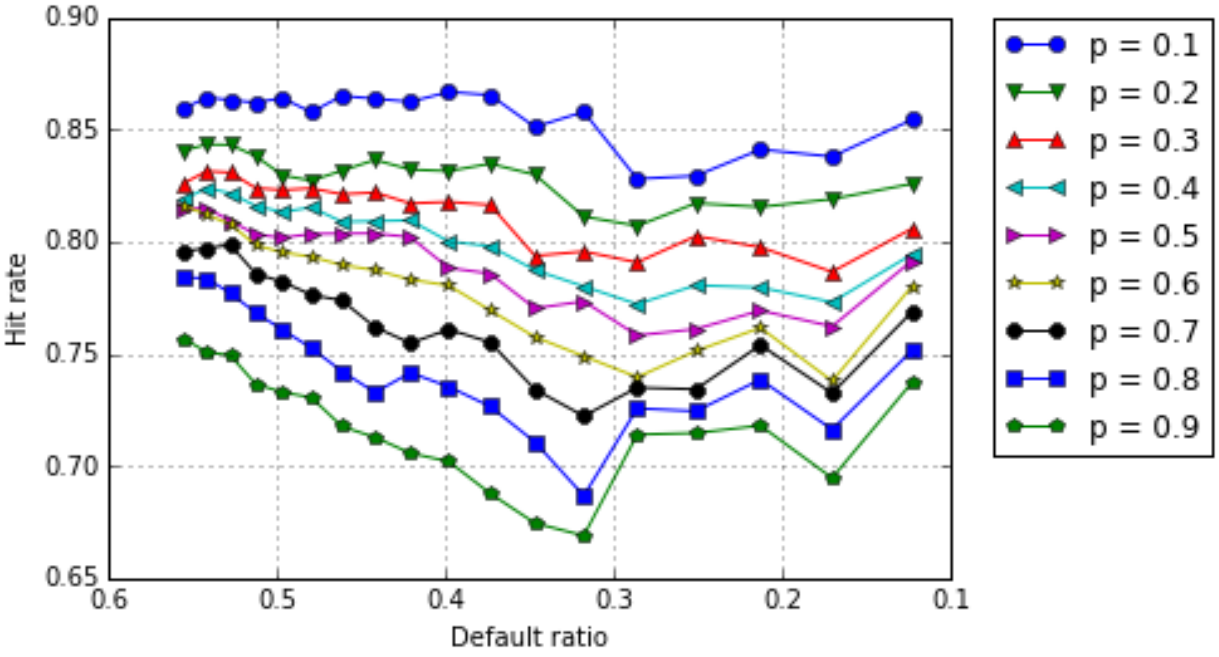

Figure 5.18: Performance of Gaussian classifier with Bernoulli priors: Australian z-scored data

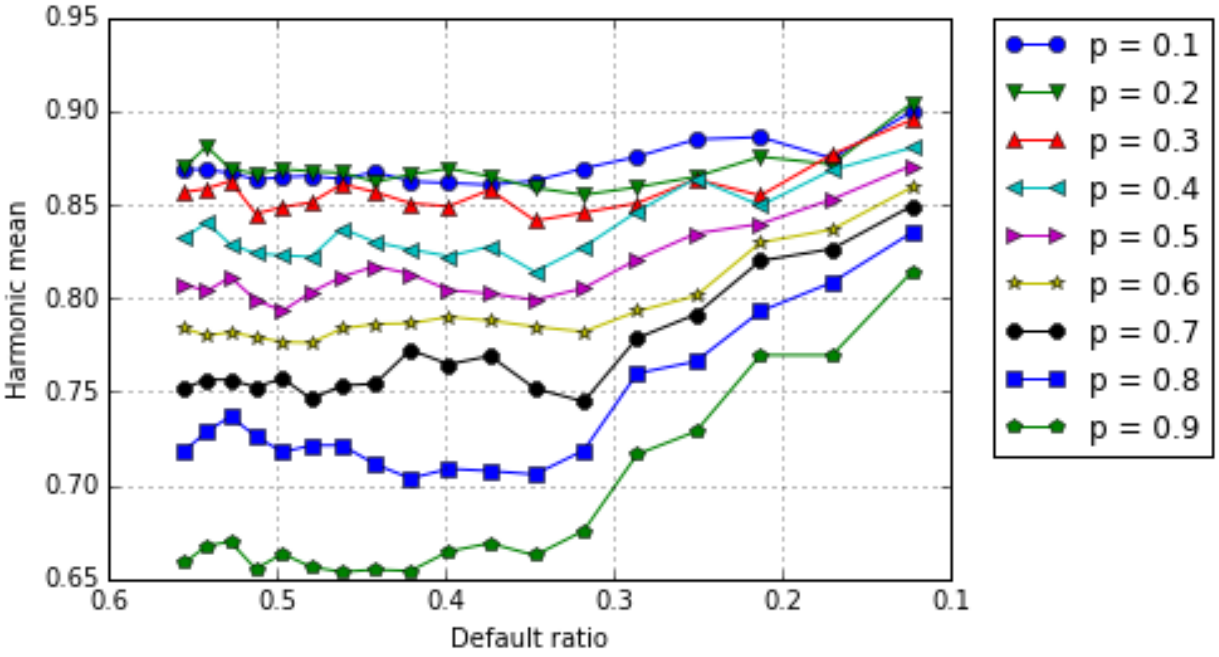

Figure 5.19: Performance of Gaussian classifier with Bernoulli priors: Australian PCA 95% data

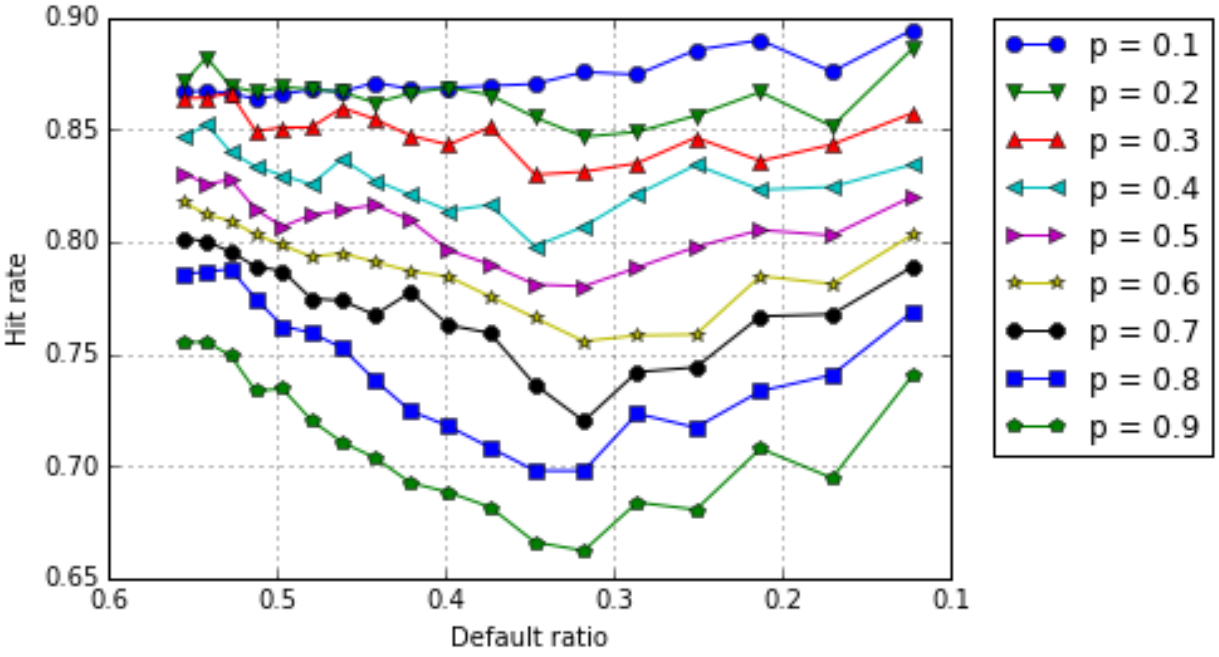

Figure 5.20: Performance of Gaussian classifier with Bernoulli priors: Australian PCA 95% data

The performance of the NB classifier decrease as the prior parameter decrease for the z-scored data. This is not the case for the PCA 95% data. This implies that a prior of $p = 0.1$ results in the optimal performance in terms of the hit rate as well as the harmonic mean for the z-scored data. The harmonic mean of the NB classifier applied to the z-scored data remains fairly constant up to a default ratio of 0.3178 after which it increases. The only prior that does not result in the harmonic mean behaving in the before mentioned fashion is $p = 0.1$. This prior remains fairly constant regardless of the class imbalance. This is reflected in Figure 5.21.

The hit rate of NB classifier for the z-scored data generally decrease up to a default ratio of 0.3178, after which it either increase or remain constant. A final increase is noticeable between the default ratios 0.1026 and 0.0789. Regardless of the prior used the hit rate at the starting default ratio is lower than the final default ratio of 0.0789. The decrease in hit rate and simultaneous increase in harmonic mean may be attributed to an increase in the proportion of defaulters correctly classified as such. See Figure 5.22.

Observing the hit rate of the NB classifier applied to the PCA95% data it is evident that the priors $p = 0.7$ to $p = 0.9$ initially decrease after which at a default ratio of 0.3178 it increases. The priors $p = 0.6$ and $p = 0.5$ follow a similar pattern. However, the slope at which the hit rate initially decrease and finally increase is not as great as that of the priors $p = 0.7$ to $p = 0.9$. The priors $p = 0.1$ to $p = 0.4$ do not exhibit the same initial decreasing behaviour. In fact as the value of the prior decrease from $p = 0.4$ to $p = 0.1$ the slope with which it increase, increases. However, the hit rate resulting from the use of the prior $p = 0.1$

is considerably lower than that of the prior $p = 0.2$ when the data set is balanced. The result is that the prior $p = 0.1$ only results in optimal performance for the greatest evaluated class imbalance. The prior $p = 0.2$ performs consistently well for lower default ratios. These results are clear from Figure 5.24.

The performance of the priors $p = 03$ to $p = 0.6$ applied to the NB classifier remain fairly constant with a resulting harmonic mean of 0.85. These priors outperform the other evaluated priors. It is interesting to note the dip in performance at a default ratio of 0.3178, with larger dips for the priors $p = 0.7$ and $p = 0.8$. See Figure 5.23.

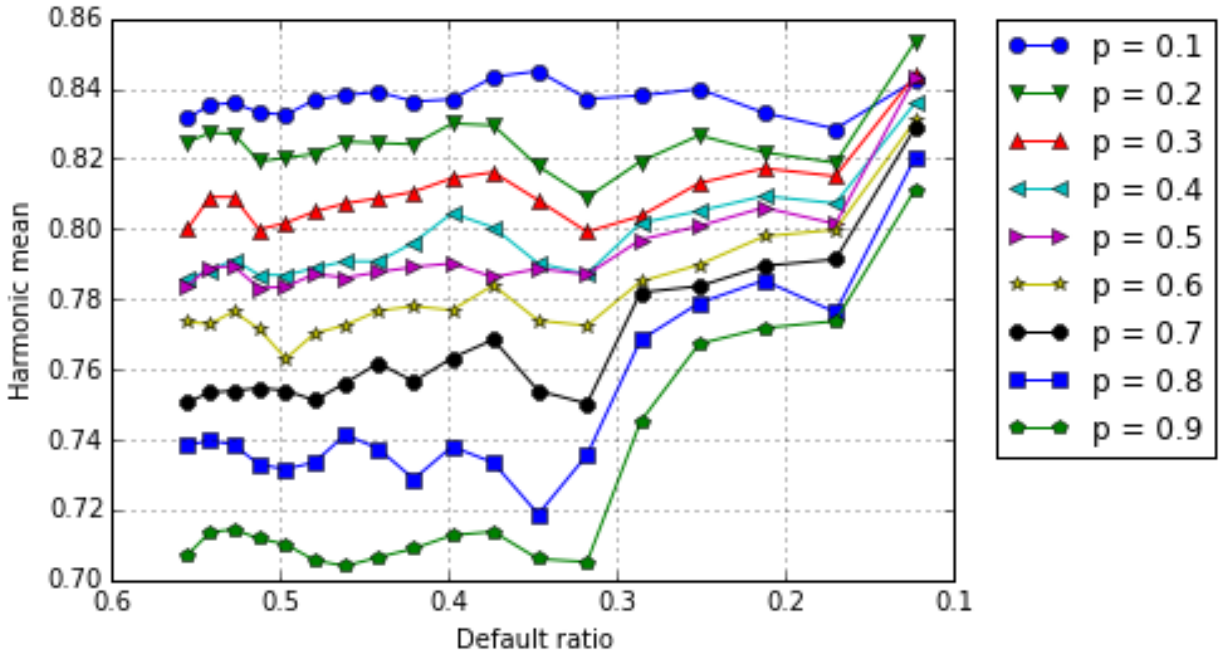

Figure 5.21: Performance of NB classifier with Bernoulli priors: Australian z-scored data

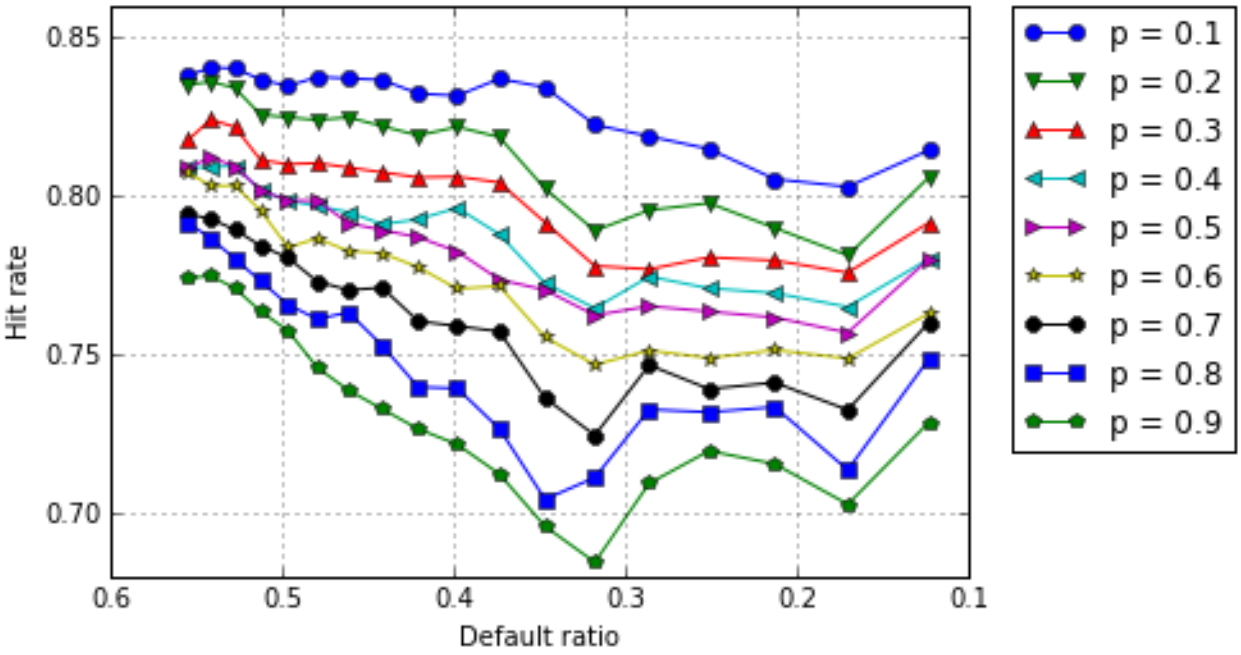

Figure 5.22: Performance of NB classifier with Bernoulli priors: Australian z-scored data

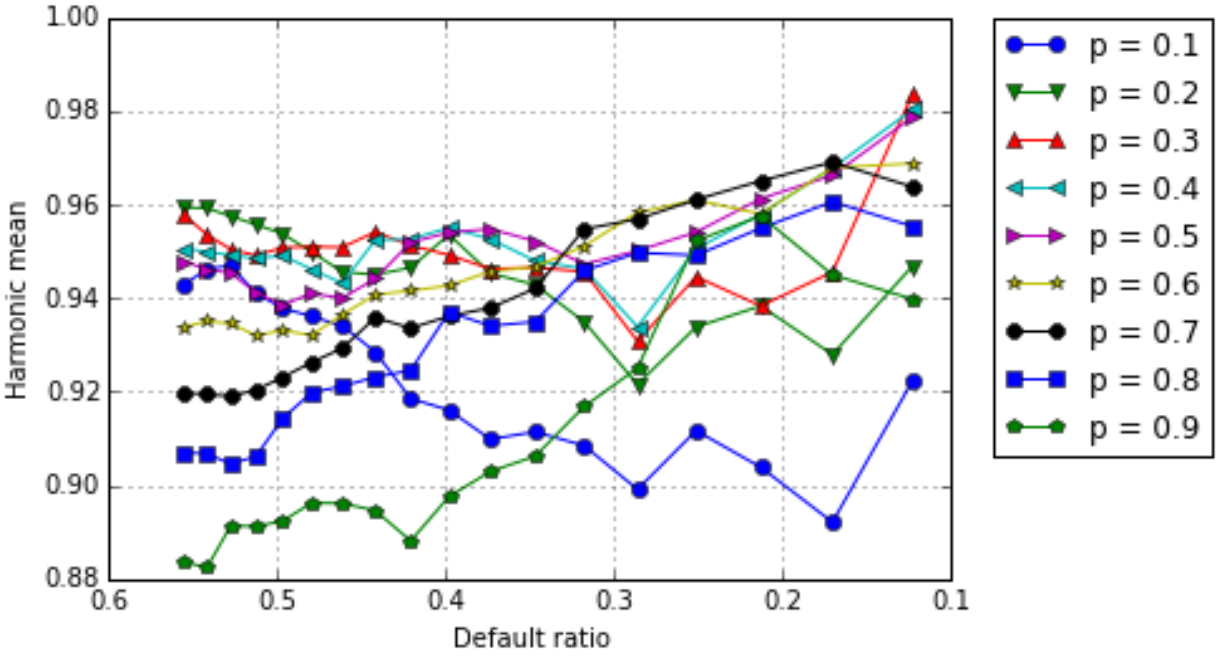

Figure 5.23: Performance of NB classifier with Bernoulli priors: Australian PCA 95% data

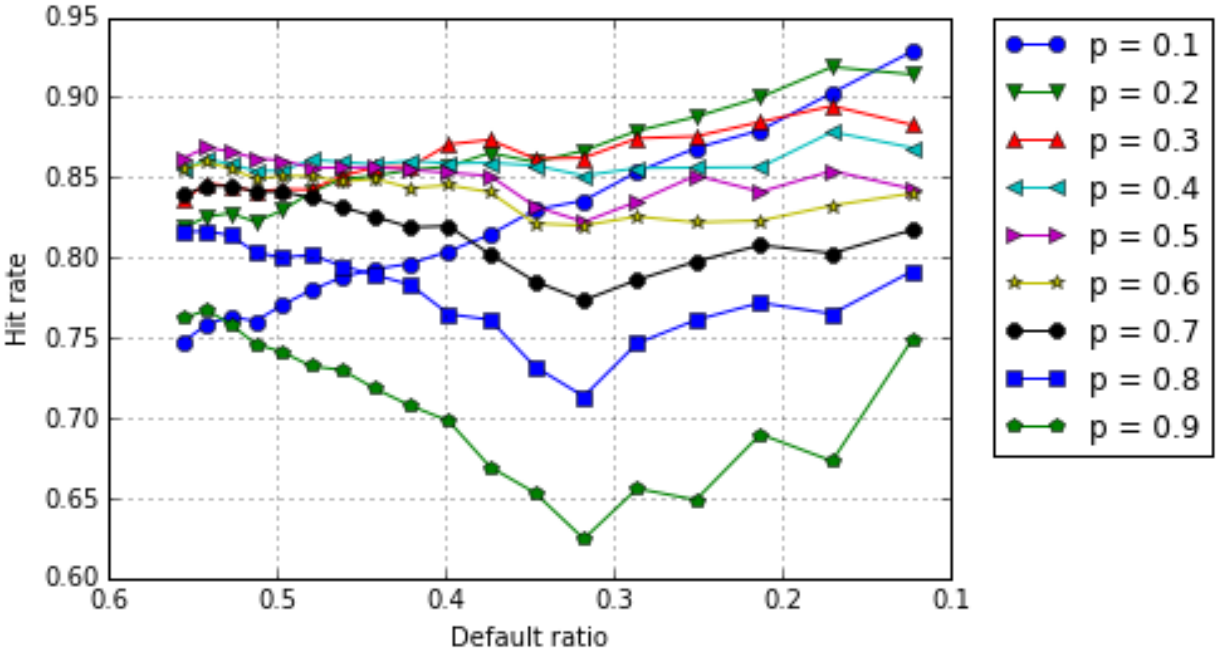

Figure 5.24: Performance of NB classifier with Bernoulli priors: Australian PCA 95% data

Considering the harmonic mean of the Silverman classifier for the z-scored data it is seen that the priors $p = 0.1$ to $p = 0.4$ remain fairly constant. These priors perform well in comparison to the other investigated priors. The prior $p = 0.2$ outperforms all other priors over the majority of default ratios. It is interesting to note a jump that occurs in performance at a default ratio of 0.286. The jump is more discernible for the poor performing priors. See Figure 5.25.

The hit rate of the Silverman classifier with priors $p = 0.2$ and 0.3, applied to the z-scored data, perform well and remain fairly constant over the various default ratios. A slight upward trend in hit rate is visible for the prior $p = 0.1$. For greater class imbalances the prior $p = 0.1$ results in the highest hit rate. The poor performing priors have a general decreasing trend up to a default ratio of 0.286, at which a jump in performance occurs. This is reflected in Figure 5.26.

The priors $p = 0.2$ to $p = 0.5$ remain relatively constant in terms of harmonic mean with respect to the PCA95% data. These priors outperform the other considered priors for the majority of default ratios. Exceptions to this are the priors $p = 0.6$ and $p = 0.7$ which perform well for larger class imbalances. As with the z-scored data a jump in harmonic mean at a default ratio of 0.286 is evident for the prior $p = 0.9$. It is also interesting to note that whereas the majority of priors result in an overall upward trend, the prior $p = 0.1$ is decreasing from the default ratio 0.2512 onward. See Figure 5.27.

Similar to the z-scored data the priors $p = 0.2$ to $p = 0.4$ used in conjunction with the Silverman classifier remain fairly constant in terms of hit rate for the PCA 95% data. In

contrast to the hit rate for the z-scored data the prior $p = 0.1$ has a lower initial value and increase significantly as the class imbalance increase. For greater class imbalances the prior $p = 0.2$ maximises the hit rate. Just like a jump in the harmonic mean is evident at a default ratio of 0.286 for the prior $p = 0.9$, a jump in the hit rate is evident at the same default ratio for the prior $p = 0.9$.

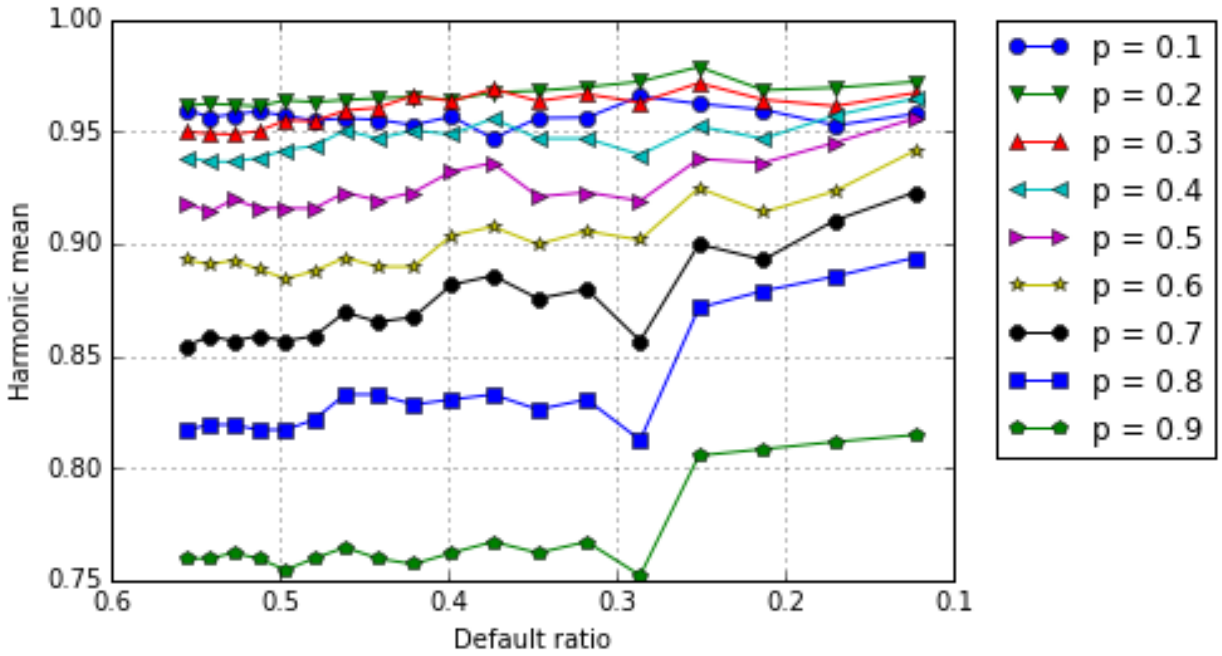

Figure 5.25: Performance of Silverman classifier with Bernoulli priors: Australian z-scored data

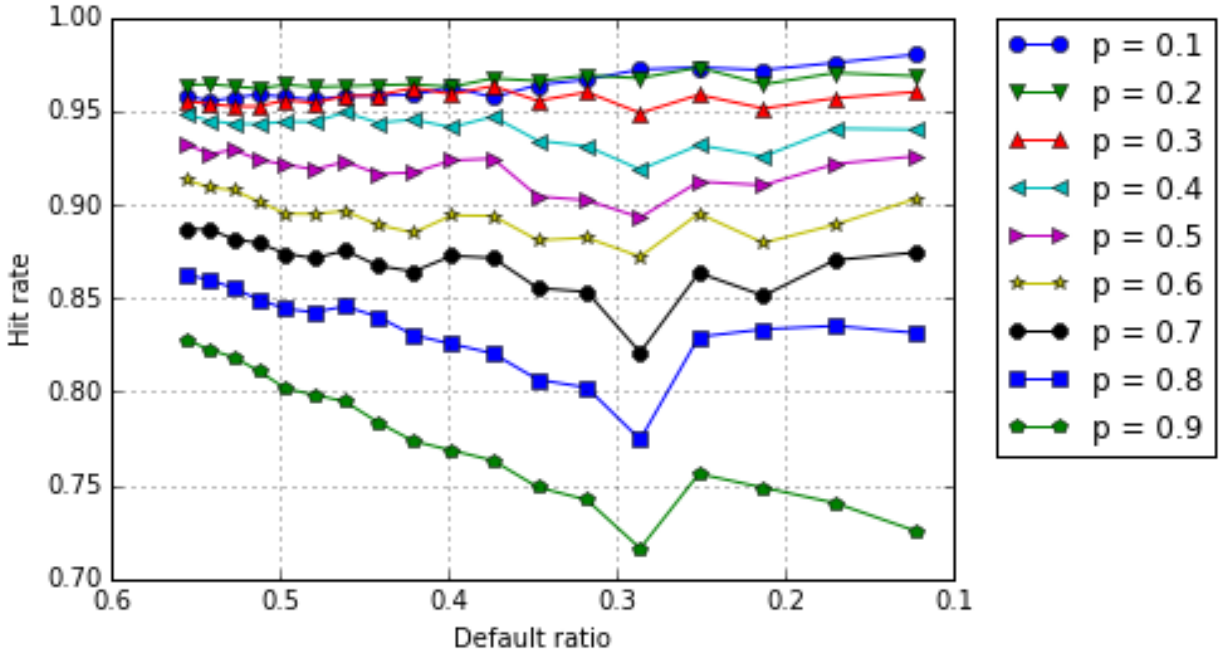

Figure 5.26: Performance of Silverman classifier with Bernoulli priors: Australian z-scored data

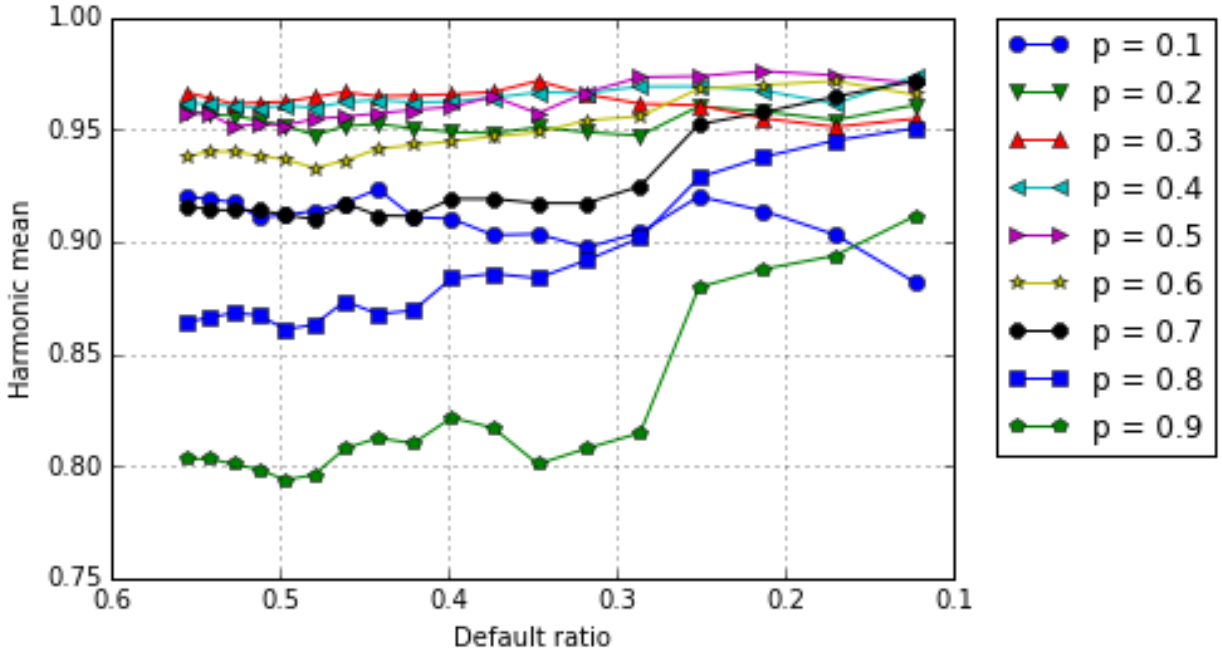

Figure 5.27: Performance of Silverman classifier with Bernoulli priors: Australian PCA 95% data

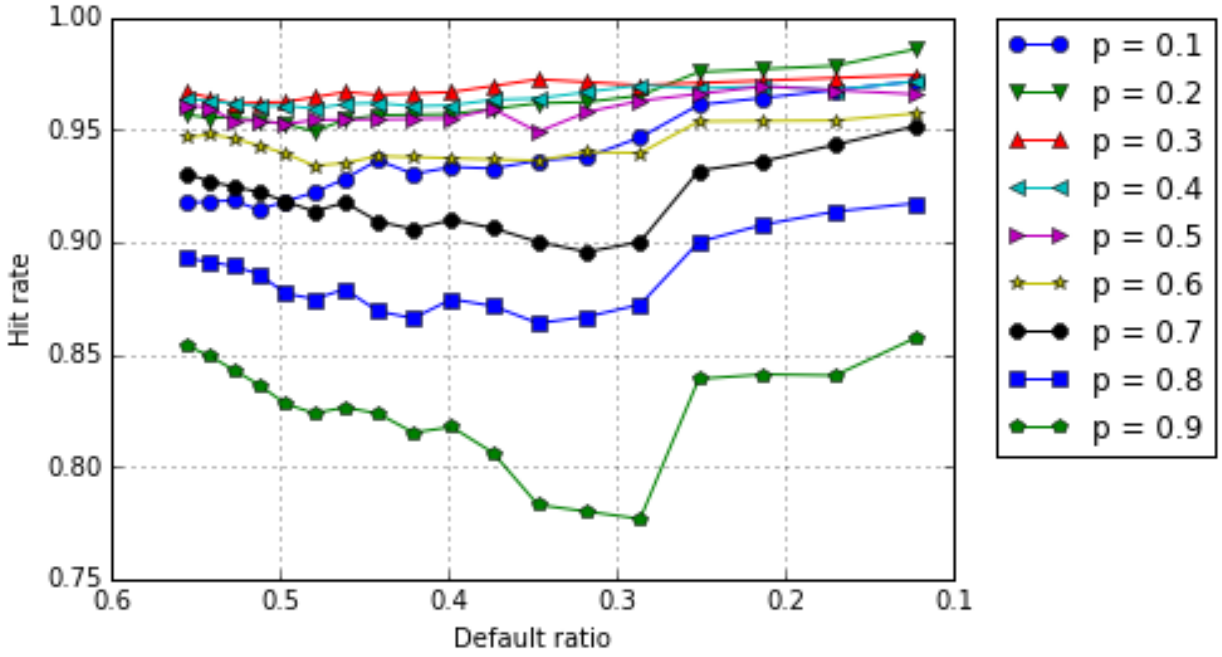

Figure 5.28: Performance of Silverman classifier with Bernoulli priors: Australian PCA 95% data

Considering the z-scored data the ranking (in terms of performance) of the priors for the MLE classifier are similar regardless of whether the performance is measured in terms of the hit rate or the harmonic mean. The only exception being the prior $p = 0.1$. This prior results in a general upward trend in terms of the hit rate, with such a steepness that it

outperforms all other priors for the greatest evaluated class imbalance. The fact that the prior optimise the hit rate but not the harmonic mean at large class imbalances indicate that too much weight is placed on the majority class. This results in the number of instances in the majority class being correctly classified and hence a high hit rate. However, a small proportion of instances in the minority class are correctly classified for which the harmonic mean penalise the classifier. The harmonic mean as well as the hit rate is optimised for the majority of evaluated default ratios by applying the prior $p = 0.2$ to the MLE classifier. This is of extreme value since it implies that the prior not only result in the most instances being correctly classified, but also that this prior best assists the classifier in classifying true positive as well as true negative instances. It is worth noting that the performance of the MLE classifier decrease as the prior increase, with the exception of the prior $p = 0.1$. This is reflected in figures 5.29 and 5.30

The impact of a prior on the performance of the MLE classifier is much more erratic for the PCA 95% data than that of the z-scored data, as evident from figures 5.31 and 5.32. The priors $p = 0.9$ to $p = 0.5$ in general have increasing trends for the PCA 95% data. However, selecting a prior to optimise the harmonic mean would greatly depend on the default ratio present in the data. The priors $p = 0.2$ to $p = 0.4$ result in higher hit rates, in comparison to the other evaluated classifiers, for the PCA 95% data. The priors $p = 0.6$ to $p = 0.9$ result in a decrease in hit rate for some of the smaller default ratios evaluated, whereas the priors $p = 0.1$ to $p = 0.5$ maintain an increasing trend over these default ratios. In general a prior of $p = 0.4$ would perform sufficiently well across the various default ratios.

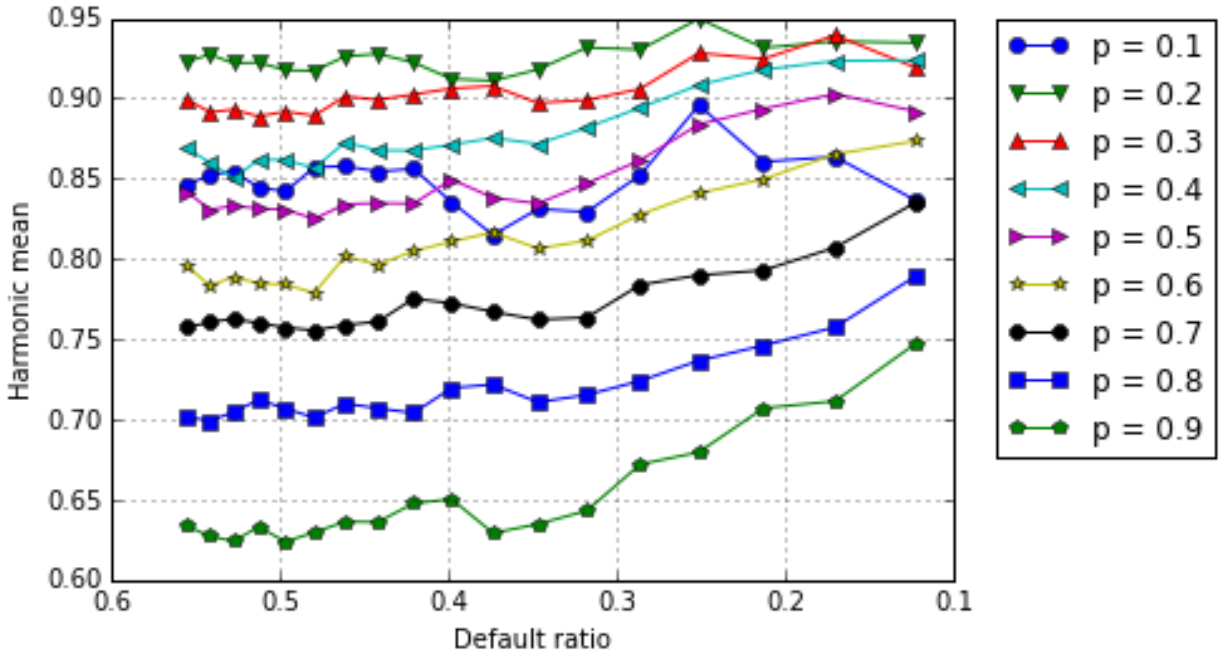

Figure 5.29: Performance of MLE classifier with Bernoulli priors: Australian z-scored data

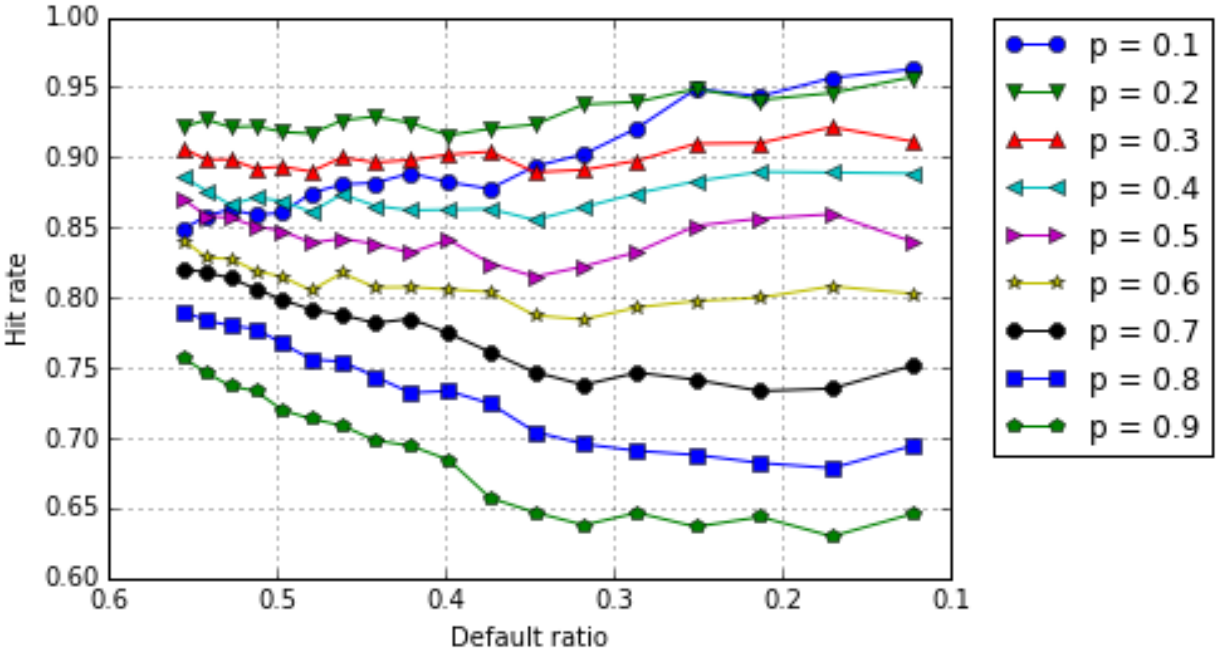

Figure 5.30: Performance of MLE classifier with Bernoulli priors: Australian z-scored data

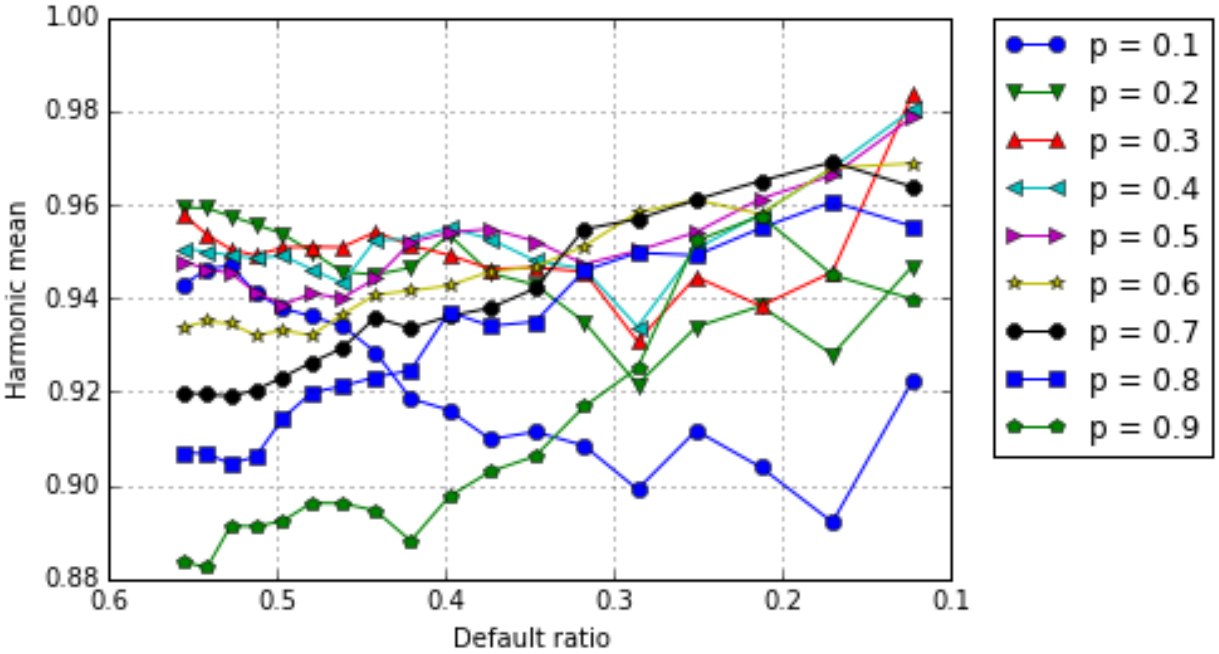

Figure 5.31: Performance of MLE classifier with Bernoulli priors: Australian PCA 95% data

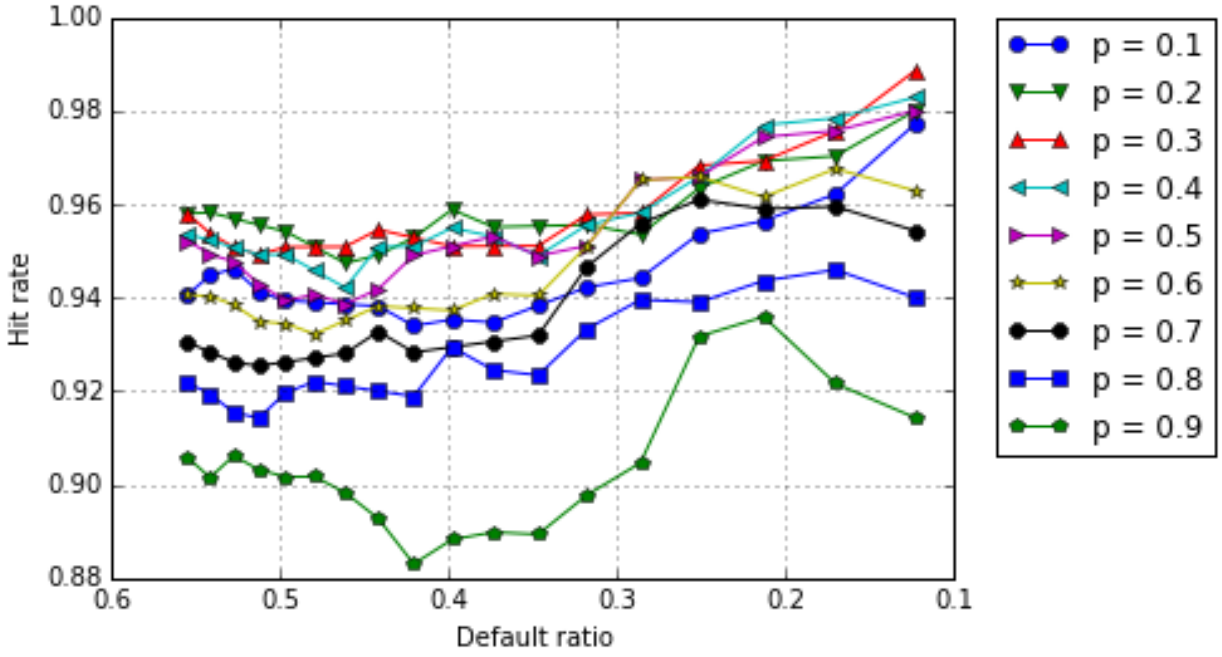

Figure 5.32: Performance of MLE classifier with Bernoulli priors: Australian PCA 95% data

5.2.2 Comparison of various priors

The following subsection compares the performance of Bernoulli priors, no priors and frequentist priors. The performance of the frequentist prior is indicated by the "Frequentist prior" column, the performance of the use of no prior is indicated by the "$p = 0.5$" column and the remaining column indicates the performance of the Bernoulli prior.

German data

Viewing Table 5.1 no single prior used in conjunction with the Gaussian classifier results in the optimal harmonic mean across all class-imbalances. However, it would appear that the frequentist prior rarely results in the optimal harmonic mean. This suggests that applying a Bernoulli prior with a finer tuned prior parameter, say to three or four decimal places, may be more appropriate. This is also an indication that a fixed Bernoulli parameter prior used regardless of class-imbalance is inefficient. Instead the prior parameter should be dependent on the class-imbalance.

The performance, measured in terms of the hit rate, suggests that the Bernoulli prior outperforms the frequentist prior as well as the use of no prior for the majority of class-imbalances. It is also clear that the frequentist prior performs better than the classifier used without a prior for all evaluated class-imbalances. The fact that the frequentist prior outperforms the Bernoulli prior for greater class imbalances suggest that the Bernoulli prior parameter may depend on the class-imbalance. See Table 5.2. The difference in conclusions made between

the performance of the harmonic mean and that of the hit rate emphasise that the user should decided whether it is more important to optimise the hit rate or the harmonic mean. It is evident that the Gaussian classifier performs better, as indicated by the bold values, on the z-scored data than the PCA95% data. This is due to the fact that by applying PCA to the data the correlation is removed between features and some variation is lost due to the dimensionality reduction. However, the Gaussian classifier is able to take into account covariances, and hence correlations, between features. The classifier therefore performs better on the z-scored data in which the features are correlated and all of the original variation is present in the data.

	PCA 95%			Z-Scored		
Default ratio	**Frequentist prior**	$p = 0.4$	$p = 0.5$	**Frequentist prior**	$p = 0.4$	$p = 0.5$
0,3000	0,6749	0,7328	0,7287	0,7393	**0,7709**	0,7610
0,2857	0,6967	0,7398	0,7302	**0,7705**	0,7546	0,7426
0,2708	0,7149	0,7493	0,7417	0,7474	**0,7646**	0,7580
0,2553	0,7180	0,7429	0,7450	0,7609	0,7586	**0,7672**
0,2391	0,7690	0,7531	0,7465	0,7592	**0,7758**	0,7595
0,2222	0,7703	0,7480	0,7397	**0,7799**	0,7679	0,7603
0,2045	0,7673	0,7499	0,7452	**0,7875**	0,7673	0,7702
0,1860	0,7420	0,7561	0,7506	0,7660	**0,7701**	0,7553
0,1667	0,7392	0,7743	0,7785	0,7605	0,7921	**0,7944**
0,1463	0,7344	0,8036	0,8004	0,7482	**0,8133**	0,8101
0,1250	0,7292	0,8088	0,8243	0,7561	0,8209	**0,8558**
0,1026	0,7319	0,8362	0,8282	0,7567	**0,8469**	0,8451
0,0789	0,7310	0,8374	0,8619	0,7604	0,8524	**0,8712**

Table 5.1: Harmonic mean of the Gaussian classifier: German data

	PCA 95%			Z-Scored		
Default ratio	**Frequentist prior**	$p = 0.1$	$p = 0.5$	**Frequentist prior**	$p = 0.1$	$p = 0.5$
0,3000	0,7550	0,7890	0,7090	0,7800	**0,7990**	0,7430
0,2857	0,7612	0,7990	0,7061	0,7857	**0,8061**	0,7235
0,2708	0,7781	0,8063	0,7188	0,7948	**0,8177**	0,7406
0,2553	0,7904	0,8160	0,7234	0,8011	**0,8287**	0,7564
0,2391	0,7935	**0,8326**	0,7207	0,8087	0,8315	0,7413
0,2222	0,7956	0,8322	0,7078	0,8133	**0,8389**	0,7433
0,2045	0,8034	0,8455	0,7102	0,8216	**0,8545**	0,7443
0,1860	0,8209	0,8535	0,7140	0,8337	**0,8581**	0,7372
0,1667	0,8524	**0,8607**	0,7464	0,8488	0,8595	0,7738
0,1463	0,8659	**0,8732**	0,7841	0,8659	0,8720	0,8000
0,1250	0,8875	**0,8950**	0,8000	0,8825	0,8925	0,8388
0,1026	0,9090	0,9051	0,8308	0,9154	**0,9179**	0,8513
0,0789	0,9368	0,9316	0,8579	**0,9342**	0,9316	0,8750

Table 5.2: Hit rate of the Gaussian classifier: German data

The NB classifier performs better on the PCA95% data than the z-scored data. This can be explained by the fact that there is no correlation present in the PCA95% data. Since the NB classifier does not take account of the covariances, and hence correlations, it follows intuitively that the classifier will perform better on the PCA95% data than the z-scored data.
The NB classifier performs best, in terms of the harmonic mean, on the PCA95% data when no prior is used. In fact from Table 5.3 it is evident that the frequentist prior actually reduce the harmonic mean of the classifier regardless of the class imbalance. Considering the z-scored data; the classifier has the optimal harmonic mean for the majority of default ratios when no prior is used, with the exception of a few default ratios for which the frequentist prior results in the optimal harmonic mean. The largest class-imbalance is the only case in which the Bernoulli prior of $p = 0.6$ results in the optimal harmonic mean. It is clear that for the z-scored data the optimal prior is dependent on the class-imbalance. However, since the frequentist prior, of which the value of the prior depends on the class-imbalance, does not result in the optimal harmonic mean the relationship between the prior and the class-imbalance differ from that given by the frequentist prior.
The hit rate of the PCA95% data is optimised by the Bernoulli prior or the frequentist prior depending on the default ratio. This suggests that a Bernoulli prior with prior parameter dependent on the class-imbalance will result in the optimal performance. Considering the z-scored data, the hit rate is optimised using a Bernoulli prior of $p = 0.1$ for the majority of class-imbalances. The exception being the smallest default ratio, which is optimised by the frequentist prior. It is clear that the use of a prior can improve the number of instances correctly classified. The reader is referred to Table 5.4.
As with the Gaussian classifier there is a difference in the conclusions made for the performance evaluated in terms of the harmonic mean versus that of the hit rate; once again emphasising the choice the user of the model have to make between optimising the hit rate versus optimising the harmonic mean.

	PCA 95%		Z-Scored		
Default ratio	**Frequentist prior**	$p = 0.5$	**Frequentist prior**	$p = 0.5$	$p = 0.6$
0,3000	0,6097	**0,7197**	0,7073	0,7044	0,6909
0,2857	0,5901	**0,7162**	0,6906	0,7043	0,6877
0,2708	0,5824	**0,7262**	0,6935	0,7037	0,6915
0,2553	0,5859	**0,7182**	0,6986	0,7021	0,6932
0,2391	0,5526	**0,7268**	0,7032	0,6989	0,6876
0,2222	0,5289	**0,7233**	0,7012	0,6994	0,6894
0,2045	0,4978	**0,7379**	0,7069	0,6939	0,6822
0,1860	0,4635	**0,7374**	0,6981	0,7010	0,6893
0,1667	0,4388	**0,7248**	0,6830	0,7005	0,6922
0,1463	0,4267	**0,7294**	0,6369	0,7108	0,7019
0,1250	0,4450	**0,7542**	0,6477	0,7403	0,7291
0,1026	0,3317	**0,7415**	0,5663	0,7337	0,7164
0,0789	0,3978	0,7522	0,5234	0,7712	**0,7757**

Table 5.3: Harmonic mean of the NB classifier: German data

	PCA 95%			Z-Scored		
Default ratio	**Frequentist prior**	$p = 0.2$	$p = 0.5$	**Frequentist prior**	$p = 0.1$	$p = 0.5$
0,3000	**0,7630**	0,7490	0,7300	0,7380	0,7570	0,6890
0,2857	**0,7673**	0,7561	0,7255	0,7327	0,7541	0,6857
0,2708	**0,7719**	0,7688	0,7333	0,7458	0,7667	0,6865
0,2553	**0,7840**	0,7809	0,7255	0,7617	0,7734	0,6872
0,2391	0,7924	**0,7935**	0,7315	0,7587	0,7870	0,6772
0,2222	0,8033	**0,8067**	0,7311	0,7600	0,7933	0,6744
0,2045	0,8148	**0,8216**	0,7477	0,7545	0,8045	0,6591
0,1860	0,8326	**0,8372**	0,7500	0,7744	0,8221	0,6628
0,1667	0,8357	**0,8429**	0,7548	0,8024	0,8274	0,6798
0,1463	**0,8524**	0,8488	0,7598	0,8220	0,8317	0,7012
0,1250	**0,8725**	0,8713	0,7738	0,8400	0,8463	0,7200
0,1026	**0,8923**	0,8821	0,7667	0,8526	0,8551	0,7308
0,0789	**0,9158**	0,8868	0,7855	0,8711	0,8671	0,7750

Table 5.4: Hit rate of the NB classifier: German data

The Silverman classifier performs better for the majority of default ratios on the PCA95% data. This may be explained in a similar fashion as with the NB classifier. Silverman's rule of thumb estimates a diagonal bandwidth matrix and therefore does not take into account covariances, and hence correlations, between features. Since the PCA95% data's features are uncorrelated the classifier performs better on this data set.

It is worth noting that the frequentist prior reduces the harmonic mean of the Silverman classifier. For both the z-scored as well as the PCA95% data whether the Bernoulli prior results in the optimal harmonic mean depends on the default ratio. However, the dependency

is not explained by the relationship given by the frequentist prior. Table 5.5 also suggests that the use of a Bernoulli prior with a finer tuned parameter might be more appropriate. The frequentist prior reduce the hit rate of the Silverman classifier on the z-scored data, regardless of the default ratio. Considering the PCA95% data the frequentist prior reduce the hit rate of the Silverman classifier for the greater class imbalances. The hit rate for both the z-scored as well as the PCA95% data is optimised using a Bernoulli prior. However, there are a few default ratios for which not using a prior optimises the hit rate.

	PCA 95%			Z-Scored		
Default ratio	**Frequentist prior**	$p = 0.6$	$p = 0.5$	**Frequentist prior**	$p = 0.8$	$p = 0.5$
0,3000	0,9528	**0,9589**	0,9586	0,9486	0,9518	0,9586
0,2857	0,9474	**0,9606**	0,9572	0,9473	0,9532	0,9583
0,2708	0,9512	**0,9620**	0,9574	0,9444	0,9554	0,9587
0,2553	0,9474	**0,9620**	0,9591	0,9403	0,9561	0,9605
0,2391	0,9478	**0,9620**	0,9594	0,9452	0,9546	0,9610
0,2222	0,9442	0,9594	**0,9622**	0,9457	0,9574	0,9615
0,2045	0,9496	0,9577	**0,9612**	0,9469	0,9561	0,9605
0,1860	0,9267	0,9613	**0,9641**	0,9235	0,9593	0,9634
0,1667	0,9241	0,9663	**0,9691**	0,9182	0,9664	0,9648
0,1463	0,9012	0,9669	0,9697	0,8950	**0,9725**	0,9661
0,1250	0,8804	0,9675	**0,9675**	0,8799	0,9627	0,9637
0,1026	0,8631	0,9617	**0,9624**	0,8396	0,9583	0,9435
0,0789	0,7870	0,9608	0,9435	0,7230	**0,9615**	0,8561

Table 5.5: Harmonic mean of the Silverman classifier: German data

	PCA 95%			Z-Scored		
Default ratio	**Frequentist prior**	$p = 0.4$	$p = 0.5$	**Frequentist prior**	$p = 0.4$	$p = 0.5$
0,3000	0,9660	**0,9690**	0,9650	0,9630	0,9650	0,9650
0,2857	0,9653	**0,9684**	0,9653	0,9633	0,9663	0,9653
0,2708	0,9698	**0,9677**	0,9667	0,9635	**0,9677**	0,9667
0,2553	**0,9691**	**0,9691**	0,9681	0,9628	0,9681	0,9681
0,2391	0,9707	**0,9728**	0,9696	0,9663	0,9717	0,9696
0,2222	0,9722	**0,9744**	0,9722	0,9678	0,9733	0,9711
0,2045	0,9750	**0,9761**	0,9716	0,9705	0,9750	0,9705
0,1860	0,9698	**0,9791**	0,9733	0,9640	0,9756	0,9721
0,1667	0,9714	**0,9774**	**0,9774**	0,9667	**0,9774**	0,9750
0,1463	0,9683	**0,9793**	0,9780	0,9646	**0,9793**	0,9780
0,1250	0,9688	**0,9838**	0,9813	0,9675	0,9800	0,9825
0,1026	0,9705	**0,9846**	0,9833	0,9692	0,9769	0,9821
0,0789	0,9697	0,9816	**0,9842**	0,9645	0,9776	0,9776

Table 5.6: Hit rate of the Silverman classifier: German data

The MLE classifier performs better on the z-scored data rather than the PCA95% data.

This might be attributed to the fact that the MLE density estimator is able to capture correlation between variables owing to the adaptive kernel bandwidths. The MLE is able to capture correlation between variables since the MLE estimates a unique bandwidth in each dimension for each kernel. This is despite the fact that diagonal bandwidth matrices are estimated in the experiment.

The frequentist prior used in conjunction with the MLE classifier decrease the harmonic mean on the z-scored data set, whereas a Bernoulli prior of $p = 0.4$ results in the optimal harmonic mean for the majority of default ratios. The fact that there are a few default ratios for which the use of no prior results in the optimal performance suggest that a finer tuned Bernoulli prior parameter might be more appropriate. In the case of the PCA95% data the frequentist prior performs optimally, in terms of the harmonic mean, for the smaller class-imbalances. However, as the class-imbalance increase the frequentist prior results in a lower harmonic mean than that of the classifier used without a prior. Regardless of the class-imbalance the Bernoulli prior with parameter $p = 0.9$ has a higher harmonic mean than the classifier used without a prior. For the greater class-imbalances the Bernoulli prior outperforms, in terms of the harmonic mean, the frequentist prior. This, as with the other classifiers, suggest that the optimal prior depends on the default ratio. Yet the dependency is not the same as that given by the frequentist prior. See Table 5.7.

The hit rate of the MLE classifier is optimised on the z-scored data by the Bernoulli prior with parameter $p = 0.4$, regardless of the default ratio. On the other hand, the frequentist prior results in a lower hit rate than the MLE classifier with no prior, for the majority of default ratios. Considering the PCA95% data the frequentist prior results in the optimal hit rate regardless of the default ratio, with the only exception being the smallest evaluated default ratio. For the majority of evaluated default ratios the Bernoulli prior, with parameter $p = 0.2$, improves the hit rate of the classifier when compared to the classifier used without a prior.

	PCA 95%			Z-Scored		
Default ratio	**Frequentist prior**	$p = 0.9$	$p = 0.5$	**Frequentist prior**	$p = 0.4$	$p = 0.5$
0,3000	0,9528	0,9413	0,9306	0,9486	**0,9616**	0,9609
0,2857	0,9474	0,9389	0,9304	0,9473	0,9594	**0,9610**
0,2708	0,9512	0,9407	0,9350	0,9444	**0,9642**	0,9638
0,2553	0,9474	0,9405	0,9328	0,9403	0,9596	**0,9683**
0,2391	0,9478	0,9399	0,9394	0,9452	0,9624	**0,9638**
0,2222	0,9442	0,9423	0,9281	0,9457	**0,9641**	0,9610
0,2045	0,9496	0,9443	0,9335	0,9469	**0,9655**	0,9620
0,1860	0,9267	0,9426	0,9245	0,9235	**0,9740**	0,9669
0,1667	0,9241	0,9536	0,9259	0,9182	**0,9749**	0,9686
0,1463	0,9012	0,9486	0,9186	0,8950	**0,9761**	0,9733
0,1250	0,8804	0,9513	0,9110	0,8799	**0,9741**	0,9678
0,1026	0,8631	0,9528	0,8988	0,8396	**0,9682**	0,9627
0,0789	0,7870	0,9506	0,9031	0,7230	**0,9602**	0,9561

Table 5.7: Harmonic mean of the MLE classifier: German data

	PCA 95%			Z-Scored		
Default ratio	**Frequentist prior**	$p = 0.2$	$p = 0.5$	**Frequentist prior**	$p = 0.4$	$p = 0.5$
0,3000	0,9660	0,9480	0,9460	0,9630	**0,9710**	0,9640
0,2857	0,9653	0,9469	0,9469	0,9633	**0,9704**	0,9643
0,2708	0,9698	0,9563	0,9531	0,9635	**0,9750**	0,9667
0,2553	0,9691	0,9585	0,9532	0,9628	**0,9713**	0,9691
0,2391	0,9707	0,9620	0,9598	0,9663	**0,9717**	0,9663
0,2222	**0,9722**	0,9578	0,9556	0,9678	**0,9722**	0,9644
0,2045	**0,9750**	0,9602	0,9636	0,9705	**0,9750**	0,9659
0,1860	0,9698	0,9628	0,9605	0,9640	**0,9814**	0,9698
0,1667	0,9714	0,9655	0,9619	0,9667	**0,9821**	0,9714
0,1463	0,9683	0,9646	0,9610	0,9646	**0,9829**	0,9780
0,1250	0,9688	0,9688	0,9650	0,9675	**0,9850**	0,9738
0,1026	0,9705	0,9718	0,9705	0,9692	**0,9833**	0,9731
0,0789	0,9697	0,9803	0,9737	0,9645	**0,9842**	0,9763

Table 5.8: Hit rate of the MLE classifier: German data

Australian data

It is evident that the Gaussian classifier performs better, as indicated by the bold values, on the PCA95% data than the z-scored% data. This is unexpected since the Gaussian classifier is capable of modelling correlations between features and the PCA95% data does not contain correlated features, whereas the z-scored data does contain correlated features. It is therefore expected that the Gaussian classifier performs better on the z-scored data.

The performance of Gaussian classifier on both the PCA95% as well as the z-scored data is optimised by using a Bernoulli prior. Compared to the classifier used without a prior,

the frequentist prior decrease the performance for larger default ratios and increase the performance for smaller default ratios. This is true regardless of whether the performance is measured in terms of the hit rate or the harmonic mean. The reader is referred to tables 5.9 and 5.10.

	PCA 95%			Z-Scored		
Default ratio	**Frequentist prior**	$p = 0.2$	$p = 0.5$	**Frequentist prior**	$p = 0.1$	$p = 0.5$
0,5551	0,7915	**0,8698**	0,8074	0,7768	0,8547	0,7794
0,5418	0,7921	**0,881**	0,8036	0,7743	0,8609	0,7831
0,5277	0,7958	**0,8689**	0,8112	0,7796	0,8614	0,7818
0,5127	0,7948	**0,8666**	0,799	0,7802	0,8597	0,7791
0,4967	0,7952	**0,8689**	0,7931	0,782	0,8629	0,782
0,4797	0,8045	**0,8677**	0,8031	0,7958	0,858	0,7893
0,4614	0,8158	**0,8669**	0,8114	0,8017	0,8659	0,7937
0,4418	0,8163	0,8621	0,817	0,8059	**0,8651**	0,7999
0,4208	0,8183	**0,866**	0,8129	0,8136	0,8638	0,8027
0,398	0,8124	**0,8692**	0,8042	0,8083	0,869	0,7947
0,3735	0,8247	0,8649	0,8024	0,8171	**0,8671**	0,7986
0,3468	0,8229	**0,8589**	0,799	0,8078	0,8579	0,7892
0,3178	0,8383	0,8551	0,8053	0,8163	**0,8669**	0,8001
0,286	0,8525	**0,8592**	0,8207	0,8268	0,847	0,7947
0,2512	0,8625	**0,8648**	0,8345	0,8519	0,8509	0,8071
0,2128	0,8669	**0,8755**	0,8396	0,8535	0,8657	0,822
0,1703	0,8691	**0,8714**	0,8527	0,8636	0,8679	0,8258
0,1229	0,8964	**0,9041**	0,8708	0,883	0,8848	0,8438

Table 5.9: Harmonic mean of the Gaussian classifier: Australian data

	PCA 95%			Z-Scored		
Default ratio	**Frequentist prior**	$p = 0.1$	$p = 0.5$	**Frequentist prior**	$p = 0.1$	$p = 0.5$
0,5551	0,8217	**0,8667**	0,8304	0,8145	0,8594	0,8145
0,5418	0,8179	**0,8672**	0,8254	0,8090	0,8642	0,8149
0,5277	0,8169	**0,8662**	0,8277	0,8077	0,8631	0,8092
0,5127	0,8111	**0,8635**	0,8143	0,8048	0,8619	0,8032
0,4967	0,8082	**0,8656**	0,8066	0,8016	0,8639	0,8016
0,4797	0,8119	**0,8678**	0,8119	0,8085	0,8576	0,8034
0,4614	0,8175	**0,8667**	0,8140	0,8088	0,8649	0,8035
0,4418	0,8145	**0,8709**	0,8164	0,8073	0,8636	0,8036
0,4208	0,8132	**0,8679**	0,8094	0,8094	0,8623	0,8019
0,398	0,8039	**0,8686**	0,7961	0,8000	0,8667	0,7882
0,3735	0,8143	**0,8694**	0,7898	0,8041	0,8653	0,7857
0,3468	0,8085	**0,8702**	0,7809	0,7894	0,8511	0,7702
0,3178	0,8222	**0,8756**	0,7800	0,7933	0,8578	0,7733
0,286	0,8349	**0,8744**	0,7884	0,7930	0,8279	0,7581
0,2512	0,8488	**0,8854**	0,7976	0,8122	0,8293	0,7610
0,2128	0,8538	**0,8897**	0,8051	0,8154	0,8410	0,7692
0,1703	0,8568	**0,8757**	0,8027	0,8243	0,8378	0,7622
0,1229	0,8886	**0,8943**	0,8200	0,8514	0,8543	0,7914

Table 5.10: Hit rate of the Gaussian classifier: Australian data

The NB classifier performs better on the PCA95% data than the z-scored data. This can be explained by the fact that there is no correlation present in the PCA95% data. Since the NB classifier does not take account of the covariances, and hence correlations, it follows intuitively that the classifier will perform better on the PCA95% data than the z-scored data. A Bernoulli prior with parameter $p = 0.1$ results in the optimal harmonic mean for the NB classifier applied to the z-scored data, regardless of the default ratio. The only exception being the largest evaluated class-imbalance. There are a few default ratios for which the frequentist prior increase the performance of the classifier. However, for larger default ratios the frequentist prior has a lower harmonic mean than the classifier used without a prior. Considering the harmonic mean of the NB classifier applied to the PCA95% data, there is no single type of prior that results in the optimal harmonic mean regardless of the class-imbalance. The prior is therefore dependent on the class imbalance. However, the dependency is not the same as that given by the frequentist prior. This is reflected in Table 5.11.
The hit rate of the NB classifier applied to the z-scored data is optimised by the Bernoulli prior with parameter $p = 0.1$. This is the same prior that optimises the harmonic mean. For greater default ratios the frequentist prior results in a lower hit rate than the NB classifier used with no prior. However, as the default ratio decrease the frequentist prior improves the hit rate. Considering the PCA95% data, the frequentist prior as well as the Bernoulli prior results in a higher hit rate for smaller default ratios. On the other hand, for larger default

ratios the NB classifier used without any prior results in the optimal hit rate. This once again suggests that there exists a dependency between the prior used and the class-imbalance. See Table 5.12.

	PCA 95%			Z-Scored		
Default ratio	**Frequentist prior**	$p = 0.4$	$p = 0.5$	**Frequentist prior**	$p = 0.1$	$p = 0.5$
0,5551	0,8566	0,8572	**0,8606**	0,7815	0,8316	0,7839
0,5418	0,8587	0,8624	**0,8665**	0,7828	0,8352	0,7891
0,5277	0,8579	0,8591	**0,8653**	0,7872	0,8359	0,7893
0,5127	0,8476	0,8539	**0,8615**	0,7822	0,8331	0,7832
0,4967	0,8505	0,8536	**0,8603**	0,7835	0,8325	0,7835
0,4797	0,8561	**0,8595**	0,8563	0,7853	0,8368	0,7872
0,4614	0,8528	**0,8587**	0,8571	0,7881	0,8380	0,7861
0,4418	**0,8586**	0,8559	0,8578	0,7866	0,8390	0,7877
0,4208	**0,8640**	0,8577	0,8575	0,7961	0,8363	0,7891
0,3980	**0,8621**	0,8548	0,8572	0,8045	0,8369	0,7901
0,3735	**0,8605**	0,8514	0,8544	0,8064	0,8433	0,7862
0,3468	0,8452	**0,8500**	0,8389	0,8022	0,8449	0,7887
0,3178	0,8400	**0,8416**	0,8266	0,7972	0,8370	0,7870
0,2860	**0,8500**	0,8447	0,8403	0,8076	0,8380	0,7969
0,2512	0,8515	0,8453	**0,8554**	0,8208	0,8398	0,8008
0,2128	0,8397	**0,8470**	0,8462	0,8199	0,8331	0,8060
0,1703	**0,8767**	0,8561	0,8553	0,8226	0,8285	0,8014
0,1229	**0,8456**	0,8439	0,8404	0,8387	0,8423	0,8432

Table 5.11: Harmonic mean of the NB classifier: Australian data

	PCA 95%			Z-Scored		
Default ratio	**Frequentist prior**	$p = 0.2$	$p = 0.5$	**Frequentist prior**	$p = 0.1$	$p = 0.5$
0,5551	**0,8623**	0,8188	**0,8623**	0,8101	0,8377	0,8087
0,5418	0,8627	0,8254	**0,8687**	0,8075	0,8403	0,8119
0,5277	0,8600	0,8277	**0,8662**	0,8077	0,8400	0,8092
0,5127	0,8492	0,8222	**0,8619**	0,8016	0,8365	0,8016
0,4967	0,8508	0,8295	**0,8607**	0,7984	0,8344	0,7984
0,4797	**0,8559**	0,8407	**0,8559**	0,7949	0,8373	0,7983
0,4614	0,8526	0,8491	**0,8561**	0,7930	0,8368	0,7912
0,4418	**0,8600**	0,8509	0,8564	0,7873	0,8364	0,7891
0,4208	**0,8660**	0,8547	0,8547	0,7925	0,8321	0,7868
0,3980	**0,8667**	0,8569	0,8529	0,7961	0,8314	0,7824
0,3735	**0,8714**	0,8653	0,8510	0,7939	0,8367	0,7735
0,3468	**0,8638**	0,8596	0,8319	0,7851	0,8340	0,7702
0,3178	**0,8711**	0,8667	0,8222	0,7756	0,8222	0,7622
0,2860	**0,8837**	0,8791	0,8349	0,7814	0,8186	0,7651
0,2512	**0,8927**	0,8878	0,8512	0,7902	0,8146	0,7634
0,2128	0,8897	**0,9000**	0,8410	0,7872	0,8051	0,7615
0,1703	0,9162	**0,9189**	0,8541	0,7865	0,8027	0,7568
0,1229	**0,9200**	0,9143	0,8429	0,8086	0,8143	0,7800

Table 5.12: Hit rate of the NB classifier: Australian data

There is no definitive conclusion that can be made regarding whether the Silverman classifier performs better on the z-scored or the PCA95% data. For some default ratios the classifier performs better on the z-scored data and others the PCA95% data. The default ratios for which it has a higher hit rate on the z-scored data do not even coincide with the default ratios for which it has a higher harmonic mean on the z-scored data. This is unexpected since the nature of the Silverman classifier suggests it should perform better on a data set with no correlation between features.

A Bernoulli prior with parameter $p = 0.2$, used in conjunction with the Silverman classifier, results in the optimal harmonic mean for the z-scored data, regardless of the class imbalance. The frequentist prior improves on the harmonic mean of the classifier for default ratios of 0.4797 and lower. No specific prior results in the optimal harmonic mean regardless of class-imbalance for the PCA95% data. This suggests that a Bernoulli prior with the parameter dependent upon the class-imbalance might be appropriate. This is reflected in Table 5.13.

The hit rate of the Silverman classifier applied to the z-scored data is the highest for the Bernoulli prior with parameter $p = 0.2$, for all evaluated class-imbalances. Although the frequentist prior improves the hit rate of the Silverman classifier for default ratios smaller and equal to 0.4797, it does not improve the hit rate by such a margin that the frequentist prior outperforms the Bernoulli prior with parameter $p = 0.2$. The frequentist prior results in the optimal hit rate for the PCA95% data with default ratios between 0.4967 and 0.2128 (both

included). For the majority of evaluated default ratios the Bernoulli prior with parameter $p = 0.2$ improves on the Silverman classifier without a prior. The aforementioned suggest that a Bernoulli prior with the parameter dependent upon the default ratio might be appropriate. See Table 5.14.

	PCA 95%			Z-Scored		
Default ratio	**Frequentist prior**	$p = 0.4$	$p = 0.5$	**Frequentist prior**	$p = 0.2$	$p = 0.5$
0,5551	0,9463	**0,9620**	0,9572	0,9025	0,9616	0,9179
0,5418	0,9459	0,9610	0,9568	0,9083	**0,9623**	0,9140
0,5277	0,9460	0,9604	0,9512	0,9140	**0,9618**	0,9197
0,5127	0,9507	0,9597	0,9524	0,9158	**0,9612**	0,9158
0,4967	0,9518	0,9606	0,9518	0,9157	**0,9638**	0,9157
0,4797	0,9597	0,9597	0,9546	0,9194	**0,9631**	0,9156
0,4614	0,9606	0,9621	0,9557	0,9303	**0,9640**	0,9228
0,4418	0,9594	0,9631	0,9571	0,9377	**0,9647**	0,9192
0,4208	0,9637	0,9620	0,9587	0,9468	**0,9653**	0,9228
0,3980	**0,9657**	0,9624	0,9600	0,9486	0,9640	0,9322
0,3735	**0,9694**	0,9641	0,9646	0,9568	0,9673	0,9359
0,3468	**0,9716**	0,9665	0,9571	0,9576	0,9681	0,9209
0,3178	0,9524	0,9662	0,9665	0,9665	**0,9697**	0,9228
0,2860	0,9472	0,9691	**0,9732**	0,9591	0,9722	0,9190
0,2512	0,9603	0,9691	0,9737	0,9703	**0,9787**	0,9377
0,2128	0,9545	0,9673	**0,9759**	0,9667	0,9684	0,9360
0,1703	0,9217	0,9614	**0,9740**	0,9630	0,9695	0,9450
0,1229	0,8672	0,9737	0,9704	0,9668	**0,9721**	0,9558

Table 5.13: Harmonic mean of the Silverman classifier: Australian data

	PCA 95%			Z-Scored		
Default ratio	**Frequentist prior**	$p = 0.2$	$p = 0.5$	**Frequentist prior**	$p = 0.2$	$p = 0.5$
0,5551	0,9522	0,9565	0,9609	0,9025	**0,9638**	0,9319
0,5418	0,9507	0,9552	0,9597	0,9083	**0,9642**	0,9269
0,5277	0,9492	0,9554	0,9538	0,9140	**0,9631**	0,9292
0,5127	0,9524	0,9540	0,9540	0,9158	**0,9619**	0,9238
0,4967	0,9525	0,9525	0,9525	0,9157	**0,9639**	0,9213
0,4797	0,9593	0,9492	0,9542	0,9194	**0,9627**	0,9186
0,4614	0,9596	0,9544	0,9544	0,9303	**0,9632**	0,9228
0,4418	0,9582	0,9564	0,9545	0,9377	**0,9636**	0,9164
0,4208	0,9623	0,9566	0,9547	0,9468	**0,9642**	0,9170
0,398	**0,9647**	0,9569	0,9549	0,9486	0,9627	0,9235
0,3735	**0,9714**	0,9592	0,9592	0,9568	0,9673	0,9245
0,3468	**0,9723**	0,9617	0,9489	0,9576	0,9660	0,9043
0,3178	0,9644	0,9622	0,9578	0,9665	**0,9689**	0,9022
0,286	0,9651	0,9651	0,9628	0,9591	**0,9674**	0,8930
0,2512	**0,9756**	0,9756	0,9659	0,9703	0,9732	0,9122
0,2128	**0,9795**	0,9769	0,9692	0,9667	0,9641	0,9103
0,1703	0,9730	**0,9784**	0,9676	0,9630	0,9703	0,9216
0,1229	0,9686	**0,9857**	0,9657	0,9668	0,9686	0,9257

Table 5.14: Hit rate of the Silverman classifier: Australian data

The MLE classifier performs better on the PCA95% data rather than the z-scored data. This is due to the fact that the version of the MLE classifier used estimates diagonal bandwidth matrices. Although the MLE classifier can still detect correlations due to the MLE's ability to estimates an unique bandwidth in each dimension for each kernel, it would appear that on the Australian data set the fact that the MLE estimates diagonal bandwidths outweighs this aforementioned ability. See tables 5.15 and 5.16.

The harmonic mean of the MLE classifier applied to the z-scored data suggests that a Bernoulli prior with parameter $p = 0.2$ is most appropriate. The frequentist prior improves the harmonic mean of the MLE classifier for the z-scored data with evaluated default ratios less than 0.5127. Considering the PCA95% data there is no definitive optimal prior regardless of default ratio. This suggests that a Bernoulli prior that is dependent upon the default ratio is most appropriate. This can be seen by observing Table 5.15. The Bernoulli prior with parameter $p = 0.4$ improves the hit rate of the MLE classifier for both the PCA95% as well as the z-scored data , regardless of the class-imbalance. The prior results in the optimal hit rate for the majority of default ratios in both cases. The frequentist prior is shown to be ineffective for the PCA95% and often reduce the hit rate of the MLE classifier. The reader is referred to Table 5.16.

	PCA 95%			Z-Scored		
Default ratio	**Frequentist prior**	$p = 0.7$	$p = 0.5$	**Frequentist prior**	$p = 0.2$	$p = 0.5$
0,5551	0,9409	0,9198	**0,9479**	0,8244	0,9229	0,8417
0,5418	**0,9460**	0,9194	**0,9460**	0,8155	0,9274	0,8306
0,5277	0,9402	0,9190	**0,9454**	0,8224	0,9220	0,8331
0,5127	**0,9463**	0,9204	0,9411	0,8207	0,9224	0,8313
0,4967	**0,9404**	0,9231	0,9385	0,8317	0,9176	0,8308
0,4797	**0,9411**	0,9262	0,9410	0,8279	0,9172	0,8249
0,4614	0,9395	0,9293	**0,9400**	0,8444	0,9261	0,8341
0,4418	**0,9460**	0,9358	0,9441	0,8457	0,9278	0,8348
0,4208	0,9515	0,9336	**0,9519**	0,8578	0,9224	0,8344
0,398	0,9509	0,9361	**0,9540**	0,8712	0,9122	0,8495
0,3735	0,9431	0,9379	**0,9548**	0,8831	0,9111	0,8381
0,3468	0,9401	0,9420	**0,9522**	0,8850	0,9181	0,8345
0,3178	0,9404	**0,9547**	0,9472	0,8973	0,9317	0,8470
0,286	0,9308	**0,9569**	0,9501	0,9037	0,9304	0,8615
0,2512	0,9284	**0,9610**	0,9541	0,9411	0,9497	0,8838
0,2128	0,9262	**0,9650**	0,9612	0,9393	0,9315	0,8936
0,1703	0,9017	**0,9690**	0,9662	0,9381	0,9356	0,9028
0,1229	0,8961	**0,9638**	0,9786	0,8922	0,9347	0,8920

Table 5.15: Harmonic mean of the MLE classifier: Australian data

	PCA 95%			Z-Scored		
Default ratio	**Frequentist prior**	$p = 0.4$	$p = 0.5$	**Frequentist prior**	$p = 0.2$	$p = 0.5$
0,5551	0,9464	**0,9536**	0,9522	0,8594	0,9217	0,8696
0,5418	0,9493	**0,9522**	0,9493	0,8478	0,9269	0,8582
0,5277	0,9431	**0,9508**	0,9477	0,8492	0,9215	0,8569
0,5127	0,9476	**0,9492**	0,9429	0,8429	0,9222	0,8508
0,4967	0,9410	**0,9492**	0,9393	0,8475	0,9180	0,8475
0,4797	0,9407	**0,9458**	0,9407	0,8407	0,9169	0,8390
0,4614	0,9386	**0,9421**	0,9386	0,8491	0,9263	0,8421
0,4418	0,9455	**0,9509**	0,9418	0,8455	0,9291	0,8382
0,4208	**0,9509**	**0,9509**	0,9491	0,8528	0,9245	0,8321
0,398	0,9510	**0,9549**	0,9510	0,8627	0,9157	0,8412
0,3735	0,9469	**0,9531**	**0,9531**	0,8735	0,9204	0,8245
0,3468	0,9468	**0,9489**	**0,9489**	0,8723	0,9234	0,8149
0,3178	0,9533	**0,9556**	0,9511	0,8889	0,9378	0,8222
0,286	0,9581	0,9581	**0,9651**	0,8977	0,9395	0,8326
0,2512	0,9610	**0,9659**	**0,9659**	0,9317	0,9488	0,8512
0,2128	0,9667	**0,9769**	0,9744	0,9462	0,9410	0,8564
0,1703	0,9649	**0,9784**	0,9757	0,9622	0,9459	0,8595
0,1229	0,9743	**0,9829**	0,9800	0,9657	0,9571	0,8400

Table 5.16: Hit rate of the MLE classifier: Australian data

5.2.3 Parametric versus non-parametric classifiers

German data

Figure 5.33 indicates that the non-parametric classifiers have higher hit rates than the parametric classifiers for the PCA95% data. All the classifiers, with the exception of the BLR classifier, exhibit a general increasing trend. The increasing trend is expected as the number of instances classified correctly when simply classifying all instances to the majority class, increases as the class-imbalance increase. It is possible that the Silverman classifier result in higher hit rates than the MLE classifier due to the fact that only a single iteration is used to update the MLE bandwidth. Another explanation is that the MLE performs better in terms of density estimation and not necessarily classification. It is interesting to note that even though the PCA95% data do not contain correlation between features, the Gaussian classifier still outperforms the NB classifier. It was expected that these two classifiers result in the same hit rate for the PCA95% data set. The trend with which the NB as well as the Gaussian classifiers increase is greater than that of the LR classifier. The result is that the NB and Gaussian classifiers perform considerably better than the LR classifier for the greater evaluated class-imbalances. In comparison the BLR classifier have much lower respective hit rates than the LR classifier. This might be due to the fact that the posterior predictive distribution of the BLR classifier is approximated.
In terms of the harmonic mean the non-parametric classifiers outperform the parametric classifiers. The non-parametric classifiers maintain a fairly constant harmonic mean regardless of the class-imbalance. As is the case with the hit rate, the harmonic mean of the Silverman classifier is higher than that of the MLE classifier. The slope of the Gaussian classifier is steeper than that of the NB classifier. Both these classifiers exhibit a general upward trend. The Gaussian classifier has a higher harmonic mean than the NB classifier regardless of the default ratio. It is interesting to note that even though the LR classifier results in a higher hit rate than the BLR classifier for every default ratio, the BLR results in a higher harmonic mean than the LR for every default ratio. That is to say that the BLR has a higher quality of classification. This suggest that although the LR classifier might classify a greater proportion of classes correctly, should a cost matrix be involved it might be that the BLR classifier will result in more cost effective classifications. It is important to notice that the LR as well as the BLR classifier have decreasing trends in terms of the harmonic mean. This is reflected in Figure 5.34.

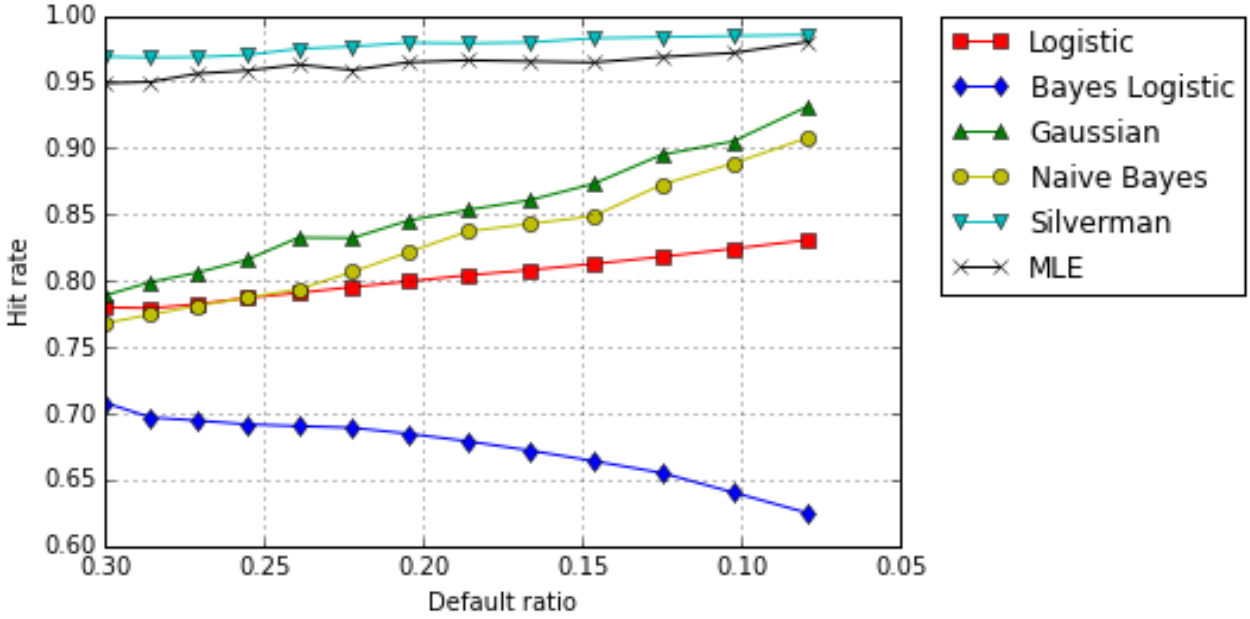

Figure 5.33: Performance of parametric versus non-parametric classifiers: German PCA 95% data

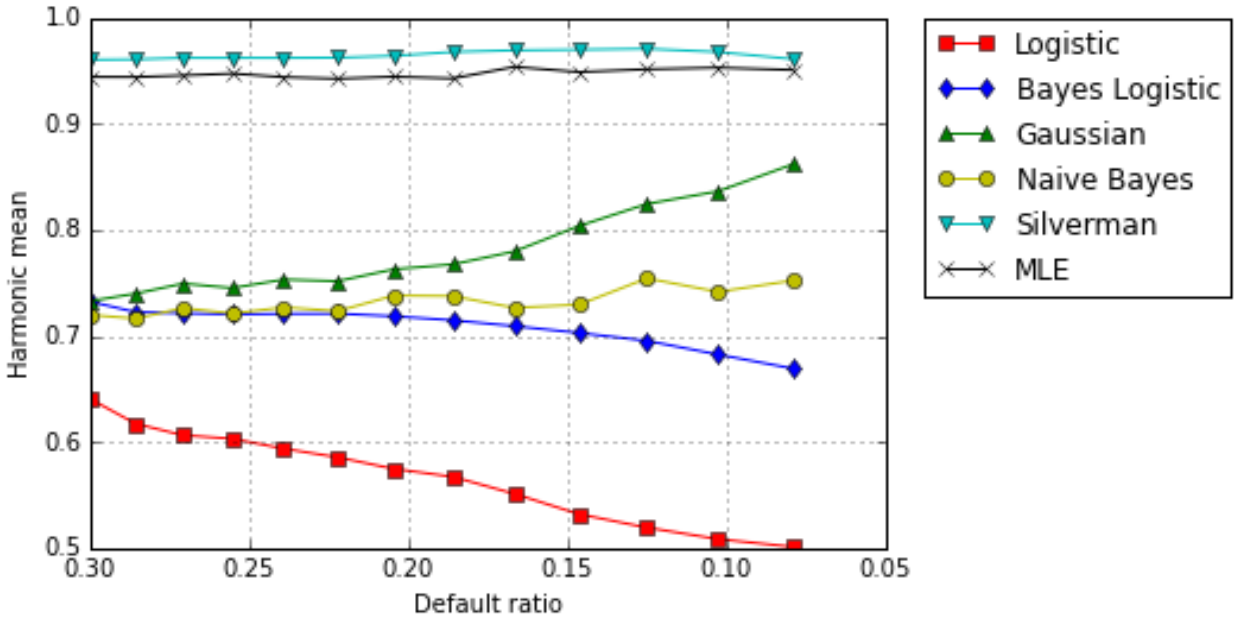

Figure 5.34: Performance of parametric versus non-parametric classifiers: German PCA 95% data

Comparing Figure 5.35 to Figure 5.33 it is seen that whereas the Silverman classifier performs much better than the MLE classifier on the PCA95% data, the hit rate of these classifiers are much more similar for the z-scored data. This may be attributed to the MLE's ability to capture correlations between features, which the Silverman classifier is unable to do. The hit rate for the non-parametric classifiers are higher than that of the parametric classifiers across all default ratios, for the z-scored data. It is interesting to note that the LR classifier

outperforms the NB classifier for smaller evaluated class-imbalances. The slope of the NB classifier is so steep that the hit rate of the classifier surpasses that of the LR classifier for larger evaluated class-imbalances. The hit rate of the BLR classifier is the lowest of all the classifiers for the z-scored data regardless of default ratio. It is also the only classifier that has a negative trend in terms of the hit rate.

Consider Figure 5.36. As in the case of the hit rate the two non-parametric classifier have very similar harmonic means, with the MLE marginally outperforming the Silverman classifier for a few default ratios. This may once again be explained by the fact that the MLE is capable of modelling correlations between features, which the Silverman classifier is not able to do. The non-parametric classifiers outperform all the other evaluated classifiers in terms of the harmonic mean, for the z-scored data. The conclusions regarding the other classifiers are similar to that of the harmonic mean for the PCA95% data. However, the default ratio at which the NB classifier surpasses that of the BLR classifier is smaller than for the PCA95% data.

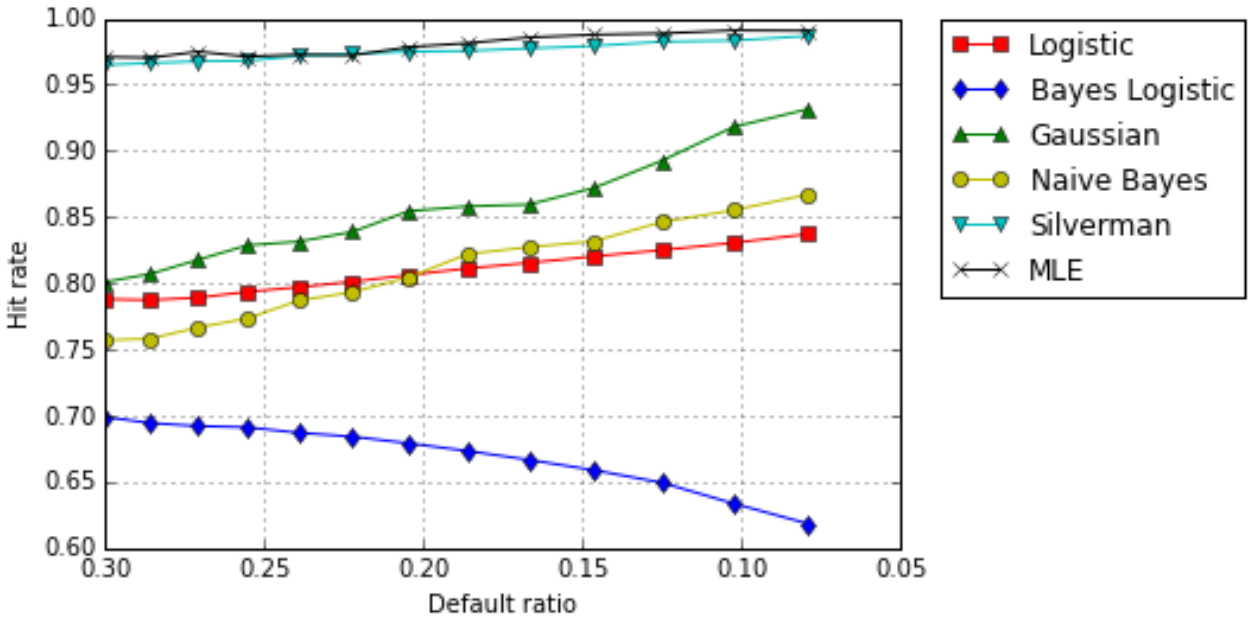

Figure 5.35: Performance of parametric versus non-parametric classifiers: German z-scored data

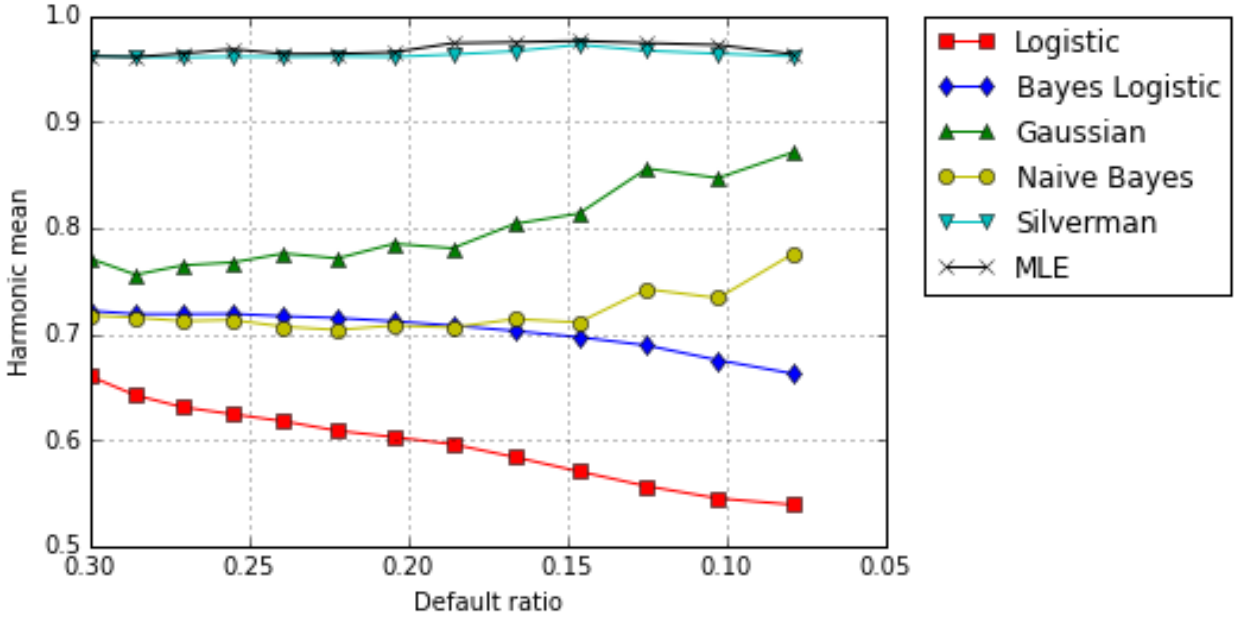

Figure 5.36: Performance of parametric versus non-parametric classifiers: German z-scored data

Australian data

The non-parametric classifiers outperform the parametric classifiers, in terms of the hit rate, for the PCA95% data. The Silverman classifier has a higher hit rate than the MLE classifier for the majority of default ratios, with the only exception being the two smallest evaluated default ratios. This may be attributed to the single, instead of multiple, iterations used to update the MLE bandwidth. The BLR and the LR classifier exhibit very similar performance in terms of the hit rate. The hit rate of both these classifiers remain fairly constant regardless of the class-imbalance. Originally these classifiers outperform the Gaussian as well as the NB classifiers. It is possible for the NB classifier to have a higher hit rate than the Gaussian classifier, at larger class-imbalances, since the PCA95% data do not contain correlations between features and the Gaussian classifier can mistakenly model covariances, which do not exist, whereas the NB classifier does not take covariances into account. See Figure 5.37.
Very similar conclusions can be made for the harmonic mean of the classifiers applied to the PCA95% data as those made for the hit rate applied to the same data. The non-parametric classifiers outperform the parametric classifiers. Furthermore, the LR and the BLR classifiers have very similar harmonic means regardless of the default ratio. These two classifiers have a fairly constant, yet slightly negative trend. The main difference in the performance evaluation in terms of the hit rate and the harmonic mean is surrounding the NB classifier. All the other evaluated classifiers outperform the NB classifier regardless of the default ratio. This is interesting since it is expected that due to the absence of correlations between features in the data set, that the NB and the Gaussian classifier would have similar harmonic means

regardless of the default ratio. It could even be expected that the NB classifier results in higher harmonic means than the Gaussian classifier due to the potential of the Gaussian classifier incorrectly modelling covariances between features, which do not exist. See Figure 5.38.

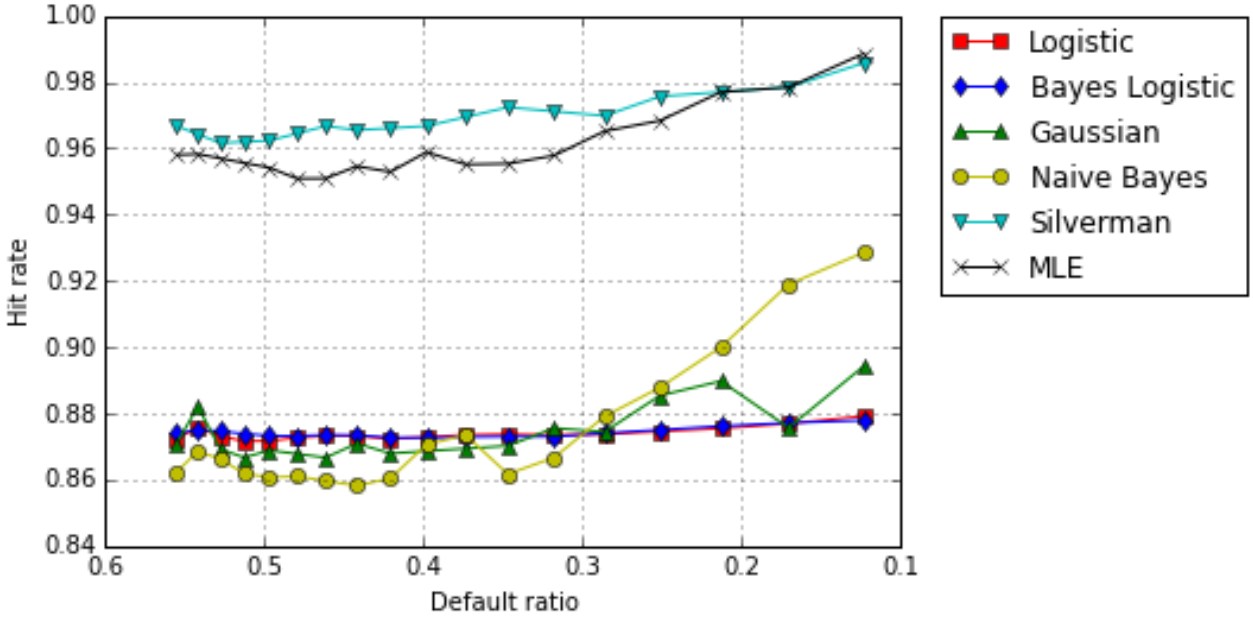

Figure 5.37: Performance of parametric versus non-parametric classifiers: Australian PCA 95% data

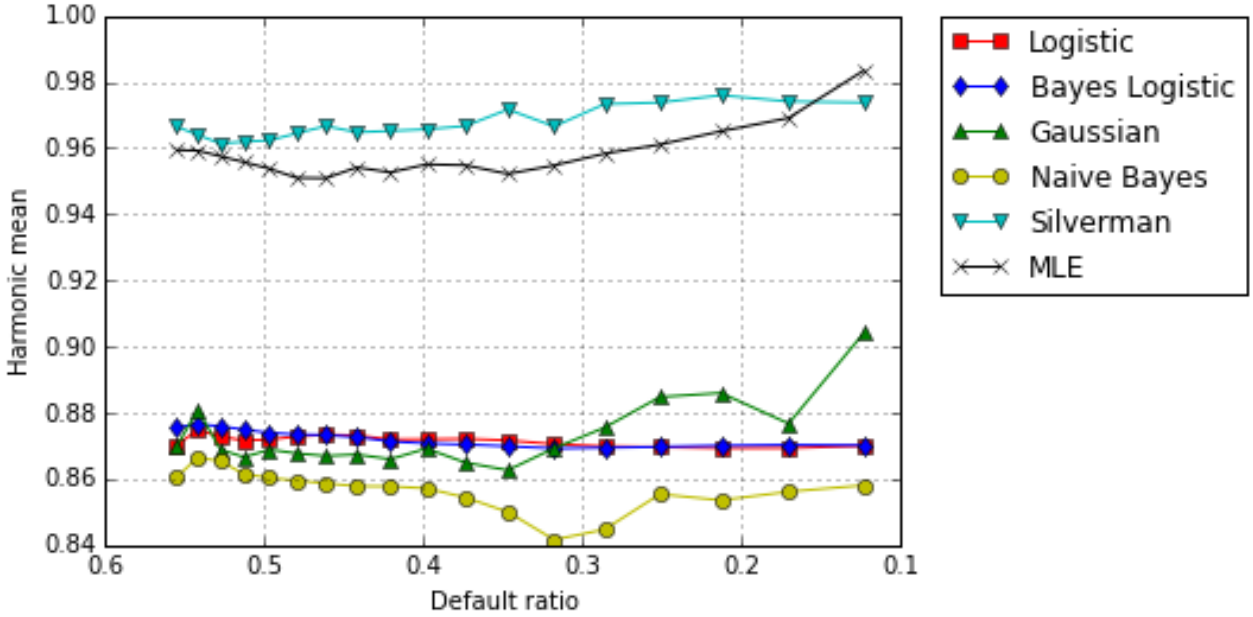

Figure 5.38: Performance of parametric versus non-parametric classifiers: Australian PCA 95% data

The non-parametric classifiers have higher hit rates than the parametric classifiers when applied to the z-scored data. Both the Silverman as well as the MLE classifier have a general

increasing trend, with that of the MLE being steeper than that of the Silverman classifier. The Silverman classifier have greater hit rates for all the evaluated default ratios than that of the MLE classifier. Although it is shown that the MLE is a better density estimator than Silverman Van der Walt 2014, it may be that the Silverman classifier is better at classification than the MLE classifier. The hit rate of the BLR as well as the LR classifier remain fairly constant for the z-scored data. These classifiers perform very similar with the LR classifier marginally outperforming the BLR classifier. These classifiers outperform the Gaussian as well as the NB classifiers across all evaluated default ratios. As expected the Gaussian classifier results in a higher hit rate, regardless of default ratio, than the NB classifier. This is due to the Gaussian classifier's ability to model correlations between features in the z-scored data set. Refer to Figure 5.39.

Very similar conclusions regarding the harmonic mean of the classifiers can be made as those made in terms of the hit rate. As with the hit rate, the non-parametric classifier have higher harmonic means than that of the parametric classifiers and the harmonic mean of the Silverman classifier is higher than that of the MLE classifier, regardless of the default ratio. Although the harmonic mean of the Logistic regression and BLR classifiers are very similar with the LR marginally outperforming the BLR classifier, they both have slightly decreasing trend. As with the hit rate the NB and Gaussian classifier have the lowest harmonic means of all the evaluated classifiers, with the Gaussian classifier outperforming the NB classifier as expected. This is reflected in Figure 5.40.

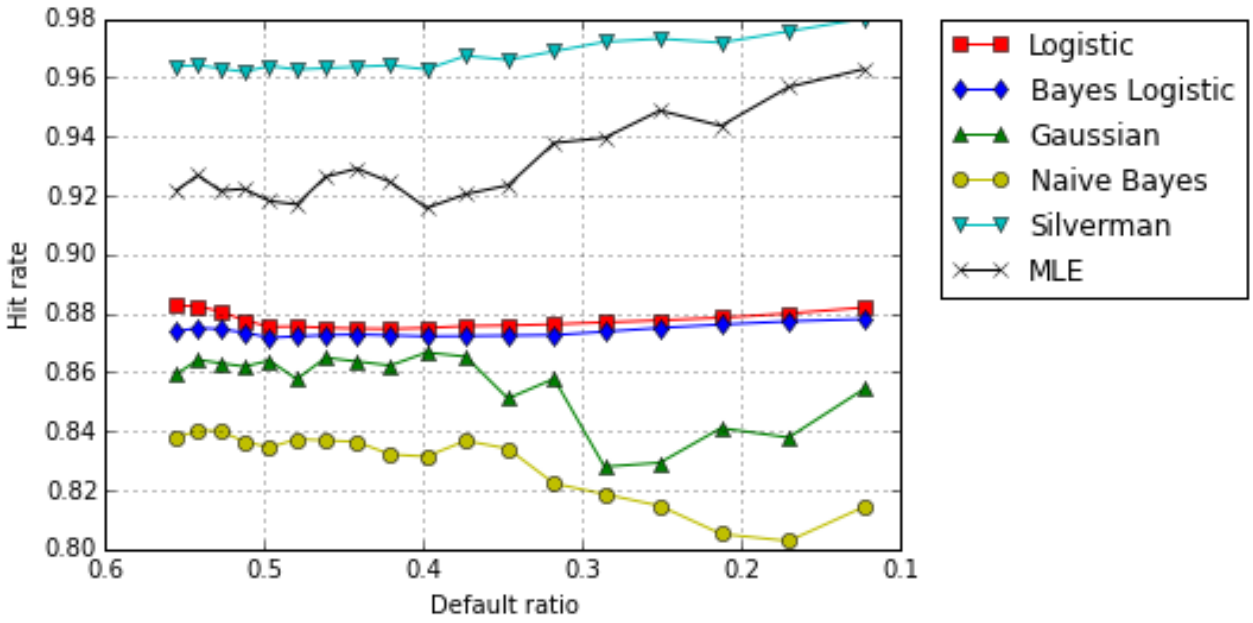

Figure 5.39: Performance of parametric versus non-parametric classifiers: Australian z-scored data

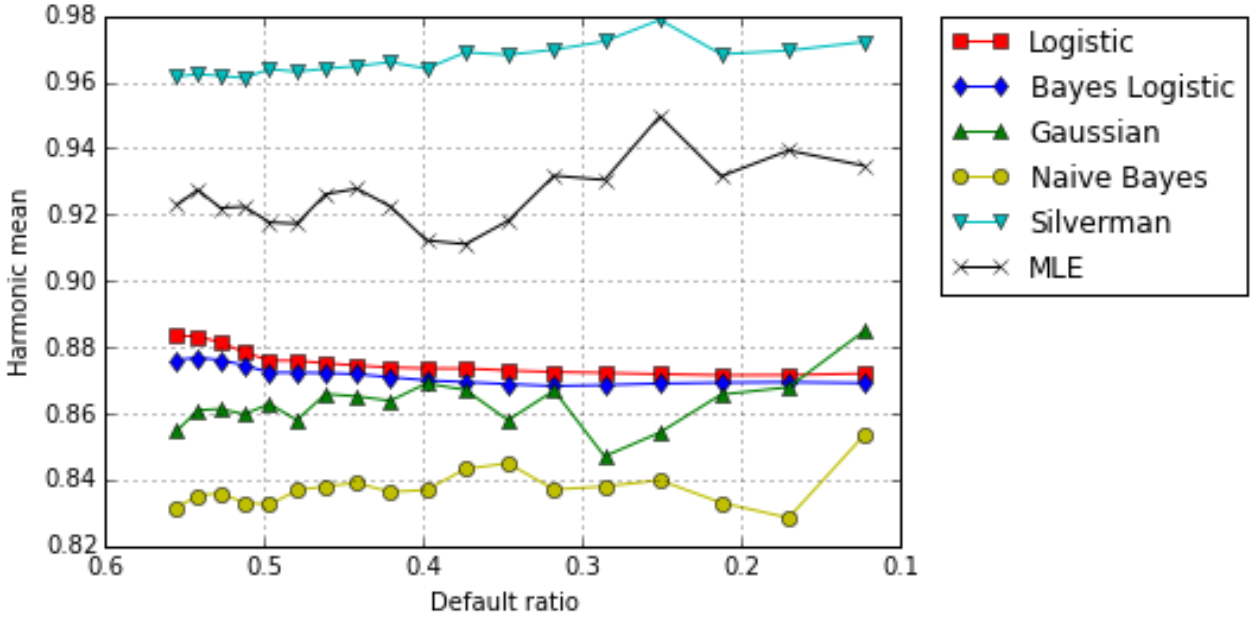

Figure 5.40: Performance of parametric versus non-parametric classifiers: Australian z-scored data

Lending club data

Considering Figure 5.41 it is important to note that the hit rates of the Silverman, NB, Gaussian as well as the LR classifiers closely resemble the proportion of non-defaulting instances. This suggest that these classifiers simply assign the majority of instances to the majority class. The BLR classifier is the only classifier that maintains a fairly constant hit rate regardless of the default ratio. It is interesting to observe that unlike for the other data sets where the number of observations as well as the default ratios were considerably larger, the hit rate of the parametric classifiers closely resemble that of the Silverman classifier. In particular the NB classifier performs very similar to the Silverman classifier, in terms of the hit rate. This might be due to the fact that NB classifier is incapable of modelling covariances and similarly the Silverman classifier's bandwidth is incapable of taking covariances into account.

The BLR classifier maintains a constant harmonic mean regardless of the class imbalance for the PCA95% data. This classifier outperforms all other evaluated classifiers in terms of the harmonic mean. This may be due to the added structure provided by the use of a prior distribution in order to obtain a posterior predictive distribution. The prior distribution ensures that the effect of the class imbalance is reduced. As expected the harmonic mean of the NB and the Gaussian classifiers are extremely similar. This is due to the fact that the PCA95% data do not contain correlations between features, rendering the Gaussian classifier's superior power of taking covariances into account redundant. The Silverman classifier performs similarly to the Gaussian and NB classifiers, only marginally outperforming them in terms

of the harmonic mean. The LR classifier performs extremely poor in terms of the harmonic mean, being practically naught, regardless of the class imbalance. This only emphasise the role the prior distribution plays in the BLR classifier. The reader is referred to Figure 5.42.

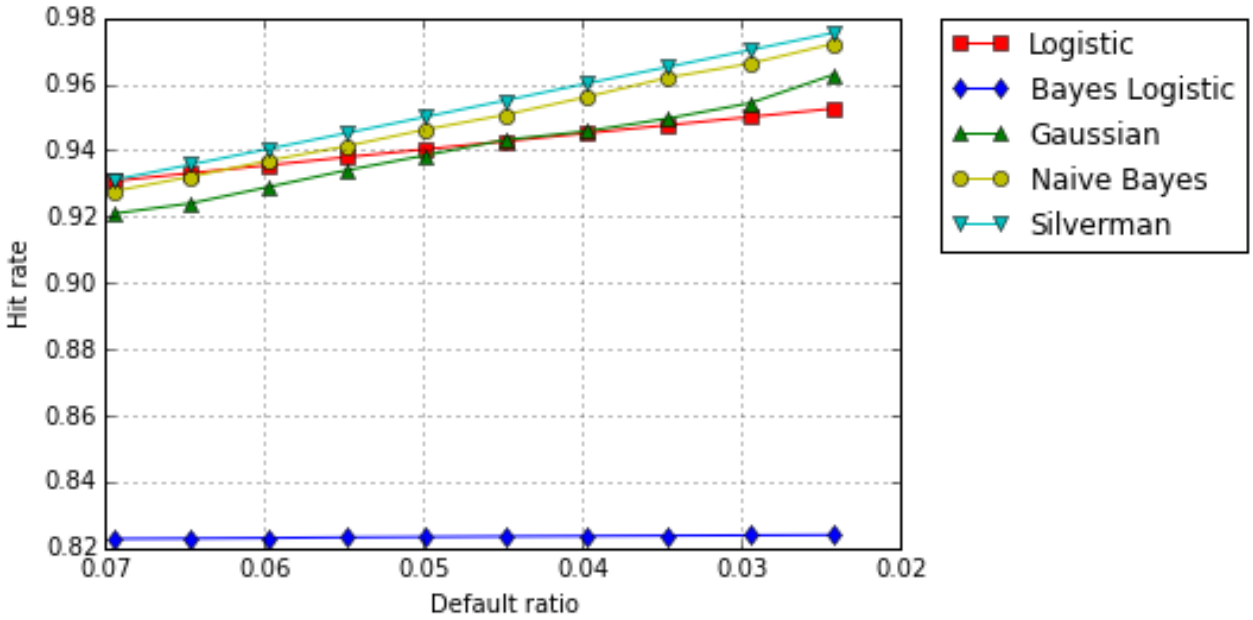

Figure 5.41: Performance of parametric versus non-parametric classifiers: Lending club PCA95% data

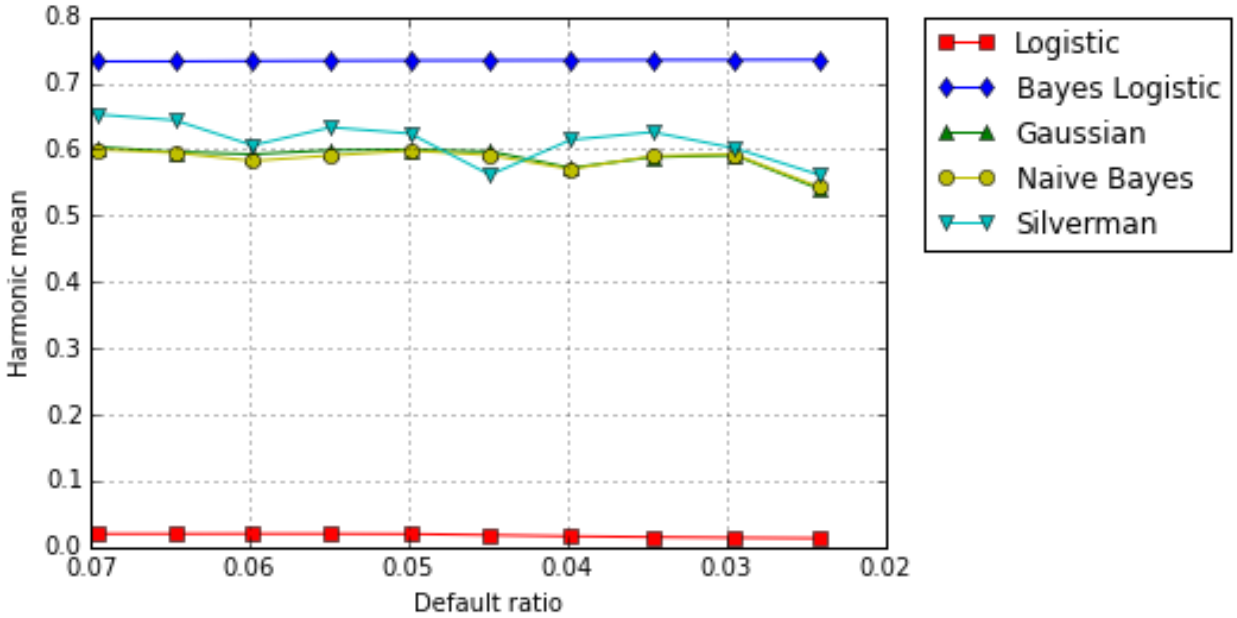

Figure 5.42: Performance of parametric versus non-parametric classifiers: Lending club PCA95% data

As with the PCA95% data the hit rate of the Silverman and NB classifiers closely resemble one another for the z-scored data. These classifiers have higher hit rates than all the other evaluated classifiers, regardless of the default ratio. Unlike for the PCA95% data the LR

classifier has a higher hit rate than the Gaussian classifier, regardless of the default ratio. This is unexpected since the Gaussian classifier is capable of modelling covariances and it is therefore expected to perform better on data that contains correlated features. Furthermore, the BLR classifier has an increasing trend, in terms of the hit rate, for the z-scored data. See Figure 5.43

The harmonic mean of the Gaussian classifier marginally outperforms the BLR classifier. As expected the Gaussian classifier exhibits a higher harmonic mean than the NB classifier, since the PCA95% data do not contain correlations. It is surprising that the NB, Gaussian as well as the BLR classifiers outperform the Silverman classifier, in terms of the harmonic mean. It is worth noting that the Silverman classifier does exhibit an general upward trend as the default ratio decrease. The LR classifier exhibit extremely poor performance in terms of the harmonic mean, with all the other evaluated classifiers significantly outperforming it. Even worse, the LR classifier decrease in the harmonic mean as the class imbalance increase. See Figure 5.44.

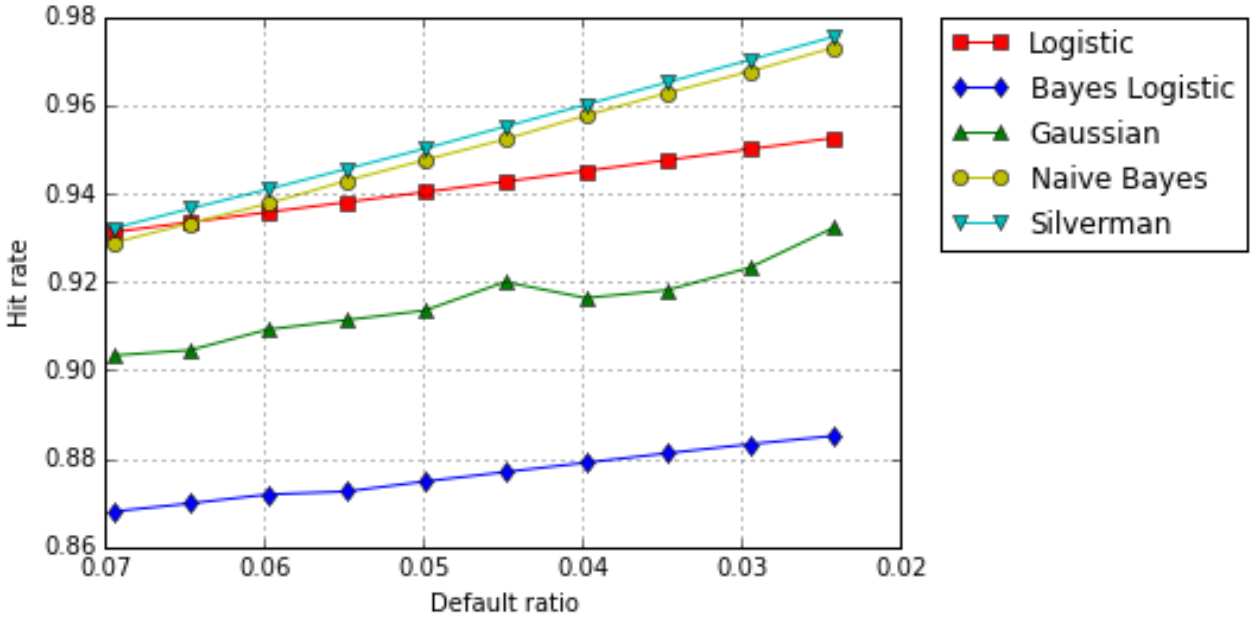

Figure 5.43: Performance of parametric versus non-parametric classifiers: Lending club z-scored data

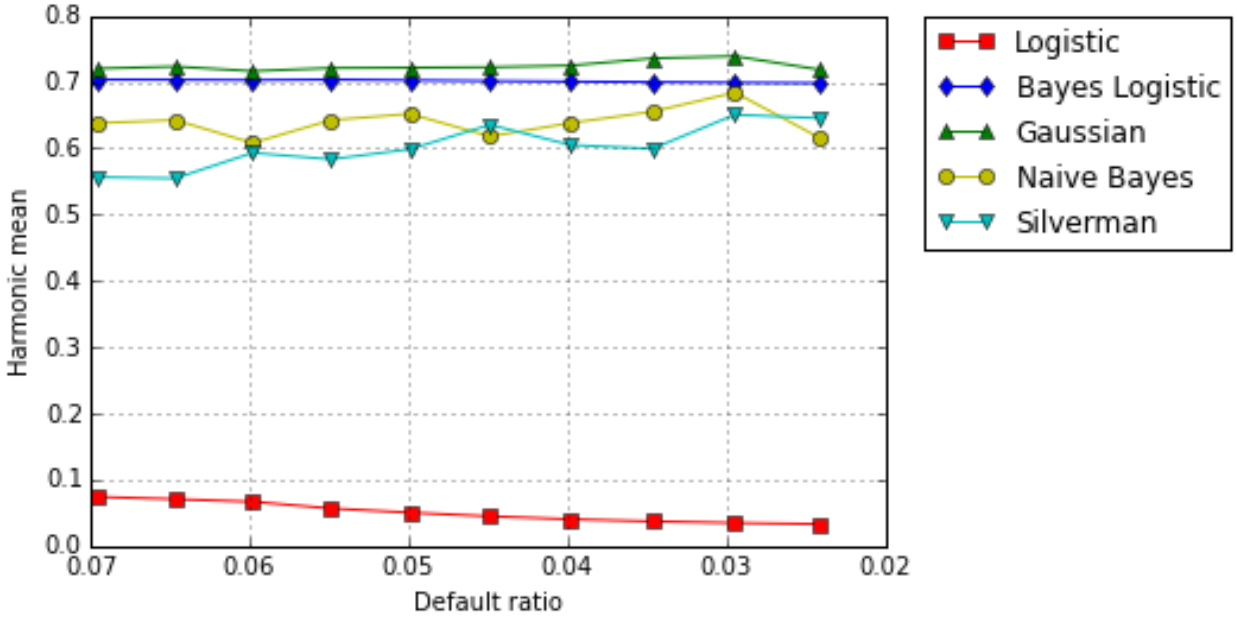

Figure 5.44: Performance of parametric versus non-parametric classifiers: Lending club z-scored data

5.3 Conclusion

This chapter investigated the effect of the frequentist as well as the Bernoulli priors on the performance of classifiers in the class imbalance setting. The chapter went further to compare the performance of parametric versus non-parametric classifiers in the class imbalance setting. Section 5.2.1 clearly illustrated that the use of an optimal Bernoulli prior can not only improve the general performance of the classifier, but also to some extent prevent the performance of the classifier deteriorating as the class imbalance increase. This was shown to be true for all investigated classifiers. It was shown that at worst the optimal Bernoulli prior will result in the same performance of the classifier when used without a prior.

Section 5.2.2 showed that the prior that optimise the harmonic mean of a classifier does not necessarily optimise the hit rate of that classifier. It is therefore the user's decision to decide whether it is of greater importance to optimise the hit rate or the harmonic mean. The section also illustrated that although there exists a relationship between the class imbalance and the optimal prior, that this relationship is not that given by the frequentist prior. Instead a Bernoulli prior with the prior parameter tuned for a specific class imbalance should be used.

In Section 5.2.3 parametric classifiers with optimal Bernoulli prior were compared to the non-parametric classifiers with optimal Bernoulli prior. The section showed that the parametric classifiers are outperformed by the non-parametric classifiers, thus showing the classification power of kernel density estimation. The only exception being the BLR classifier applied to the Lending club data, which outperformed the Silverman classifier. This emphasised the importance of the structure brought to the classifier, when data in one of the classes are

scarce, through the use of a prior distribution. There is thus a great desire for a Bayesian approach to kernel density estimation.

Chapter 6

Bayesian non-parametric classification

In Chapter 5 it was shown that the kernel density estimator performs exceptionally well as a classifier. However, there exhibits a desire for kernel density estimators to poses a predictive power. Bernardo (1999) observed this desire and developed an univariate approach to Bayesian kernel density estimation. Bernardo's method was extrapolated to the uncorrelated multivariate case (De Lima, Pereira, and Souza 2013). This chapter starts off by reviewing Bernardo's univariate method in Section 6.1 and then deriving a theoretical approach to the correlated multivariate method in Section 6.2.

6.1 Univariate bayesian kernel density estimation

6.1.1 Overview

This subsection serves to provide a step-by-step overview of the derivation of the univariate Bayesian kernel density estimation method, as developed by Bernardo (1999).
The steps are as follow:

1. The data is split into two random partitions. See Section 6.1.2.
2. The likelihood of the kernel density estimation is expressed in terms of the product of m mixtures of k inverse gamma distributions. See Equation 6.3 to Equation 6.9.
3. Since the product of various mixtures of inverse gamma distributions are complicated to work with, the mixture of inverse gamma distributions are approximated by a single inverse gamma distribution.
    (a) An expression for the Kullback-Leiber divergence between an inverse gamma distribution and some density function $g(h)$ is obtained. See Equation 6.10 to Equation 6.14.

(b) The expressions for $\mathbb{E}[\ln(h)]$ and $\mathbb{E}[h^{-1}]$, that result in the minimum Kullback-Leiber divergence is obtained. See Equation 6.15 to Equation 6.17.

(c) The function $g(h)$ is set equal to a mixture of inverse gamma distributions and the expeced values in terms of this function $g(h)$ of $\ln(h)$ and h^{-1} is equated to the abovementioned expected values. Thereby approximating the mixture of inverse gamma distributions with a single inverse gamma distribution. See Equation 6.18 to Equation 6.34.

4. The approximation of the mixture of inverse gamma distributions with a single inverse gamma distribution is substitued into the expression for the likelihood. Thereby expressing the likelihood in terms of the product of inverse gamma distributions. See Equation 6.35 to Equation 6.40.

5. The approximate reference distribution is calculated

   (a) The reference prior is determined.

      i. The approximate maximum likelihood estimate of h is calculated. See Equation 6.41 to Equation 6.44.

      ii. The approximate likelihood is rewritten in terms of the maximum likelihood estimate. See Equation 6.45 to Equation 6.47.

      iii. The asymptotic posterior distribution of h is axpproximated for some function $u(h)$. See Equation 6.48 to Equation 6.54.

   (b) Bayes' rule is applied to the reference prior and the likelihood to obtain the approximate posterior distribution. See Equation 6.55 to Equation 6.57.

6. The approximate reference predictive distribution is calculated. See Equation 6.58 to Equation 6.67.

7. The approximate reference predictive distribution is calculated for n_p random partitions of the data, $\mathbf{X}$, and averaged over. See Equation 6.68.

8. Repeat the process for various number of mixture components, k, and choose the value of k that minimises the entropy. See Equation 6.69 and Equation 6.70.

6.1.2 Preliminaries

Consider the data $\mathbf{x} = \{x_1, \ldots, x_n\}$ consisting of n univariate instances. Let $\mathbf{x}^{(tr)} = \{x_1^{(tr)}, \ldots, x_k^{(tr)}\}$, for some $0 < k < n$, be a subset, of size k, of $\mathbf{x}$. Let $\mathbf{x}^{(te)} = \{x_1^{(te)}, \ldots, x_m^{(te)}\}$ be a subset, of size $m = n - k$, of $\mathbf{x}$. Note that $\mathbf{x} = \{\mathbf{x}^{(te)}, \mathbf{x}^{(tr)}\}$ so that $\mathbf{x}^{(te)}$ and $\mathbf{x}^{(tr)}$ are non-overlapping

subsets of $\mathbf{x}$. Let $q(.)$ be a kernel function, then using kernel density estimation a reasonable approximation to density of $\mathbf{X}^{(te)}$ given the bandwidth, h, may be obtained

$$\begin{aligned} p\left(\mathbf{x}^{(te)}|h\right) &= \prod_{i=1}^{m} p\left(x_i^{(te)}|h\right) \\ &\approx \prod_{i=1}^{m}\sum_{j=1}^{k} q\left(x_i^{(te)}|x_j^{(tr)},h\right) \end{aligned}$$

Thus, using Bayes' rule the posterior distribution of the bandwidth is

$$\begin{aligned} p\left(h|\mathbf{x}^{(tr)},\mathbf{x}^{(te)}\right) &\propto p(h)\, p\left(\mathbf{x}^{(te)}|\mathbf{x}^{(tr)},h\right) \\ &= p(h)\prod_{i=1}^{m}\sum_{j=1}^{k} q\left(x_i^{(te)}|x_j^{(tr)},h\right) \end{aligned}$$

The posterior predictive distribution for some data point x is thus given by

$$\begin{aligned} p\left(x|\mathbf{x}^{(tr)},\mathbf{x}^{(te)}\right) &= \int_{\infty}^{\infty} p(x|h)\, p\left(h|\mathbf{x}^{(tr)},\mathbf{x}^{(te)}\right) dh \\ &\approx \int_{\infty}^{\infty} \frac{1}{k}\sum_{j=1}^{k} q\left(x|x_j^{(tr)},h\right) p\left(h|\mathbf{x}^{(tr)},\mathbf{x}^{(te)}\right) dh \\ &= \frac{1}{k}\sum_{j=1}^{k}\int_{\infty}^{\infty} q\left(x|x_j^{(tr)},h\right) p\left(h|\mathbf{x}^{(tr)},\mathbf{x}^{(te)}\right) dh \end{aligned} \tag{6.1}$$

Equation 6.1 can be viewed as the average of the k integrated kernels with respect to the posterior distribution of the smoothing parameter h. Since this is true for any partition $\mathbf{x} = \{\mathbf{x}^{(te)}, \mathbf{x}^{(tr)}\}$ it follows that an estimate for posterior predictive distribution may be calculated as

$$p(x|k,\mathbf{x}) = \frac{1}{n_p}\sum_{l=1}^{n_p} p\left(x|\mathbf{x}_{(l)}^{(te)},\mathbf{x}_{(l)}^{(tr)}\right) \tag{6.2}$$

where n_p is the number of partitions with the form $\mathbf{x} = \{\mathbf{x}^{(te)}, \mathbf{x}^{(tr)}\}$. The number of partitions are chosen to be the same as the number of observations in the data set, that is to say $n_p = n$. The notation $\mathbf{x}_{(l)}^{(te)}$ and $\mathbf{x}_{(l)}^{(tr)}$ are used to denote the l^{th} partition of the $\mathbf{x}^{(te)}$ subset and $\mathbf{x}^{(tr)}$ subset, respectively.

The value of k is of utter importance. In Section 6.1.3 it will be shown that the likelihood is proportional to the product of m mixture models, each consisting of k inverted gamma distributions. Therefore, the value of k not only represent the size of the subset $\mathbf{x}^{(tr)}$, but also represent the number of mixture components in each of the m mixture models. It makes sense that the final posterior predictive distribution explicitly depends on the value of k. The choice of k will be addressed in Section 6.1.6. It is clear that the posterior predictive distribution and the posterior distribution of the bandwidth is dependent on the choice of the kernel function. The following sections assume a univariate normal kernel function.

6.1.3 Approximate likelihood

This section elaborates on the calculations required to obtain the approximate likelihood function of the data. The section is started off by expressing the Gaussian kernel function, found in the expression for the likelihood, in terms of an inverse gamma function. It is therefore shown that the likelihood is proportional to the product of the mixture of inverse gamma functions. This is done in Equation 6.3 to Equation 6.9. In order to simplify the expression for the likelihood, the mixture of inverse gamma functions (found within the likelihood expression) is approximated with a single inverse gamma function. This approximation is performed using the Kullback-Leibler divergence. This is done in Equation 6.10 to Equation 6.34. Finally in Equation 6.35 to Equation 6.40 the obtained approximation to the mixture of inverse gamma functions is substituted back into the likelihood function, resulting in the desired approximate likelihood.

Consider the probability density function of the inverse gamma distribution,

$$Ig\left(h|\alpha,\beta\right)=\frac{\beta^{\alpha}}{\Gamma\left(\alpha\right)}h^{-\alpha-1}e^{-\frac{\beta}{h}}$$

with $\alpha>0$ and $\beta>0$. It is possible to write the likelihood function in terms of the inverse gamma distribution,

$$L\left(h|\alpha,\beta\right)=\prod_{i=1}^{m}\frac{1}{k}\sum_{j=1}^{k}q\left(x_i^{(te)}|x_j^{(tr)},h\right) \tag{6.3}$$

$$=\prod_{i=1}^{m}\frac{1}{k}\sum_{j=1}^{k}\frac{1}{\sqrt{2\pi h}}e^{-\frac{\left(x_i^{(te)}-x_j^{(tr)}\right)^2}{2h}} \tag{6.4}$$

$$=\prod_{i=1}^{m}\frac{1}{k}\sum_{j=1}^{k}\frac{h}{\left(x_i^{(te)}-x_j^{(tr)}\right)}\frac{\left(x_i^{(te)}-x_j^{(tr)}\right)}{\sqrt{2}\Gamma\left(\frac{1}{2}\right)}h^{-\frac{3}{2}}e^{-\frac{\left(x_i^{(te)}-x_j^{(tr)}\right)^2}{2h}} \tag{6.5}$$

$$=\prod_{i=1}^{m}\frac{1}{k}\sum_{j=1}^{k}\frac{h}{\left(x_i^{(te)}-x_j^{(tr)}\right)}Ig\left(h|\frac{1}{2},\frac{\left(x_i^{(te)}-x_j^{(tr)}\right)^2}{2}\right) \tag{6.6}$$

$$\propto\prod_{i=1}^{m}\sum_{j=1}^{k}\frac{h}{\left(x_i^{(te)}-x_j^{(tr)}\right)}Ig\left(h|\frac{1}{2},\frac{\left(x_i^{(te)}-x_j^{(tr)}\right)^2}{2}\right) \tag{6.7}$$

$$=\prod_{i=1}^{m}\sum_{j=1}^{k}\frac{h}{\sqrt{d_{ij}}}Ig\left(h|\frac{1}{2},\frac{d_{ij}}{2}\right) \tag{6.8}$$

$$\propto h^m\prod_{i=1}^{m}\sum_{j=1}^{k}w_{ij}Ig\left(h|\frac{1}{2},\frac{d_{ij}}{2}\right) \tag{6.9}$$

where $w_{ij} = \left(\sqrt{d_{ij}} \sum_{j=1}^{k} \frac{1}{\sqrt{d_{ij}}}\right)^{-1}$ and Equation 6.5 follows from the fact that $\Gamma\left(\frac{1}{2}\right) = \sqrt{\pi}$. This shows that the likelihood function can be written in such a way that it is proportional to the product of m mixture models. These mixture models consist out of k inverse gamma functions with parameters $\alpha = \frac{1}{2}$, $\beta_{ij} = \frac{d_{ij}}{2} = \frac{\left(x_i^{(te)} - x_j^{(tr)}\right)^2}{2}$ and weights equal to $w_{ij} = \left(\sqrt{d_{ij}} \sum_{j=1}^{k} \frac{1}{\sqrt{d_{ij}}}\right)^{-1}$.

In order to approximate the above mentioned mixture of inverse gamma distributions, the Kulback-Leibler divergence between the mixture and an inverse gamma distribution $Ig\left(h|, \alpha, \beta\right)$ needs to be minimised. The Kullback-Leibler divergence between the inverse gamma distribution $Ig\left(h|, \alpha, \beta\right)$ and some density $g(h)$ is given by

$$\begin{aligned}
\delta\left(\alpha, \beta\right) &= \int_0^\infty g\left(h\right) \ln\left(\frac{g\left(h\right)}{Ig\left(h|\alpha, \beta\right)}\right) dh && (6.10)\\
&= \int_0^\infty g\left(h\right) \ln g\left(h\right) dh - \int_0^\infty g\left(h\right) \ln Ig\left(h|\alpha, \beta\right) dh && (6.11)\\
&= \int_0^\infty g\left(h\right) \ln g\left(h\right) dh - \int_0^\infty g\left(h\right) \ln\left(\frac{\beta^\alpha}{\Gamma\left(\alpha\right)} h^{-\alpha-1} e^{-\frac{\beta}{h}}\right) dh && (6.12)\\
&= \int_0^\infty g\left(h\right) \ln g\left(h\right) dh - \alpha \ln\left(\beta\right) \int_0^\infty g\left(h\right) dh + \ln\left(\Gamma\left(\alpha\right)\right) \int_0^\infty g\left(h\right) dh \\
&\quad + \left(\alpha + 1\right) \int_0^\infty g\left(h\right) \ln\left(h\right) dh + \beta \int_0^\infty g\left(h\right) h^{-1} dh && (6.13)\\
&= c - \alpha \ln\left(\beta\right) + \ln\left(\Gamma\left(\alpha\right)\right) + \left(\alpha + 1\right) \mathbf{E}\left[\ln\left(h\right)\right] + \beta \mathbf{E}\left[h^{-1}\right] && (6.14)
\end{aligned}$$

where c is some irrelevant constant. Equation 6.14 follows from the fact that since $g(h)$ is a density function, $\int_0^\infty g\left(h\right) dh = 1$. In order to minimise the Kullback-Leibler divergence we differentiate with respect to the respective parameters and set the equations equal to zero.

$$\begin{aligned}
\frac{\partial \delta\left(\alpha, \beta\right)}{\alpha} &= -\ln(\beta) + \psi\left(\alpha\right) + \mathbf{E}\left[\ln\left(h\right)\right] = 0 && (6.15)\\
\frac{\partial \delta\left(\alpha, \beta\right)}{\beta} &= -\frac{\alpha}{\beta} + \mathbf{E}\left[h^{-1}\right] = 0 && (6.16)
\end{aligned}$$

where $\psi\left(.\right) = \frac{\Gamma`(.)}{\Gamma(.)}$ is the digamma function. Solving Equation 6.15 and Equation 6.16 it is seen that the Kullback-Leilber divergence is minimised if, and only if

$$\mathbf{E}\left[\ln\left(h\right)\right] = \ln(\beta) - \psi\left(\alpha\right), \qquad \mathbf{E}\left[h^{-1}\right] = \frac{\alpha}{\beta} \qquad (6.17)$$

Let $g\left(h\right) = \sum_j \rho_j Ig\left(h|\frac{1}{2}, \beta_j\right)$ be a mixture of inverse gamma distributions with weights ρ_k. The best approximation to this mixture of inverse gamma distributions by a single inverse

gamma distribution $Ig\left(h|\alpha,\beta\right)$ is obtained by matching the expected values of $\ln\left(h\right)$ and h^{-1}

$$\mathbf{E}_g\left[h^{-1}\right] = \sum_j \rho_j \frac{1}{2\beta_j} \tag{6.18}$$

$$\mathbf{E}_g\left[\ln\left(h\right)\right] = \sum_j \rho_j \left\{\ln\left(\beta_j\right) - \psi\left(\frac{1}{2}\right)\right\} \tag{6.19}$$

$$= \sum_j \rho_j \ln(\beta_j) - \psi\left(\frac{1}{2}\right) \tag{6.20}$$

Equation 6.20 follows from the fact that $\sum_j \rho_j = 1$. Equating the equations in Equation 6.17 to Equation 6.20 and Equation 6.18 respectively we obtain

$$\mathbf{E}\left[h^{-1}\right] = \mathbf{E}_g\left[h^{-1}\right] \tag{6.21}$$

$$\frac{\alpha}{\beta} = \sum_j \rho_j \frac{1}{2\beta_j} \tag{6.22}$$

$$\beta = 2\alpha \left(\sum_j \rho_j \frac{1}{\beta_j}\right)^{-1} \tag{6.23}$$

and

$$\mathbf{E}\left[\ln\left(h\right)\right] = \mathbf{E}_g\left[\ln\left(h\right)\right] \tag{6.24}$$

$$\ln(\beta) - \psi\left(\alpha\right) = \sum_j \rho_j \ln(\beta_j) - \psi\left(\frac{1}{2}\right) \tag{6.25}$$

$$\ln\left(2\alpha\left(\sum_j \rho_j \frac{1}{\beta_j}\right)^{-1}\right) - \psi\left(\alpha\right) = \sum_j \rho_j \ln(\beta_j) - \psi\left(\frac{1}{2}\right) \tag{6.26}$$

$$\ln\left(\alpha\right) - \psi\left(\alpha\right) = -\ln\left(2\left(\sum_j \rho_j \frac{1}{\beta_j}\right)^{-1}\right) + \sum_j \rho_j \ln(\beta_j) - \psi\left(\frac{1}{2}\right) \tag{6.27}$$

$$\ln\left(\alpha\right) - \psi\left(\alpha\right) = \ln\left(\frac{1}{2}\right) - \ln\left(\left(\sum_j \rho_j \frac{1}{2\beta_j}\right)^{-1}\right) + \ln\left(e^{\sum_j \rho_j \ln(\beta_j)}\right) - \psi\left(\frac{1}{2}\right) \tag{6.28}$$

$$\ln\left(\alpha\right) - \psi\left(\alpha\right) = \ln\left(\frac{1}{2}\right) - \psi\left(\frac{1}{2}\right) + \ln\left(\frac{e^{\sum_j \rho_j \ln(\beta_j)}}{\left(\sum_j \rho_j \frac{1}{\beta_j}\right)^{-1}}\right) \tag{6.29}$$

Now consider the asymptotic series expansion of the digamma function for some $x \in \mathbb{R}$

$$\psi\left(x\right) = \ln(x) - \frac{1}{2x} - \sum_{n=1}^{\infty} \frac{B_{2n}}{2nx^{2n}} \tag{6.30}$$

where B_s is the s^{th} Bernoulli number. The digamma function can therefore be approximated by

$$\psi\left(x\right) \approx \ln(x) - \frac{1}{2x} \tag{6.31}$$

Using this approximation to the digamma function an approximate solution can be obtained for Equation 6.29

$$\frac{1}{2\alpha} \approx 1 + \ln\left(\frac{e^{\sum_j \rho_j \ln(\beta_j)}}{\left(\sum_j \rho_j \frac{1}{\beta_j}\right)^{-1}}\right) \tag{6.32}$$

$$\alpha \approx \frac{1}{2}\left(1 + \ln\left(\frac{e^{\sum_j \rho_j \ln(\beta_j)}}{\left(\sum_j \rho_j \frac{1}{\beta_j}\right)^{-1}}\right)\right)^{-1} \tag{6.33}$$

Substituting 6.33 into 6.23 we obtain

$$\beta \approx \left(1 + \ln\left(\frac{e^{\sum_j \rho_j \ln(\beta_j)}}{\left(\sum_j \rho_j \frac{1}{\beta_j}\right)^{-1}}\right)\right)^{-1}\left(\sum_j \rho_j \frac{1}{\beta_j}\right)^{-1} \tag{6.34}$$

The likelihood can therefore now be expressed as

$$L\left(h, \mathbf{x}^{(te)}, \mathbf{x}^{(tr)}\right) \propto h^m \prod_{i=1}^{m}\sum_{j=1}^{k} w_{ij} Ig\left(h|\frac{1}{2}, \frac{d_{ij}}{2}\right) \tag{6.35}$$

$$\approx h^m \prod_{i=1}^{m} Ig\left(h|a_i, b_i\right) \tag{6.36}$$

$$= h^m \prod_{i=1}^{m} \frac{b_i^{a_i}}{\Gamma\left(a_i\right)} h^{-a_i-1} e^{-\frac{b_i}{h}} \tag{6.37}$$

$$\propto h^m \prod_{i=1}^{m} h^{-a_i-1} e^{-\frac{b_i}{h}} \tag{6.38}$$

$$= h^{-\sum_{i=1}^{m} a_i} e^{-\frac{\sum_{i=1}^{m} b_i}{h}} \tag{6.39}$$

where

$$a_i = \frac{1}{2}\left(1 + \ln\left(\frac{e^{\sum_j w_{ij} \ln(d_{ij})}}{\left(\sum_j w_{ij} \frac{1}{d_{ij}}\right)^{-1}}\right)\right)^{-1}, \quad b_i = \frac{1}{2}\left(1 + \ln\left(\frac{e^{\sum_j w_{ij} \ln(d_{ij})}}{\left(\sum_j w_{ij} \frac{1}{d_{ij}}\right)^{-1}}\right)\right)^{-1}\left(\sum_j w_{ij} \frac{1}{d_{ij}}\right)^{-1} \tag{6.40}$$

6.1.4 Approximate reference distribution

The approximate likelihood can be used to help determine a reference distribution for the parameter h. The approximate maximum likelihood estimate of h can be calculated as

$$L\left(h, \mathbf{x}^{(te)}, \mathbf{x}^{(tr)}\right) \propto h^{-\sum_{i=1}^{m} a_i} e^{-\frac{\sum_{i=1}^{m} b_i}{h}} \tag{6.41}$$

$$\ln\left(L\left(h, \mathbf{x}^{(te)}, \mathbf{x}^{(tr)}\right)\right) \propto -m\bar{a}\ln(h) - \frac{\bar{b}}{h} \tag{6.42}$$

$$\frac{\partial}{\partial h}\ln\left(L\left(h, \mathbf{x}^{(te)}, \mathbf{x}^{(tr)}\right)\right) \propto \frac{-m\bar{a}}{h} + \frac{m\bar{b}}{h^2} = 0 - m\bar{a}h = m\bar{b} \tag{6.43}$$

$$\hat{h} = \frac{\bar{b}}{\bar{a}} \tag{6.44}$$

where $\bar{a} = m^{-1}\sum_{i=1}^{m} a_i$ and $\bar{b} = m^{-1}\sum_{i=1}^{m} b_i$. The likelihood function can now be rewritten in terms of the approximate maximum likelihood estimate, $\hat{h} = \frac{\bar{b}}{\bar{a}}$

$$L\left(h, \mathbf{x}^{(te)}, \mathbf{x}^{(tr)}\right) = \frac{\left(m\bar{b}\right)^{m\bar{a}}}{\Gamma\left(\bar{a}m\right)} h^{-(m\bar{a}+1)} e^{\frac{-m\bar{b}}{h}} \tag{6.45}$$

$$L\left(h, \mathbf{x}^{(te)}, \mathbf{x}^{(tr)}\right) = \frac{\left(m\bar{a}\frac{\bar{b}}{\bar{a}}\right)^{m\bar{a}}}{\Gamma\left(\bar{a}m\right)} h^{-(m\bar{a}+1)} e^{\frac{-m\bar{a}\frac{\bar{b}}{\bar{a}}}{h}} \tag{6.46}$$

$$L\left(h, \mathbf{x}^{(te)}, \mathbf{x}^{(tr)}\right) = \frac{\left(m\bar{a}\hat{h}\right)^{m\bar{a}}}{\Gamma\left(\bar{a}m\right)} h^{-(m\bar{a}+1)} e^{\frac{-m\bar{a}\hat{h}}{h}} \tag{6.47}$$

The asymptotic posterior distribution of h can now be approximated, for some positive function $u(h)$, using Bayes' rule

$$\pi\left(h|\hat{h}\right) \propto \frac{\left(m\bar{a}\hat{h}\right)^{m\bar{a}}}{\Gamma\left(\bar{a}m\right)} h^{-(m\bar{a}+1)} e^{\frac{-m\bar{a}\hat{h}}{h}} u\left(h\right) \tag{6.48}$$

Suppose for instance $u\left(h\right) = 1$, then the posterior reduces to

$$\pi\left(h|\hat{h}\right) \propto \frac{\left(m\bar{a}\hat{h}\right)^{m\bar{a}}}{\Gamma\left(\bar{a}m\right)} h^{-(m\bar{a}+1)} e^{\frac{-m\bar{a}\hat{h}}{h}} \tag{6.49}$$

It therefore follows that

$$f\left(h\right) = \pi\left(h|\hat{h}\right)|_{\hat{h}=h} \tag{6.50}$$

$$= \frac{\left(m\bar{a}h\right)^{m\bar{a}}}{\Gamma\left(\bar{a}m\right)} h^{-(m\bar{a}+1)} e^{\frac{-m\bar{a}h}{h}} \tag{6.51}$$

$$= \frac{\left(m\bar{a}\right)^{m\bar{a}}}{\Gamma\left(\bar{a}m\right)} h^{-1} e^{-m\bar{a}} \tag{6.52}$$

$$= ch^{-1} \tag{6.53}$$

for some constant c. From Equation 6.53 it is clear that the reference prior is given by

$$\pi\left(h\right) = h^{-1} \tag{6.54}$$

The approximate reference posterior distribution is finally obtained by applying Bayes' theorem to Equation 6.54 and Equation 6.39

$$\pi\left(h|\mathbf{x}^{(te)},\mathbf{x}^{(tr)}\right) \propto \pi\left(h|\hat{h}\right) L\left(h,\mathbf{x}^{(te)},\mathbf{x}^{(tr)}\right) \tag{6.55}$$

$$\propto h^{-1}h^{-\sum_{i=1}^{m} a_i} e^{-\frac{\sum_{i=1}^{m} b_i}{h}} \tag{6.56}$$

$$= h^{-m\bar{a}-1} e^{-\frac{m\bar{b}}{h}} \tag{6.57}$$

The approximate reference posterior distribution is an inverse gamma distribution with parameters $m\bar{a}$ and $m\bar{b}$, that is $Ig\left(h|m\bar{a},m\bar{b}\right)$.

6.1.5 Approximate reference predictive distribution

The approximate reference posterior distribution is used to calculate the approximate reference predictive distribution. In order to calculate the approximate reference predictive distribution the fact that a normal distribution with variance following an inverse gamma

distribution results in a Student t distribution, is proven.

$$\pi\left(x|\mathbf{x}^{(te)},\mathbf{x}^{(tr)}\right)=\int_0^\infty\frac{1}{k}\sum_{j=1}^{k}N\left(x|x_j^{(tr)},h\right)Ig\left(h|m\bar{a},m\bar{b}\right)dh \tag{6.58}$$

$$=\frac{1}{k}\sum_{j=1}^{k}\int_0^\infty N\left(x|x_j^{(tr)},h\right)Ig\left(h|m\bar{a},m\bar{b}\right)dh \tag{6.59}$$

$$=\frac{1}{k}\sum_{j=1}^{k}\int_0^\infty\frac{1}{\sqrt{2\pi h}}e^{-\frac{\left(x-x_j^{(tr)}\right)^2}{2h}}\frac{\left(m\bar{b}\right)^{m\bar{a}}}{\Gamma\left(m\bar{a}\right)}h^{-(m\bar{a}+1)}e^{-\frac{m\bar{b}}{h}}dh \tag{6.60}$$

$$=\frac{1}{k}\sum_{j=1}^{k}\int_0^\infty\frac{\left(m\bar{b}\right)^{m\bar{a}}}{\sqrt{2\pi}\Gamma\left(m\bar{a}\right)}h^{-(m\bar{a}+1)-\frac{1}{2}}e^{-\frac{\left(x-x_j^{(tr)}\right)^2+2m\bar{b}}{2h}}dh \tag{6.61}$$

$$=\frac{1}{k}\sum_{j=1}^{k}\frac{\left(m\bar{b}\right)^{m\bar{a}}}{\sqrt{2\pi}\Gamma\left(m\bar{a}\right)}\left(\frac{\left(x-x_j^{(tr)}\right)^2+2m\bar{b}}{2}\right)^{-\left(m\bar{a}+\frac{1}{2}\right)}\Gamma\left(m\bar{a}+\frac{1}{2}\right)$$
$$\times\int_0^\infty\left(\frac{\left(x-x_j^{(tr)}\right)^2+2m\bar{b}}{2}\right)^{m\bar{a}+\frac{1}{2}}\frac{1}{\Gamma\left(m\bar{a}+\frac{1}{2}\right)}h^{-(m\bar{a}+\frac{1}{2}+1)}e^{-\frac{\left(x-x_j^{(tr)}\right)^2+2m\bar{b}}{2h}}dh \tag{6.62}$$

$$=\frac{1}{k}\sum_{j=1}^{k}\frac{\left(m\bar{b}\right)^{m\bar{a}}}{\sqrt{2\pi}\Gamma\left(m\bar{a}\right)}\left(\frac{\left(x-x_j^{(tr)}\right)^2+2m\bar{b}}{2}\right)^{-\left(m\bar{a}+\frac{1}{2}\right)}\Gamma\left(m\bar{a}+\frac{1}{2}\right) \tag{6.63}$$

$$=\frac{1}{k}\sum_{j=1}^{k}\frac{\Gamma\left(\frac{2m\bar{a}+1}{2}\right)\left(m\bar{b}\right)^{m\bar{a}+\frac{1}{2}}}{\sqrt{2m\bar{a}\pi}\Gamma\left(\frac{2m\bar{a}}{2}\right)\left(\frac{m\bar{b}}{m\bar{a}}\right)^{\frac{1}{2}}}\left(m\bar{b}+\frac{\left(x-x_j^{(tr)}\right)^2}{2}\right)^{-\left(m\bar{a}+\frac{1}{2}\right)} \tag{6.64}$$

$$=\frac{1}{k}\sum_{j=1}^{k}\frac{\Gamma\left(\frac{2m\bar{a}+1}{2}\right)}{\sqrt{2m\bar{a}\pi}\Gamma\left(\frac{2m\bar{a}}{2}\right)\left(\frac{m\bar{b}}{m\bar{a}}\right)^{\frac{1}{2}}}\left(1+\frac{\left(x-x_j^{(tr)}\right)^2}{2m\bar{b}}\right)^{-\left(m\bar{a}+\frac{1}{2}\right)} \tag{6.65}$$

$$=\frac{1}{k}\sum_{j=1}^{k}\frac{\Gamma\left(\frac{2m\bar{a}+1}{2}\right)}{\sqrt{2m\bar{a}\pi}\Gamma\left(\frac{2m\bar{a}}{2}\right)\left(\frac{m\bar{b}}{m\bar{a}}\right)^{\frac{1}{2}}}\left(1+\frac{1}{2m\bar{a}}\frac{\left(x-x_j^{(tr)}\right)^2}{\frac{m\bar{b}}{m\bar{a}}}\right)^{-\left(m\bar{a}+\frac{1}{2}\right)} \tag{6.66}$$

$$=\frac{1}{k}\sum_{j=1}^{k}t_{2m\bar{a}}\left(x|x_j^{(tr)},\frac{\bar{b}}{\bar{a}}\right) \tag{6.67}$$

The reference predictive distribution is therefore approximated by a mixture of Student t distributions, each centred at $x_j^{(tr)}$. The non-central parameter $\frac{\bar{b}}{\bar{a}}$ in the mixture model has the same function as the bandwidth in traditional kernel density estimation.

Note that Equation 6.62 follows from the fact that

$$\int_0^\infty \left(\frac{\left(x - x_j^{(tr)}\right)^2 + 2m\bar{b}}{2} \right)^{\alpha+\frac{1}{2}} \frac{1}{\Gamma\left(m\bar{a} + \frac{1}{2}\right)} h^{-(m\bar{a}+\frac{1}{2}+1)} e^{-\frac{\left(x-x_j^{(tr)}\right)^2+2m\bar{b}}{2h}} dh$$
$$= \int_0^\infty Ig\left(h|m\bar{a} + \frac{1}{2}, \frac{\left(x - x_j^{(tr)}\right)^2 + 2m\bar{b}}{2} \right) dh$$
$$= 1$$

Calculating the approximate reference predictive distribution for each of the n_p random partitions of the form $\mathbf{x} = \{\mathbf{x}_{(l)}^{(te)}, \mathbf{x}_{(l)}^{(tr)}\}$ with $l = 1, \ldots, n_p$, the desired model is obtained by substituting Equation 6.67 into Equation 6.2.

$$\begin{aligned} p\left(x|k, \mathbf{x}\right) &= \frac{1}{n_p} \sum_{l=1}^{n_p} p\left(x|\mathbf{x}_{(l)}^{(te)}, \mathbf{x}_{(l)}^{(tr)}\right) \\ &= \frac{1}{n_p} \sum_{l=1}^{n_p} \frac{1}{k} \sum_{j=1}^{k} t_{2m\bar{a}_{(l)}} \left(x|x_{j(l)}^{(tr)}, \frac{\bar{b}_{(l)}}{\bar{a}_{(l)}} \right) \end{aligned} \tag{6.68}$$

where $\bar{a}_{(l)}$ and $\bar{b}_{(l)}$ are the respective parameters $\bar{a}$ and $\bar{b}$ calculated using the l^{th} random partition of $\mathbf{x}$. It is important to note that for each of the n_p partitions the value of k remains fixed.

6.1.6 Number of mixture components, k

As previously mentioned the choice of k, the number of instances in the $\mathbf{x}^{(te)}$ subset and hence the number of mixture components, is of great importance. Bernardo suggests that the expected utility with form

$$u\left(\hat{p}\right) = a \int_X p\left(x\right) \ln\left[\hat{p}\left(x\right)\right] dx + b$$

with b and $a > 0$ some arbitrary constants, may be considered the appropriate method to evaluate the performance of the model. However, since the true model $p(x)$ is unknown and a random sample , $\mathbf{x} = \{x_1, \ldots, x_n\}$, from the model is known, the Monte Carlo approximation may be used

$$\hat{u}\left(\hat{p}\right) \approx a\frac{1}{n} \sum_{j=1}^{n} \ln\left[\hat{p}\left(x_j|\mathbf{x}_{-j}\right)\right] + b \tag{6.69}$$

with $\hat{p}\left(x_j|\mathbf{x}_{-j}\right)$ the predictive density of x_j given all the observations excluding x_j.
The expected utility is calculated for models with various values of k. The value of k resulting in the highest expected utility is chosen as the optimal value.

Since the values of a and b are arbitrary and the method is used in order to compare the performance of the fitted models, it may be argued that entropy calculated as

$$H\left(\hat{p}\right) = -\frac{1}{n}\sum_{j=1}^{n}\ln\left[\hat{p}\left(x_j|\mathbf{x}_{-j}\right)\right] \tag{6.70}$$

may be used instead of expected utility. This resolves the issue of selecting values for a and b. The value of k that results in the minimum entropy results in the optimal model.

6.1.7 Example

In order to illustrate the working of the abovementioned Bayesian kernel density estimation model, the example performed in (Bernardo 1999) is repeated with the added component of choosing k with respect to minising the entropy. Fourteen observations are generated using the mixture model $p\left(x\right) = 0.7N\left(x|0,1\right) + 0.3N\left(x|5,1\right)$ to form the data set

$$\mathbf{x} = \{-1.39, -0.85, -0.54, -0.32, -0.31, -0.30, -0.19, -0.02, 0.54, 3.65, 4.21, 4.30, 4.98, 5.51\}$$

The entropy of the model $\pi\left(x|k,\mathbf{x}\right)$ is evaluated for $k = 1,\ldots,12$. From this the model resulting in the minimum entropy is fitted to the data. Since the data set is small cross-validation is not performed.

Considering Table 6.1 it is interesting to note that based on the average sample entropy the optimal value for the partition size is $k = 8$, whereas based on the average utility as given in (Bernardo 1999) the optimal value for the partition size is $k = 7$. This may be explained by considering the standard deviations of both the average utility as well as the average sample entropy. The average sample entropy for $k = 8$ is -0.166638064 with a standard deviation of 0.00274308701. This suggests that a three standard deviation confidence interval would be $H(\hat{p}) \in (-0.174867325; -0.158408803)$. The confidence interval for the sample entropy when $k = 8$ suggests that $k = 6,\ldots,10$ may have also resulted in appropriate models.

k	$\bar{H}(\hat{p})$	$S_{H(\hat{p})}$	$\bar{u}$	s_u
1	-0.0770924261	0.00218547205	0.623	0.007
2	-0.114479455	0.00403995765	0.701	0.011
3	-0.134445024	0.0034220724	0.742	0.009
4	-0.149378097	0.00587592722	0.761	0.01
5	-0.156339656	0.00417289488	0.764	0.005
6	-0.162431812	0.00280058455	0.764	0.008
7	-0.164942029	0.00316037422	0.767	0.006
8	-0.166638064	0.00274308701	0.766	0.006
9	-0.165711269	0.00242677633	0.762	0.005
10	-0.162168165	0.00255138971	0.753	0.007
11	-0.155215316	0.00421163317	0.739	0.006
12	-0.145759578	0.00713781333	0.698	0.008

Table 6.1: Mean and standard deviations of the sample entropy as well as the utility, using 20 reference predictive estimates for each of the $k = 1, \ldots, 12$ partitions.

In Figure 6.1 the various estimated pdf's are compared to the actual mixture model. The figure clearly illustrates Silverman's rule of thumb (given in the cyan coloured dotted line) over fitting the data. This bandwidth estimate results in an under smoothed estimated density. Both the densities estimated using the Bayesian KDE method, with $k = 7$ and $k = 8$ respectively, perform very similar. It is seen that both these densities come much closer to estimating the actual density (represented by a thicker red line). It is seen that for this particular set of random partitioning of, $\mathbf{x} = \left(\mathbf{x}^{(\mathbf{te})}, \mathbf{x}^{(\mathbf{tr})}\right)$, the model with $k = 8$ (selected using average sample entropy) performs slightly better than the model with $k = 7$ (selected using average utility). This goes to show that using sample entropy to select the optimal value for k is a competitive alternative to using utility.

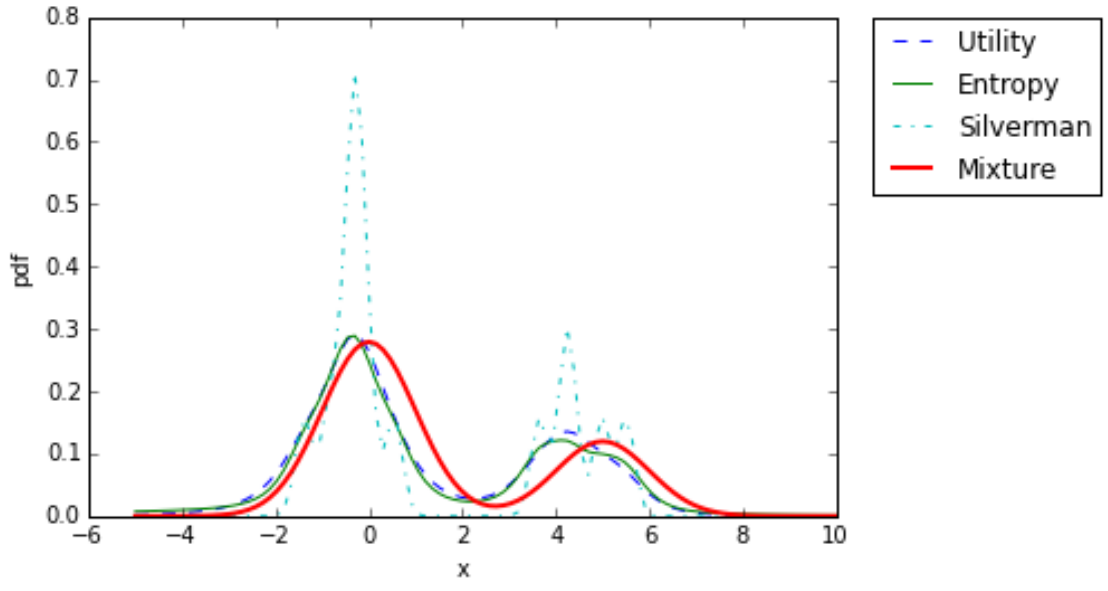

Figure 6.1: Estimated pdf's for the simulated data $\mathbf{x}$

6.2 Correlated multivariate Bayesian kernel density estimation

6.2.1 Overview

This subsection serves to provide a step-by-step overview of the derivation of the correlated multivariate Bayesian kernel density estimation method. It follows a similar approach as developed by Bernardo (1999) for the univariate case.
The steps are as follow:

1. The data is split into two random partitions. See Section 6.2.2.
2. The likelihood of the kernel density estimation is expressed in terms of the product of m mixtures of k inverse Wishart distributions. See Equation 6.73 to Equation 6.80.
3. Since the product of various mixtures of inverse Wishart distributions are complicated to work with, the mixture of inverse Wishart distributions are approximated by a single inverse Wishart distribution.
   (a) An expression for the Kullback-Leiber divergence between an inverse Wishart distribution and some density function $g(\mathbf{H})$ is obtained. See Equation 6.81 to Equation 6.85.
   (b) The expressions for $\mathbb{E}[\ln|\mathbf{H}|]$ and $\mathbb{E}[\mathbf{H}^{-1}]^T$, that result in the minimum Kullback-Leiber divergence is obtained. See Equation 6.86 to Equation 6.88.
   (c) The function $g(\mathbf{H})$ is set equal to a mixture of inverse Wishart distributions and the expeced values in terms of this function $g(\mathbf{H})$ of $\ln|\mathbf{H}|$ and $\mathbf{H}^{-1}$ is equated to the abovementioned expected values. Thereby approximating the mixture of inverse Wishart distributions with a single inverse Wishart distribution. See Equation 6.89 to Equation 6.102.
4. The approximation of the mixture of inverse Wishart distributions with a single inverse Wishart distribution is substitued into the expression for the likelihood. Thereby expressing the likelihood in terms of the product of inverse Wishart distributions. See Equation 6.103 to Equation 6.109.
5. The approximate posterior distribution is calculated
   (a) The prior is an inverse Wishart distribution. See Equation 6.110 and Equation 6.111.
   (b) Bayes' rule is applied to the prior and the likelihood to obtain the approximate posterior distribution. See Equation 6.112 to Equation 6.114.

6. The approximate posterior predictive distribution is calculated. See Equation 6.115 to Equation 6.124.

7. The approximate posterior predictive distribution is calculated for n_p random partitions of the data, $\mathbf{X}$, and averaged over. See Equation 6.125.

8. Repeat the process for various number of mixture components, k, and choose the value of k that minimises the entropy. See Equation 6.126.

6.2.2 Preliminaries

Consider the data $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ consisting of n instances and p features. Let $\mathbf{X}^{(tr)} = \{\mathbf{x}_1^{(tr)}, \ldots, \mathbf{x}_k^{(tr)}\}$,for some $0 < k < n$, be a subset, of size k, of $\mathbf{X}$. Let $\mathbf{X}^{(te)} = \{\mathbf{x}_1^{(te)}, \ldots, \mathbf{x}_m^{(te)}\}$ be a subset, of size $m = n - k$, of $\mathbf{X}$. Note that $\mathbf{X} = \{\mathbf{X}^{(te)}, \mathbf{X}^{(tr)}\}$ so that $\mathbf{X}^{(te)}$ and $\mathbf{X}^{(tr)}$ are non-overlapping subsets of $\mathbf{X}$. Let $q(.)$ be a kernel function, then using kernel density estimation a reasonable approximation to density of $\mathbf{X}^{(te)}$ given the bandwidth, $\mathbf{H}$, may be obtained

$$\begin{aligned} p\left(\mathbf{X}^{(te)}|\mathbf{H}\right) &= \prod_{i=1}^{m} p\left(\mathbf{x}_i^{(te)}|\mathbf{H}\right) \\ &\approx \prod_{i=1}^{m}\sum_{j=1}^{k} q\left(\mathbf{x}_i^{(te)}|\mathbf{x}_j^{(tr)}, \mathbf{H}\right) \end{aligned}$$

Thus, using Bayes' rule the posterior distribution of the bandwidth is

$$\begin{aligned} p\left(\mathbf{H}|\mathbf{X}^{(tr)}, \mathbf{X}^{(te)}\right) &\propto p(\mathbf{H})\, p\left((\mathbf{X}^{(te)}|(\mathbf{X}^{(tr)}, \mathbf{H}\right) \\ &= p(\mathbf{H}) \prod_{i=1}^{m}\sum_{j=1}^{k} q\left(\mathbf{x}_i^{(te)}|\mathbf{x}_j^{(tr)}, \mathbf{H}\right) \end{aligned}$$

The posterior predictive distribution for some data point $\mathbf{x}$ is thus given by

$$\begin{aligned} p\left(\mathbf{x}|\mathbf{X}^{(tr)}, \mathbf{X}^{(te)}\right) &= \int_{\mathbf{H}>0} p(\mathbf{x}|\mathbf{H})\, p\left(\mathbf{H}|\mathbf{X}^{(tr)}, \mathbf{X}^{(te)}\right) d\mathbf{H} \\ &\approx \int_{\mathbf{H}>0} \frac{1}{k}\sum_{j=1}^{k} q\left(\mathbf{x}|\mathbf{x}_j^{(tr)}, \mathbf{H}\right) p\left(\mathbf{H}|\mathbf{X}^{(tr)}, \mathbf{X}^{(te)}\right) d\mathbf{H} \\ &= \frac{1}{k}\sum_{j=1}^{k} \int_{\mathbf{H}>0} q\left(\mathbf{x}|\mathbf{x}_j^{(tr)}, \mathbf{H}\right) p\left(\mathbf{H}|\mathbf{X}^{(tr)}, \mathbf{X}^{(te)}\right) d\mathbf{H} \end{aligned} \tag{6.71}$$

Equation 6.71 can be viewed as the average of the k integrated kernels with respect to the posterior distribution of the bandwidth matrix $\mathbf{H}$. Since this is true for any partition $\mathbf{X} = \{\mathbf{X}^{(te)}, \mathbf{X}^{(tr)}\}$ it follows that an estimate for posterior predictive distribution may be calculated as

$$p(\mathbf{x}|k, \mathbf{X}) = \frac{1}{n_p}\sum_{l=1}^{n_p} p\left(\mathbf{x}|\mathbf{X}_{(l)}^{(te)}, \mathbf{X}_{(l)}^{(tr)}\right) \tag{6.72}$$

where n_p is the number of partitions with the form $\mathbf{X} = \{\mathbf{X}^{(te)}, \mathbf{X}^{(tr)}\}$. The number of partitions are chosen to be the same as the number of observations in the data set, that is to say $n_p = n$. The notation $\mathbf{x}_{(l)}^{(te)}$ and $\mathbf{X}_{(l)}^{(tr)}$ are used to denote the l^{th} partition of the $\mathbf{X}^{(te)}$ subset and $\mathbf{X}^{(tr)}$ subset, respectively.

The value of k is of utter importance. In Section 6.2.3 it will be shown that the likelihood is proportional to the product of m mixture models, each consisting of k inverse Wishart distributions. Therefore, the value of k not only represent the size of the subset $\mathbf{X}^{(tr)}$, but also represent the number of mixture components in each of the m mixture models. It makes sense that the final posterior predictive distribution explicitly depends on the value of k. The choice of k will be addressed in Section 6.2.6. It is clear that the posterior predictive distribution and the posterior distribution of the bandwidth is dependent on the choice of the kernel function. The following sections assume a multivariate normal kernel function.

6.2.3 Likelihood

Consider the probability density function of the inverted Wishart distribution:

$$W^{-1}\left(\mathbf{H}|\mathbf{\Psi},\nu\right) = \frac{\left|\mathbf{\Psi}\right|^{\frac{\nu}{2}}}{2^{\frac{\nu p}{2}}\Gamma_p\left(\frac{\nu}{2}\right)}\left|\mathbf{H}\right|^{-\frac{\nu+p+1}{2}} e^{-\frac{1}{2}\operatorname{Tr}\left(\mathbf{\Psi}\mathbf{H}^{-1}\right)}$$

The likelihood function may be expressed in terms of the inverse Wishart distribution as follows:

$$L\left(\mathbf{H}, \mathbf{X}^{(tr)}, \mathbf{X}^{(te)}\right) \tag{6.73}$$

$$= \prod_{i=1}^{m} \frac{1}{k} \sum_{j=1}^{k} q(\mathbf{x}_i^{(te)} | \mathbf{x}_j^{(tr)}, \mathbf{H}) \tag{6.74}$$

$$= \prod_{i=1}^{m} \frac{1}{k} \sum_{j=1}^{k} \frac{1}{(2\pi)^{\frac{p}{2}} |\mathbf{H}|^{\frac{1}{2}}} e^{-\frac{1}{2}\left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right)^T \mathbf{H}^{-1} \left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right)} \tag{6.75}$$

$$= \prod_{i=1}^{m} \frac{1}{k} \sum_{j=1}^{k} \frac{2^{\frac{p^2}{2}} \Gamma_p\left(\frac{p}{2}\right) |\mathbf{H}|^p}{(2\pi)^{\frac{p}{2}} \left| \left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right) \left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right)^T + \lambda \mathbf{I} \right|^{\frac{p}{2}}} \frac{\left| \left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right) \left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right)^T + \lambda \mathbf{I} \right|^{\frac{p}{2}}}{2^{\frac{p^2}{2}} \Gamma_p\left(\frac{p}{2}\right)}$$
$$\times |\mathbf{H}|^{-\frac{2p+1}{2}} e^{-\frac{1}{2} \operatorname{Tr}\left[\left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right)\left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right)^T \mathbf{H}^{-1}\right] - \frac{1}{2}\lambda \operatorname{Tr}\left(\mathbf{H}^{-1}\right) + \frac{1}{2}\lambda \operatorname{Tr}\left(\mathbf{H}^{-1}\right)} \tag{6.76}$$

$$= \prod_{i=1}^{m} \frac{1}{k} \sum_{j=1}^{k} \frac{2^{\frac{p^2}{2}} \Gamma_p\left(\frac{p}{2}\right) |\mathbf{H}|^p}{(2\pi)^{\frac{p}{2}} \left| \left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right) \left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right)^T + \lambda \mathbf{I} \right|^{\frac{p}{2}}} \frac{\left| \left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right) \left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right)^T + \lambda \mathbf{I} \right|^{\frac{p}{2}}}{2^{\frac{p^2}{2}} \Gamma_p\left(\frac{p}{2}\right)}$$
$$\times |\mathbf{H}|^{-\frac{2p+1}{2}} e^{-\frac{1}{2} \operatorname{Tr}\left\{\left[\left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right)\left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right)^T + \lambda \mathbf{I}\right] \mathbf{H}^{-1}\right\} + \frac{1}{2}\lambda \operatorname{Tr}\left(\mathbf{H}^{-1}\right)} \tag{6.77}$$

$$= \prod_{i=1}^{m} \frac{1}{k} \sum_{j=1}^{k} \frac{2^{\frac{p^2}{2}} \Gamma_p\left(\frac{p}{2}\right) |\mathbf{H}|^p \, e^{\frac{1}{2}\lambda \operatorname{Tr}\left(\mathbf{H}^{-1}\right)}}{(2\pi)^{\frac{p}{2}} \left| \left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right) \left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right)^T + \lambda \mathbf{I} \right|^{\frac{p}{2}}}$$
$$\times W^{-1}\left(\mathbf{H} \left| \left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right) \left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right)^T + \lambda \mathbf{I}, p \right.\right) \tag{6.78}$$

$$\propto \prod_{i=1}^{m} \sum_{j=1}^{k} \frac{|\mathbf{H}|^p \, e^{\frac{1}{2}\lambda \operatorname{Tr}\left(\mathbf{H}^{-1}\right)}}{\left|\boldsymbol{\Psi}_{ij}\right|^{\frac{p}{2}}} W^{-1}\left(\mathbf{H} | \boldsymbol{\Psi}_{ij}, p\right) \tag{6.79}$$

$$\propto |\mathbf{H}|^{mp} \, e^{\frac{1}{2}m\lambda \operatorname{Tr}\left(\mathbf{H}^{-1}\right)} \prod_{i=1}^{m} \sum_{j=1}^{k} w_{ij} W^{-1}\left(\mathbf{H} | \boldsymbol{\Psi}_{ij}, p\right) \tag{6.80}$$

where $\left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right) \left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right)^T$ is a $p \times p$ matrix, $\mathbf{I}$ is a $p \times p$ identity matrix and λ is some value that is larger than the absolute value of the most negative eigenvalue of the matrix $\left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right) \left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right)^T$. This ensures that the matrix $\left[\boldsymbol{\Psi_{ij}} = \left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right) \left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right)^T + \lambda \mathbf{I}\right]$ is a $p \times p$ positive definite matrix. The weights, encountered in the mixture model, are given by $w_{ij} = \frac{\left|\boldsymbol{\Psi_{ij}}\right|^{-\frac{p}{2}}}{\sum_{j=1}^{k} \left|\boldsymbol{\Psi_{ij}}\right|^{-\frac{p}{2}}}$. Since $\boldsymbol{\Psi_{ij}}$ is positive definite for all i and j, all of the weights w_{ij} are greater than zero. A special case may exist in which the data are of such a nature that the matrices $\left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right) \left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right)^T$ are positive definite for all i and j. In this case the reader is reffered to Appendix B.

The Kullback-Leiber divergence of a inverse Wishart distribution $W^{-1}\left(\mathbf{H} | \boldsymbol{\Psi}, \nu\right)$ from some

density function $g(\mathbf{H})$ is given by

$$\delta(\mathbf{\Psi},\nu) = \int_{\mathbf{H}>0} g(\mathbf{H}) \ln \frac{g(\mathbf{H})}{W^{-1}(\mathbf{H}|\mathbf{\Psi},\nu)} d\mathbf{H} \tag{6.81}$$

$$= \int_{\mathbf{H}>0} g(\mathbf{H}) \ln g(\mathbf{H})\, d\mathbf{H} - \int_{\mathbf{H}>0} g(\mathbf{H}) \ln W^{-1}(\mathbf{H}|\mathbf{\Psi},\nu)\, d\mathbf{H} \tag{6.82}$$

$$= \int_{\mathbf{H}>0} g(\mathbf{H}) \ln g(\mathbf{H})\, d\mathbf{H} - \int_{\mathbf{H}>0} g(\mathbf{H}) \ln \left(\frac{|\mathbf{\Psi}|^{\frac{\nu}{2}}}{2^{\frac{\nu p}{2}} \Gamma_p\left(\frac{\nu}{2}\right)} |\mathbf{H}|^{-\frac{\nu+p+1}{2}} e^{-\frac{1}{2}\operatorname{Tr}\left(\mathbf{\Psi}\mathbf{H}^{-1}\right)} \right) d\mathbf{H} \tag{6.83}$$

$$\begin{aligned} &= \int_{\mathbf{H}>0} g(\mathbf{H}) \ln g(\mathbf{H})\, d\mathbf{H} - \frac{\nu}{2} \int_{\mathbf{H}>0} g(\mathbf{H}) \ln |\mathbf{\Psi}|\, d\mathbf{H} + \int_{\mathbf{H}>0} \frac{\nu p}{2} g(\mathbf{H}) \ln 2 d\mathbf{H} \\ &+ \int_{\mathbf{H}>0} g(\mathbf{H}) \ln \Gamma_p\left(\frac{\nu}{2}\right) d\mathbf{H} + \frac{\nu+p+1}{2} \int_{\mathbf{H}>0} g(\mathbf{H}) \ln |\mathbf{H}|\, d\mathbf{H} \\ &+ \frac{1}{2} \int_{\mathbf{H}>0} g(\mathbf{H}) \operatorname{Tr}\left(\mathbf{\Psi}\mathbf{H}^{-1}\right) d\mathbf{H} \end{aligned} \tag{6.84}$$

$$= c - \frac{\nu}{2} \ln |\mathbf{\Psi}| + \frac{\nu p}{2} \ln 2 + \ln \Gamma_p\left(\frac{\nu}{2}\right) + \frac{\nu+p+1}{2} \mathbb{E}\left[\ln |\mathbf{H}|\right] + \frac{1}{2} \operatorname{Tr}\left(\mathbf{\Psi}\mathbb{E}\left[\mathbf{H}^{-1}\right]\right) \tag{6.85}$$

where c is some constant. The minimum of the Kullback-Leiber divergence is found by differentiating with respect to the respective parameters:

$$\frac{\partial \delta(\mathbf{\Psi},\nu)}{\partial \nu} = -\frac{1}{2} \ln |\mathbf{\Psi}| + \frac{p}{2} \ln 2 + \frac{1}{2} \psi_p\left(\frac{\nu}{2}\right) + \frac{1}{2} \mathbb{E}\left[\ln |\mathbf{H}|\right] = 0 \tag{6.86}$$

$$\frac{\partial \delta(\mathbf{\Psi},\nu)}{\partial \mathbf{\Psi}} = -\frac{\nu}{2} \left(\mathbf{\Psi}^{-1}\right)^T + \frac{1}{2} \mathbb{E}\left[\mathbf{H}^{-1}\right]^T = 0 \tag{6.87}$$

Solving (6.86) and (6.87) it is seen that the Kullback-Leiber divergence is minimised if and only if

$$\mathbb{E}\left[\ln |\mathbf{H}|\right] = \ln |\mathbf{\Psi}| - p \ln 2 - \psi_p\left(\frac{\nu}{2}\right), \qquad \mathbb{E}\left[\mathbf{H}^{-1}\right]^T = \nu \left(\mathbf{\Psi}^{-1}\right)^T \tag{6.88}$$

where $\psi_p(\alpha) = \frac{\partial \ln \Gamma_p(\alpha)}{\partial \alpha}$ is the multivariate digamma function.
Let $g(\mathbf{H}) = \sum_j \rho_j W^{-1}(\mathbf{H}|\mathbf{\Psi}_{\boldsymbol{ij}}, p)$ be a mixture of inverse Wishart distributions with weights ρ_j. The best approximation to this mixture of inverse Wishart distributions by a single inverse Wishart distribution $W^{-1}(\mathbf{H}|\mathbf{\Psi},\nu)$ is obtained by matching the expected values of $\ln|\mathbf{H}|$ and $\mathbf{H}^{-1}$:

$$\mathbb{E}_g\left[\mathbf{H}^{-1}\right] = \sum_j \rho_j p \mathbf{\Psi_j}^{-1} \tag{6.89}$$

$$\mathbb{E}_g\left[\ln |\mathbf{H}|\right] = \sum_j \rho_j \left\{ \ln |\mathbf{\Psi_j}| - p \ln 2 - \psi_p\left(\frac{p}{2}\right) \right\} \tag{6.90}$$

$$= \sum_j \rho_j \ln |\mathbf{\Psi_j}| - p \ln 2 - \psi_p\left(\frac{p}{2}\right) \tag{6.91}$$

Equating the equations in Equation 6.88 to Equation 6.91 and Equation 6.89 respectively we obtain:

$$\nu\boldsymbol{\Psi}^{-1} = \sum_j \rho_j p \boldsymbol{\Psi_j}^{-1} \tag{6.92}$$

$$\boldsymbol{\Psi} = \frac{\nu}{p}\left(\sum_j \rho_j \boldsymbol{\Psi_j}^{-1}\right)^{-1} \tag{6.93}$$

and

$$\ln|\boldsymbol{\Psi}| - p\ln 2 - \psi_p\left(\frac{\nu}{2}\right) = \sum_j \rho_j \ln|\boldsymbol{\Psi_j}| - p\ln 2 - \psi_p\left(\frac{p}{2}\right) \tag{6.94}$$

$$\ln\left|\frac{\nu}{p}\left(\sum_j p_j \boldsymbol{\Psi_j}^{-1}\right)^{-1}\right| - p\ln 2 - \psi_p\left(\frac{\nu}{2}\right) = \sum_j \rho_j \ln|\boldsymbol{\Psi_j}| - p\ln 2 - \psi_p\left(\frac{p}{2}\right) \tag{6.95}$$

$$\ln\frac{\nu}{p} - \psi_p\left(\frac{\nu}{2}\right) = -\ln\left|\left(\sum_j \rho_j \boldsymbol{\Psi_j}^{-1}\right)^{-1}\right| + \sum_j p_j \ln|\boldsymbol{\Psi_j}| - \psi_p\left(\frac{p}{2}\right) \tag{6.96}$$

$$\ln\frac{\nu}{p} + \ln\frac{p}{2} - \ln\frac{p}{2} - \psi_p\left(\frac{\nu}{2}\right) = -\ln\left|\left(\sum_j \rho_j \boldsymbol{\Psi_j}^{-1}\right)^{-1}\right| + \sum_j p_j \ln|\boldsymbol{\Psi_j}| - \psi_p\left(\frac{p}{2}\right) \tag{6.97}$$

$$\ln\frac{\nu}{2} - \psi_p\left(\frac{\nu}{2}\right) = \ln\frac{p}{2} - \psi_p\left(\frac{p}{2}\right) + \ln\frac{e^{\sum_j \rho_j \ln|\boldsymbol{\Psi_j}|}}{\left|\left(\sum_j \rho_j \boldsymbol{\Psi_j}^{-1}\right)^{-1}\right|} \tag{6.98}$$

$$\tag{6.99}$$

From Equation 6.98 it is clear that a numerical method is required to determine the optimal degrees of freedom. Instead of using a numerical method to determine the optimal degrees of freedom, the degrees of freedom can be fixed at the common value, p, as suggested by West and Harrison (1997). However, this approximation is not the optimal approximation. The problem of determining the inverse Wishart distribution that approximates the mixture of inverse Wishart distributions therefore reduce to

$$\boldsymbol{\Psi} = \frac{\nu}{p}\left(\sum_j \rho_j \boldsymbol{\Psi_j}^{-1}\right)^{-1} \tag{6.100}$$

$$= \left(\sum_j \rho_j \boldsymbol{\Psi_j}^{-1}\right)^{-1} \tag{6.101}$$

An approximation the mixture of inverse Wishart distributions, $g(\mathbf{H}) = \sum_j \rho_j W^{-1}(\mathbf{H}|\boldsymbol{\Psi_{ij}}, p)$, by a single inverse Wishart distribution, $W^{-1}(\mathbf{H}|\boldsymbol{\Psi_{ij}}, p)$ is therefore given by

$$\sum_j \rho_j W^{-1}(\mathbf{H}|\boldsymbol{\Psi_{ij}}, p) \approx W^{-1}\left(\mathbf{H}|\left(\sum_j \rho_j \boldsymbol{\Psi_j}^{-1}\right)^{-1}, p\right) \tag{6.102}$$

Applying this approximation of the mixture of inverse Wishart distributions to the likelihood given in Equation 6.80, the likelihood becomes

$$L\left(\mathbf{H}, \mathbf{x}^{(tr)}, \mathbf{x}^{(te)}\right) \tag{6.103}$$

$$\propto |\mathbf{H}|^{mp} e^{\frac{1}{2}m\lambda \operatorname{Tr}\left(\mathbf{H}^{-1}\right)} \prod_{i=1}^{m}\sum_{j=1}^{k} w_{ij} W^{-1}(\mathbf{H}|\boldsymbol{\Psi_{ij}}, p) \tag{6.104}$$

$$\approx |\mathbf{H}|^{mp} e^{\frac{1}{2}m\lambda \operatorname{Tr}\left(\mathbf{H}^{-1}\right)} \prod_{i=1}^{m} W^{-1}\left(\mathbf{H}|\left(\sum_j^k w_{ij} \boldsymbol{\Psi_{ij}}^{-1}\right)^{-1}, p\right) \tag{6.105}$$

$$= |\mathbf{H}|^{mp} e^{\frac{1}{2}m\lambda \operatorname{Tr}\left(\mathbf{H}^{-1}\right)} \prod_{i=1}^{m} \frac{\left|\left(\sum_j^k w_{ij} \boldsymbol{\Psi_j}^{-1}\right)^{-1}\right|^{\frac{p}{2}}}{2^{\frac{p^2}{2}}\Gamma_p\left(\frac{p}{2}\right)} |\mathbf{H}|^{-\frac{2p+1}{2}} e^{-\frac{1}{2}\operatorname{Tr}\left[\left(\sum_j^k w_{ij} \boldsymbol{\Psi_j}^{-1}\right)^{-1}\mathbf{H}^{-1}\right]} \tag{6.106}$$

$$\propto |\mathbf{H}|^{mp} |\mathbf{H}|^{-\frac{2mp+m}{2}} e^{\frac{1}{2}\lambda \operatorname{Tr}\left(\mathbf{H}^{-1}\right)} e^{-\frac{1}{2}\sum_{i=1}^{m}\operatorname{Tr}\left[\left(\sum_j^k w_{ij} \boldsymbol{\Psi_j}^{-1}\right)^{-1}\mathbf{H}^{-1}\right]} \tag{6.107}$$

$$= |\mathbf{H}|^{-\frac{m}{2}} e^{\frac{1}{2}m\lambda \operatorname{Tr}\left(\mathbf{H}^{-1}\right)} e^{-\frac{1}{2}\operatorname{Tr}\left(\sum_{i=1}^{m}\left(\sum_j^k w_{ij} \boldsymbol{\Psi_j}^{-1}\right)^{-1}\mathbf{H}^{-1}\right)} \tag{6.108}$$

$$= |\mathbf{H}|^{-\frac{m}{2}} e^{\frac{1}{2}m\lambda \operatorname{Tr}\left(\mathbf{H}^{-1}\right)} e^{-\frac{1}{2}\operatorname{Tr}\left(\boldsymbol{\Psi}^{*}\mathbf{H}^{-1}\right)} \tag{6.109}$$

where $\boldsymbol{\Psi}^{*} = \sum_{i=1}^{m}\left(\sum_j^k w_{ij} \boldsymbol{\Psi_{ij}}^{-1}\right)^{-1}$

6.2.4 Approximate posterior distribution for the bandwidth matrix

Let the inverse Wishart distribution with $p \times p$ positive definite scale parameter $m\lambda\mathbf{I}$ and $\nu > p-1$ degrees of freedom be the prior distribution of the bandwidth matrix $\mathbf{H}$

$$W^{-1}(\mathbf{H}|m\lambda\boldsymbol{I}, \nu) = \frac{|m\lambda\boldsymbol{I}|^{\frac{\nu}{2}}}{2^{\frac{\nu p}{2}}\Gamma_p\left(\frac{\nu}{2}\right)} |\mathbf{H}|^{-\frac{\nu+p+1}{2}} e^{-\frac{1}{2}\operatorname{Tr}\left(m\lambda\boldsymbol{I}\mathbf{H}^{-1}\right)} \tag{6.110}$$

$$\propto |\mathbf{H}|^{-\frac{\nu+p+1}{2}} e^{-\frac{1}{2}m\lambda \operatorname{Tr}\left(\mathbf{H}^{-1}\right)} \tag{6.111}$$

The use of the inverse Wishart prior can be motivated by the fact that it is the conjugate prior to the covariance matrix of the multivariate normal distribution. Combining Equation 6.109 and Equation 6.111 using Bayes' rule, results in the posterior distribution of $\mathbf{H}$, which can be viewed as an inverse Wishart distribution.

$$\pi\left(\mathbf{H}|\mathbf{x}^{(tr)}, \mathbf{x}^{(te)}\right) \propto |\mathbf{H}|^{-\frac{m}{2}} e^{\frac{1}{2}m\lambda \operatorname{Tr}\left(\mathbf{H}^{-1}\right)} e^{-\frac{1}{2}\operatorname{Tr}\left(\boldsymbol{\Psi}^{*}\mathbf{H}^{-1}\right)} \times |\mathbf{H}|^{-\frac{\nu+p+1}{2}} e^{-\frac{1}{2}m\lambda \operatorname{Tr}\left(\mathbf{H}^{-1}\right)} \tag{6.112}$$

$$= |\mathbf{H}|^{-\frac{(m+\nu)+p+1}{2}} e^{-\frac{1}{2}\operatorname{Tr}\left(\boldsymbol{\Psi}^{*}\mathbf{H}^{-1}\right)} \tag{6.113}$$

$$\propto W^{-1}(\mathbf{H}|\boldsymbol{\Psi}^{*}, m+\nu) \tag{6.114}$$

It is important to take into account the restrictions $\mathbf{z}^T\mathbf{\Psi^*}\mathbf{z} > 0 \; \forall \mathbf{z} \neq \mathbf{0}$ and since $\nu > p - 1$ it is only required that $m \geq 0$. However, since $\mathbf{\Psi^*}$ is the weighted sum of positive definite matrices, with all the weights greater than zero, $\mathbf{\Psi^*}$ will be positive definite and since m represents the number of observations in a subset it can not be less than zero.

6.2.5 Approximate posterior predictive distribution

The approximate posterior distribution for the bandwidth can now be used to approximate the posterior predictive distribution

$$\pi\left(\mathbf{x}|\mathbf{x}^{(tr)},\mathbf{x}^{(te)}\right) \tag{6.115}$$

$$\approx \frac{1}{k}\sum_{j=1}^{k}\int_{\mathbf{H}>0} MVN\left(\mathbf{x}|\mathbf{x}_j^{(tr)},\mathbf{H}\right) W^{-1}\left(\mathbf{H}|\mathbf{\Psi}^*,m+\nu\right)d\mathbf{H} \tag{6.116}$$

$$= \frac{1}{k}\sum_{j=1}^{k}\int_{\mathbf{H}>0} \frac{1}{(2\pi)^{\frac{p}{2}}|\mathbf{H}|^{\frac{1}{2}}} e^{(\mathbf{x}-\mathbf{x}_j^{(tr)})^T\mathbf{H}^{-1}(\mathbf{x}-\mathbf{x}_j^{(tr)})} \frac{|\mathbf{\Psi}^*|^{\frac{m+\nu}{2}}}{2^{\frac{(m+\nu)p}{2}}\Gamma_p\left(\frac{m+\nu}{2}\right)}|\mathbf{H}|^{-\frac{(m+\nu)+p+1}{2}} e^{-\frac{1}{2}\mathrm{Tr}\left(\mathbf{\Psi}^*\mathbf{H}^{-1}\right)}d\mathbf{H} \tag{6.117}$$

$$= \frac{1}{k}\sum_{j=1}^{k}\frac{|\mathbf{\Psi}^*|^{\frac{m+\nu}{2}}\Gamma_p\left(\frac{m+\nu+1}{2}\right)}{\pi^{\frac{p}{2}}\Gamma_p\left(\frac{m+\nu}{2}\right)\left|\Psi^*+(\mathbf{x}-\mathbf{x}_j^{(tr)})(\mathbf{x}-\mathbf{x}_j^{(tr)})^T\right|^{\frac{m+\nu+1}{2}}}$$

$$\times\int_{\mathbf{H}>0}\frac{\left|\mathbf{\Psi}^*+(\mathbf{x}-\mathbf{x}_j^{(tr)})(\mathbf{x}-\mathbf{x}_j^{(tr)})^T\right|^{\frac{m+\nu+1}{2}}}{2^{\frac{(m+\nu+1)p}{2}}\Gamma_p\left(\frac{m+\nu+1}{2}\right)}|\mathbf{H}|^{-\frac{(m+\nu)+p+2}{2}}$$

$$\times e^{-\frac{1}{2}\mathrm{Tr}\left(\mathbf{\Psi}^*\mathbf{H}^{-1}+(\mathbf{x}-\mathbf{x}_j^{(tr)})(\mathbf{x}-\mathbf{x}_j^{(tr)})^T\mathbf{H}^{-1}\right)}d\mathbf{H} \tag{6.118}$$

$$= \frac{1}{k}\sum_{j=1}^{k}\frac{|\mathbf{\Psi}^*|^{\frac{m+\nu}{2}}\Gamma_p\left(\frac{m+\nu+1}{2}\right)}{\pi^{\frac{p}{2}}\Gamma_p\left(\frac{m+\nu}{2}\right)\left|\mathbf{\Psi}^*+(\mathbf{x}-\mathbf{x}_j^{(tr)})(\mathbf{x}-\mathbf{x}_j^{(tr)})^T\right|^{\frac{m+\nu+1}{2}}} \tag{6.119}$$

$$= \frac{1}{k}\sum_{j=1}^{k}\frac{|\mathbf{\Psi}^*|^{\frac{m+\nu}{2}}\Gamma_p\left(\frac{m+\nu+1}{2}\right)}{\pi^{\frac{p}{2}}\Gamma_p\left(\frac{m+\nu}{2}\right)}|\mathbf{\Psi}^*|^{-\frac{m+\nu+1}{2}}\left|\mathbf{I}+\mathbf{\Psi}^{*-1}(\mathbf{x}-\mathbf{x}_j^{(tr)})(\mathbf{x}-\mathbf{x}_j^{(tr)})^T\right|^{-\frac{m+\nu+1}{2}} \tag{6.120}$$

$$= \frac{1}{k}\sum_{j=1}^{k}\frac{\Gamma_p\left(\frac{m+\nu+1}{2}\right)}{\pi^{\frac{p}{2}}\Gamma_p\left(\frac{m+\nu}{2}\right)|\mathbf{\Psi}^*|^{\frac{1}{2}}}\left|\mathbf{I}+\mathbf{\Psi}^{*-1}(\mathbf{x}-\mathbf{x}_j^{(tr)})(\mathbf{x}-\mathbf{x}_j^{(tr)})^T\right|^{-\frac{m+\nu+1}{2}} \tag{6.121}$$

$$= \frac{1}{k}\sum_{j=1}^{k}\frac{\Gamma\left(\frac{m+\nu+1}{2}\right)}{\pi^{\frac{p}{2}}\Gamma\left(\frac{m+\nu+1-p}{2}\right)|\mathbf{\Psi}^*|^{\frac{1}{2}}}\left|\mathbf{I}+\mathbf{\Psi}^{*-1}(\mathbf{x}-\mathbf{x}_j^{(tr)})(\mathbf{x}-\mathbf{x}_j^{(tr)})^T\right|^{-\frac{m+\nu+1}{2}} \tag{6.122}$$

$$= \frac{1}{k}\sum_{j=1}^{k}\frac{\Gamma\left(\frac{m+\nu+1}{2}\right)}{\pi^{\frac{p}{2}}\Gamma\left(\frac{m+\nu+1-p}{2}\right)(m+\nu+1-p)^{\frac{p}{2}}\left|\frac{1}{m+\nu+1-p}\mathbf{\Psi}^*\right|^{\frac{1}{2}}}$$

$$\times\left[1+\frac{1}{m+\nu+1-p}(\mathbf{x}-\mathbf{x}_j^{(tr)})^T\left(\frac{1}{m+\nu+1-p}\mathbf{\Psi}^*\right)^{-1}(\mathbf{x}-\mathbf{x}_j^{(tr)})\right]^{-\frac{m+\nu+1}{2}} \tag{6.123}$$

$$= \frac{1}{k}\sum_{j=1}^{k}t_{m+\nu+1-p}\left(\mathbf{x},\mathbf{x}_j^{(tr)},\frac{1}{m+\nu+1-p}\mathbf{\Psi}^*\right) \tag{6.124}$$

The posterior predictive distribution can thus be approximated by a mixture of multivariate Student's t distributions, with each distribution having centre $\mathbf{x}_i$ and the matrix $\frac{1}{m+\nu+1-p}\mathbf{\Psi}^*$ playing the same role as the bandwidth matrix in traditional kernel density estimation. Take

note that Equation 6.119 follows from the fact that

$$\begin{aligned}
&\int_{\mathbf{H}>0} \frac{\left|\mathbf{\Psi}^* + (\mathbf{x}-\mathbf{x}_j^{(tr)})(\mathbf{x}-\mathbf{x}_j^{(tr)})^T\right|^{\frac{m+\nu+1}{2}}}{2^{\frac{(m+\nu+1)p}{2}}\Gamma_p\left(\frac{m+\nu+1}{2}\right)} |\mathbf{H}|^{-\frac{(m+\nu)+p+2}{2}} \\
&\times e^{-\frac{1}{2}\operatorname{Tr}\left(\mathbf{\Psi}^*\mathbf{H}^{-1}+(\mathbf{x}-\mathbf{x}_j^{(tr)})(\mathbf{x}-\mathbf{x}_j^{(tr)})^T\mathbf{H}^{-1}\right)} d\mathbf{H} \\
&= \int_{\mathbf{H}>0} W^{-1}\left(\mathbf{H}|\mathbf{\Psi}^* + (\mathbf{x}-\mathbf{x}_j^{(tr)})(\mathbf{x}-\mathbf{x}_j^{(tr)})^T, m+\nu+1\right) d\mathbf{H} \\
&= 1
\end{aligned}$$

and Equation 6.122 is true since

$$\begin{aligned}
\frac{\Gamma_p\left(\frac{m+\nu+1}{2}\right)}{\Gamma_p\left(\frac{m+\nu}{2}\right)} &= \frac{\pi^{\frac{p(p-1)}{4}}\prod_{i=1}^{p}\Gamma\left(\frac{m+\nu+1}{2}+\frac{1-i}{2}\right)}{\pi^{\frac{p(p-1)}{4}}\prod_{i=1}^{p}\Gamma\left(\frac{m+\nu}{2}+\frac{1-i}{2}\right)} \\
&= \frac{\Gamma\left(\frac{m+\nu+1}{2}\right)\Gamma\left(\frac{m+\nu}{2}\right)\Gamma\left(\frac{m+\nu-1}{2}\right)\ldots\Gamma\left(\frac{m+\nu+3-p}{2}\right)\Gamma\left(\frac{m+\nu+2-p}{2}\right)}{\Gamma\left(\frac{m+\nu}{2}\right)\Gamma\left(\frac{m+\nu-1}{2}\right)\ldots\Gamma\left(\frac{m+\nu+3-p}{2}\right)\Gamma\left(\frac{m+\nu+2-p}{2}\right)\Gamma\left(\frac{m+\nu+1-p}{2}\right)} \\
&= \frac{\Gamma\left(\frac{m+\nu+1}{2}\right)}{\Gamma\left(\frac{m+\nu+1-p}{2}\right)}
\end{aligned}$$

Finally Equation 6.123 can be seen to be true by considering the following identities; Let $\mathbf{A}$ be some $k \times l$, $\mathbf{B}$ some $l \times k$ and $\mathbf{D}$ some $n \times n$ matrix and let c be some constant, then

$$|\mathbf{I}_k + \mathbf{AB}| = |\mathbf{I}_l + \mathbf{BA}|$$

and

$$|c\mathbf{D}| = c^n |\mathbf{D}|$$

Calculating the approximate reference predictive distribution for each of the n_p random partitions of the form $\mathbf{X} = \{\mathbf{X}_{(l)}^{(te)}, \mathbf{X}_{(l)}^{(tr)}\}$ with $l = 1, \ldots, n_p$, the desired model is obtained by substituting Equation 6.124 into Equation 6.72

$$\begin{aligned}
p\left(\mathbf{x}|k, \mathbf{X}\right) &= \frac{1}{n_p}\sum_{l=1}^{n_p} p\left(\mathbf{x}|\mathbf{X}_{(l)}^{(te)}, \mathbf{X}_{(l)}^{(tr)}\right) \\
&= \frac{1}{n_p}\sum_{l=1}^{n_p}\frac{1}{k}\sum_{j=1}^{k} t_{m+\nu+1-p}\left(\mathbf{x}, \mathbf{x}_{(l)_j^{(tr)}}, \frac{1}{m+\nu+1-p}\mathbf{\Psi}_{(l)}^*\right) \qquad (6.125)
\end{aligned}$$

where $\mathbf{\Psi}_{(l)}^*$ is the matrix $\mathbf{\Psi}^*$ calculated using the l^{th} random partion of $\mathbf{X}$. It is important to note that for each of the n_p partitions the value of k remains fixed.

6.2.6 Number of mixture components, k

As in the univariate case, we postulate the use of sample entropy to select the optimal value of k. Using the suggestion made by Shannon (2001) for calculating sample entropy, the sample entropy for the predictive density of $\mathbf{x}_j$ is

$$H\left(\hat{p}\right) = -\frac{1}{n}\sum_{j=1}^{n}\ln\left[\hat{p}\left(\mathbf{x}_j|\mathbf{X}_{-j}\right)\right] \tag{6.126}$$

where $\hat{p}\left(\mathbf{x}_j|\mathbf{X}_{-j}\right)$ is the estimated predictive density of $\mathbf{x}_j$ given all the observations except for $\mathbf{x}_j$ itself. It can be shown that selecting the value of k that results in the minimum sample entropy is similar to selecting the value of k that maximise the likelihood function of estimated predictive density.
The method used in calculating the sample entropy can be viewed as a form of leave-one-out cross-validation.

6.3 Conclusion

This chapter started off by reviewing the univariate approach to Bayesian kernel density estimation. An example for the univariate case was included in order to demonstrate the workings of the method. The chapter went further to develop a theoretical approach to correlated multivariate Bayesian kernel density estimation.
It is important to mention that this method both the univariate as well as the multivariate case suffers from robustness issues. Various simulations of using the same data result in different density estimates. This is due to the random partitioning of the data, $\mathbf{X}$. In the univariate case the resulting densities differ marginally, depending upon the data set used. However, in the multivariate case robustness is a major problem, as we were unable to duplicate our results for various simulations.
Therefore, future work will include solving the robustness issues, investigating the effect of λ on the resulting posterior predictive distribution as well as investigating methods for reducing the computational time.

Chapter 7

Conclusion

7.1 Dissertation summary

In Chapter 1 the introduction of the dissertation was given, in which the background of credit scoring and the class-imbalance problem was defined.
Chapter 2 described the data sets used throughout the dissertation, as well as the methods used to evaluate the performance of the relevant classifiers. The processes used to prepare the data for use by the algorithms were also covered.
Chapter 3 covered the parametric classifiers and the mathematical mechanics that drive these classifiers.
Chapter 4 discussed the non-parametric classifiers, in particular the kernel density estimation method. The chapter elaborated on some of the different methods used to estimate the bandwidth used in kernel density estimation. The chapter also introduced two types of priors, the frequentist and the Bernoulli prior, that can be used in conjunction with any two-class parametric and non-parametric classifiers.
Chapter 5 set out the experimental design for various experiments as well as the experiments themselves. The aim of the chapter was to determine the effect of the Bernoulli and frequentist priors on the various classifiers, identifying the optimal prior if possible and determining whether parametric or non-parametric classifiers perform best, for various class-imbalances. The chapter went further to identify the best performing classifier over all evaluated class-imbalances.
Chapter 6 takes the idea of kernel density estimation and priors and combines it to form a fully Bayesian approach to kernel density estimation. The chapter starts off with the mathematics done by Bernardo (1999) for the univariate case, and later on extends this to a theoretical approach to Bayesian kernel density estimation for the multivariate case. The multivariate approach developed is theoretically able to handle correlated multivariate data. The chapter also provided a short example of the practical use of the Bayesian kernel density estimator in the univariate case.

7.2 Dissertation contribution

The development of a Multivariate Bayesian kernel density estimator not only provide kernel density estimators with a much desired predictive power, but also opens the world of kernel density estimation to credibility intervals. At least in theory this development takes the superior classification power of the kernel density estimator over the LR classifier, and provides it with the predictive power of classifiers such as LR. Furthermore, the dissertation introduced the frequentist as well as the Bernoulli priors. In Section 5.2.1 it was illustrated that the use of an optimal Bernoulli prior does in fact improve the performance of the evaluated classifiers. It was also shown that the Bernoulli prior provides the flexibility of acting as if there were no prior used by simply tuning the prior parameter. The section also showed that the prior parameter used to optimise the hit rate does not necessarily optimise the harmonic mean, leaving the practitioner with a choice to make. It was also shown that the Bernoulli prior may be a solution to overcoming the class-imbalance problem. Section 5.2.2 compared the performance of the Bernoulli and frequentist prior to that of the classifier used with no prior. The section highlighted that although in some instances the use of a Bernoulli prior with a fixed parameter regardless of class-imbalance results in the optimal performance, the Bernoulli prior parameter should be dependent upon the class-imbalance. The section also showed that the dependence that exists between the Bernoulli prior parameter and the class-imbalance is not that given by the frequentist prior. Section 5.2.3 emphasised that the non-parametric classifiers outperform the parametric classifiers. For the majority of data sets used the Silverman classifier outperformed the MLE classifier, indicating that although the MLE is better at density estimation the Silverman classifier is better at classification. Whether the Gaussian and NB classifiers or the LR and BLR classifiers perform better depends greatly on the data set.

7.3 Future work

There is quite a lot of future work to be done. Although the theory and mathematics for a multivariate Bayesian kernel density estimator were worked out in this dissertation, the work should be extended so that it is usable by practitioners. This would entail solving the current robustness issues. Methods of reducing computational time for the model should also be investigated. Focus can also be placed on determining the optimal value of the parameter λ used to ensure the matrix $\mathbf{\Psi}$ is positive definite.

With regard to the Bernoulli priors; the possibility of an exact relationship between the default ratio and the optimal prior parameter should be investigate. Doing this would remove the need to do a grid search and hence save computational time. The concept of Bernoulli priors can be extended to other two-class classifiers and more extensive data sets in order to

determine the limitations of the prior.

Multivariate Bayesian kernel density estimation is a promising and relatively new field. It is our hope that this dissertation will add great value to not only the specific field but also statistics in general.

Appendix A

Silverman's univariate rule of thumb

The $AMISE$ of some unknow function $\hat{f}_k$ is given by

$$AMISE(\hat{f}_k) = \frac{1}{nh}\|K\|_2^2 + \frac{h^4}{4}\{\mu_2(K)\}^2\|f''\|_2^2$$

Differentiating the $AMISE(\hat{f}_k)$ with respect to h Equation A.1 follows

$$\frac{d}{dh}AMISE(\hat{f}_k) = -\frac{1}{nh^2}\|K\|_2^2 + h^3\{\mu_2(K)\}^2\|f''\|_2^2 \tag{A.1}$$

Setting Equation A.1 equal to 0 and solving h

$$h = \left(\frac{\|K\|_2^2}{n\{\mu_2(K)\}^2\|\hat{f}_k\|_2^2}\right)^{\frac{1}{5}} \tag{A.2}$$

where K is the kernel function used, f is some unknown density function and $\mu_2(K) = \int x^2K(x)dx$.

We furthermore assume that f belongs to the family of normal distributions with mean μ and variance σ^2. Then

$$\|f''\|_2^2 = \left[\left\{\int_{-\infty}^{\infty}|f''(x)|^2dx\right\}^{\frac{1}{2}}\right]^2 \tag{A.3}$$

$$= \int_{-\infty}^{\infty}|f''(x)|^2dx \tag{A.4}$$

Equation A.3 follows from the fact that $\|j\|_2 = \left\{\int_a^b |j(x)|^2dg(x)\right\}^{\frac{1}{2}}$ with $g(x) = x$.

The probability density function of f is given by

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

The first derivative is therefore

$$\begin{aligned} f'(x) &= -\frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \left(\frac{1}{\sigma}\right)\left(\frac{x-\mu}{\sigma}\right) \\ &= -\frac{1}{\sqrt{2\pi}\sigma^2} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \left(\frac{x-\mu}{\sigma}\right) \end{aligned}$$

and the second derivative is

$$\begin{aligned} f''(x) &= \frac{1}{\sqrt{2\pi}\sigma^3} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \left(\frac{x-\mu}{\sigma}\right)^2 - \frac{1}{\sqrt{2\pi}\sigma^3} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \\ &= \frac{1}{\sqrt{2\pi}\sigma^3} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \left[\left(\frac{x-\mu}{\sigma}\right)^2 - 1\right] \end{aligned} \tag{A.5}$$

Note that the propability density function of the standard normal distribution, $p(z)$, is given by

$$p(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

with the first order derivative

$$p'(z) = -\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} z$$

and the second order derivative

$$p''(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}(z^2 - 1) \tag{A.6}$$

Set $z = \frac{x-\mu}{\sigma}$ so that $x = x\sigma + \mu$ and thus resulting in Equation A.5 being written as

$$f''(x) = \frac{1}{\sqrt{2\pi}\sigma^3} e^{-\frac{1}{2}z^2} \left[z^2 - 1\right] \tag{A.7}$$

Taking Equation A.6 into account, Equation A.7 can be expressed as

$$f''(x) = \frac{1}{\sigma^3} p''(z) \tag{A.8}$$

Substituting Equation A.8 into Equation A.4

$$\begin{aligned}
\|f''\|_2^2 &= \int_{-\infty}^{\infty} |f''(x)|^2 dx \\
&= \int_{-\infty}^{\infty} |\frac{1}{\sigma^3} p''(z)|^2 \sigma dz \qquad \text{(A.9)} \\
&= \sigma^{-5} \int_{-\infty}^{\infty} \{p''(z)\}^2 dz \\
&= \frac{\sigma^{-5}}{2\pi} \int_{-\infty}^{\infty} e^{-z^2} (z^2 - 1)^2 dz \\
&= \frac{\sigma^{-5}}{2\pi} \left[\int_{-\infty}^{\infty} e^{-z^2} z^4 dz - 2 \int_{-\infty}^{\infty} e^{-z^2} z^2 dz + \int_{-\infty}^{\infty} e^{-z^2} dz \right] \qquad \text{(A.10)}
\end{aligned}$$

Equation A.9 follows from the fact that $dx = \sigma dz$.

We start by solving $\int_{-\infty}^{\infty} e^{-ax^2} dx$:
Let $I = \int_{-\infty}^{\infty} e^{-ax^2} dx$ and similarly, let $I = \int_{-\infty}^{\infty} e^{-ay^2} dy$. Then

$$\begin{aligned}
I^2 &= \int_{-\infty}^{\infty} e^{-ax^2} dx \int_{-\infty}^{\infty} e^{-ay^2} dy \\
&= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-a(x^2+y^2)} dx dy \\
&= \int_0^{2\pi} \int_0^{\infty} e^{-ar^2} r dr d\theta \qquad \text{(A.11)} \\
&= \left[-\frac{1}{2} e^{-ar^2} \right]_0^{\infty} [\theta]_0^{2\pi} \\
&= (\frac{1}{2a})(2\pi) \\
&= \frac{\pi}{a}
\end{aligned}$$

Taking the square root on both sides, I is determined as

$$I = \frac{\sqrt{\pi}}{\sqrt{a}} \qquad \text{(A.12)}$$

Equation A.11 follows from the fact that $x = r\cos(\theta)$ and $y = r\sin(\theta)$. Setting $a = 1$ in Equation A.12 it is trivial to see that

$$\int_{-\infty}^{\infty} e^{-z^2} dz = \sqrt{\pi} \qquad \text{(A.13)}$$

The next step is to solve $\int_{-\infty}^{\infty} e^{-z^2} z^2 dz$ by utilising Feynman's trick

$$\begin{aligned}
\int_{-\infty}^{\infty} e^{-z^2} z^2 dz &= -\frac{d}{da} \sqrt{\frac{\pi}{a}} |_{a=1} \\
&= \frac{\sqrt{\pi}}{2} a^{-\frac{3}{2}} |_{a=1} \\
&= \frac{\sqrt{\pi}}{2} \qquad \text{(A.14)}
\end{aligned}$$

The integral $\int_{-\infty}^{\infty} e^{-z^2} z^4 dz$ can be solved in a similar fashion

$$\begin{aligned}\int_{-\infty}^{\infty} e^{-z^2} z^4 dz &= \frac{d^2}{da^2}\sqrt{\frac{\pi}{a}}|_{a=1} \\ &= -\frac{d}{da}\frac{\sqrt{\pi}}{2}a^{-\frac{3}{2}}|_{a=1} \\ &= \frac{3\sqrt{\pi}}{4}a^{-\frac{5}{2}}|_{a=1} \\ &= \frac{3\sqrt{\pi}}{4} \end{aligned} \tag{A.15}$$

Substituting Equation A.13, EquationA.14 and Equation A.15 back into Equation A.10 we get:

$$\begin{aligned}\|f''\| &= \frac{\sigma^{-5}}{2}\left[\int_{-\infty}^{\infty} e^{-z^2} z^4 dz - 2\int_{-\infty}^{\infty} e^{-z^2} z^2 dz + \int_{-\infty}^{\infty} e^{-z^2} dz\right] \\ &= \frac{\sigma^{-5}}{2}\left[\frac{3\sqrt{\pi}}{4} - 2\left(\frac{\sqrt{\pi}}{2}\right) + \sqrt{\pi}\right] \\ &= \sigma^{-5}\frac{3}{8\sqrt{\pi}} \end{aligned} \tag{A.16}$$

Due to the fact that σ is unknown it should be estimated by

$$\hat{\sigma} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$$

Assuming that the kernel function is the Gaussian kernel function: $p(z) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}$, then:

$$\begin{aligned}\mu_2(K) &= \mu_2(p(z)) \\ &= \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2} z^2 dz \\ &= E(Z^2) \\ &= Var(Z) + E[Z]^2 \qquad \text{(A.17)} \\ &= 1 + 0^2 \qquad \text{(A.18)} \\ &= 1 \qquad \text{(A.19)} \end{aligned}$$

where Equation A.18 follows from the fact that $Z \sim N(0,1)$. Furthermore,

$$
\begin{aligned}
\|p\|_2^2 &= \left[\left\{\int_{-\infty}^{\infty} |p(z)|^2 dz\right\}^{\frac{1}{2}}\right]^2 \\
&= \int_{-\infty}^{\infty} \frac{1}{2\pi} e^{-z^2} dz \\
&= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-z^2} dz && \text{(A.20)} \\
&= \left(\frac{1}{2\pi}\right)\left(\sqrt{\pi}\right) && \text{(A.21)} \\
&= \frac{1}{2\sqrt{\pi}} && \text{(A.22)}
\end{aligned}
$$

Equation A.21 is a result of substituting Equation A.13 into Equation A.20.

Finally we can calculate $\hat{h}$:

$$
\begin{aligned}
\hat{h} &= \left(\frac{\|p\|_2^2}{\|\hat{f}''\|_2^2 \mu_2^2(p) n}\right)^{\frac{1}{5}} && \text{(A.23)} \\
&= \left(\frac{\frac{1}{2\sqrt{\pi}}}{\left(\hat{\sigma}^{-5}\frac{3}{8\sqrt{\pi}}\right)(1)^2 n}\right)^{\frac{1}{5}} && \text{(A.24)} \\
&= \left(\frac{4}{3n}\hat{\sigma^5}\right)^{\frac{1}{5}} \\
&\approx 1.06\hat{\sigma} n^{-\frac{1}{5}}
\end{aligned}
$$

Equation A.24 is a result of substituting Equation A.16, Equation A.19 and Equation A.22 into Equation A.2. This concludes the proof for the univariate case of silverman's rule of thumb.

Appendix B

Correlated multivariate Bayesian kernel density estimation: Special case

This section follows the same general pattern as given in Section 6.2 with the exception that it is assumed that the data is of such a nature that for all i and j the matrix $\boldsymbol{\Psi_{ij}}$ is positive definite. This assumption will allow the use of an uniformative in stead of an inverse wishart prior.

Consider the data $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ consisting of n instances and p features. Let $\mathbf{X}^{(tr)} = \{\mathbf{x}_1^{(tr)}, \ldots, \mathbf{x}_k^{(tr)}\}$,for some $0 < k < n$, be a subset, of size k, of $\mathbf{X}$. Let $\mathbf{X}^{(te)} = \{\mathbf{x}_1^{(te)}, \ldots, \mathbf{x}_m^{(te)}\}$ be a subset, of size $m = n - k$, of $\mathbf{X}$. Note that $\mathbf{X} = \{\mathbf{X}^{(te)}, \mathbf{X}^{(tr)}\}$ so that $\mathbf{X}^{(te)}$ and $\mathbf{X}^{(tr)}$ are non-overlapping subsets of $\mathbf{X}$. Let $q(.)$ be a kernel function, then using kernel density estimation a reasonable approximation to density of $\mathbf{X}^{(te)}$ given the bandwidth, $\mathbf{H}$, may be obtained

$$\begin{aligned} p\left(\mathbf{X}^{(te)}|\mathbf{H}\right) &= \prod_{i=1}^{m} p\left(\mathbf{x}_i^{(te)}|\mathbf{H}\right) \\ &\approx \prod_{i=1}^{m}\sum_{j=1}^{k} q\left(\mathbf{x}_i^{(te)}|\mathbf{x}_j^{(tr)}, \mathbf{H}\right) \end{aligned}$$

Thus, using Bayes' rule the posterior distribution of the bandwidth is

$$\begin{aligned} p\left(\mathbf{H}|\mathbf{X}^{(tr)}, \mathbf{X}^{(te)}\right) &\propto p(\mathbf{H})\, p\left((\mathbf{X}^{(te)}|(\mathbf{X}^{(tr)}, \mathbf{H}\right) \\ &= p(\mathbf{H}) \prod_{i=1}^{m}\sum_{j=1}^{k} q\left(\mathbf{x}_i^{(te)}|\mathbf{x}_j^{(tr)}, \mathbf{H}\right) \end{aligned}$$

The posterior predictive distribution for some data point $\mathbf{x}$ is thus given by

$$\begin{aligned} p\left(\mathbf{x}|\mathbf{X}^{(tr)},\mathbf{X}^{(te)}\right) &= \int_{\mathbf{H}>0} p\left(\mathbf{x}|\mathbf{H}\right) p\left(\mathbf{H}|\mathbf{X}^{(tr)},\mathbf{X}^{(te)}\right) d\mathbf{H} \\ &\approx \int_{\mathbf{H}>0} \frac{1}{k}\sum_{j=1}^{k} q\left(\mathbf{x}|\mathbf{x}_j^{(tr)},\mathbf{H}\right) p\left(\mathbf{H}|\mathbf{X}^{(tr)},\mathbf{X}^{(te)}\right) d\mathbf{H} \\ &= \frac{1}{k}\sum_{j=1}^{k} \int_{\mathbf{H}>0} q\left(\mathbf{x}|\mathbf{x}_j^{(tr)},\mathbf{H}\right) p\left(\mathbf{H}|\mathbf{X}^{(tr)},\mathbf{X}^{(te)}\right) d\mathbf{H} \end{aligned}$$

It is clear that the posterior predictive distribution and the posterior distribution of the bandwidth is dependent on the choice of the kernel function. The following sections assume a multivariate normal kernel function.

B.0.1 Likelihood

Consider the probability density function of the inverted Wishart distribution:

$$W^{-1}\left(\mathbf{H}|\mathbf{\Psi},\nu\right) = \frac{\left|\mathbf{\Psi}\right|^{\frac{\nu}{2}}}{2^{\frac{\nu p}{2}}\Gamma_p\left(\frac{\nu}{2}\right)} \left|\mathbf{H}\right|^{-\frac{\nu+p+1}{2}} e^{-\frac{1}{2}\operatorname{Tr}\left(\mathbf{\Psi}\mathbf{H}^{-1}\right)}$$

Appendix B. Correlated multivariate Bayesian kernel density estimation: Special case

The likelihood function may be expressed in terms of the inverse Wishart distribution as follows:

$$
\begin{aligned}
&L\left(\mathbf{H}, \mathbf{X}^{(tr)}, \mathbf{X}^{(te)}\right) \\
&= \prod_{i=1}^{m} \frac{1}{k} \sum_{j=1}^{k} q(\mathbf{x}_i^{(te)} | \mathbf{x}_j^{(tr)}, \mathbf{H}) \\
&= \prod_{i=1}^{m} \frac{1}{k} \sum_{j=1}^{k} \frac{1}{(2\pi)^{\frac{p}{2}} |\mathbf{H}|^{\frac{1}{2}}} e^{-\frac{1}{2}\left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right)^T \mathbf{H}^{-1} \left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right)} \\
&= \prod_{i=1}^{m} \frac{1}{k} \sum_{j=1}^{k} \frac{2^{\frac{p^2}{2}} \Gamma_p\left(\frac{p}{2}\right) |\mathbf{H}|^p}{(2\pi)^{\frac{p}{2}} \left| \left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right) \left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right)^T \right|^{\frac{p}{2}}} \frac{\left| \left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right) \left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right)^T \right|^{\frac{p}{2}}}{2^{\frac{p^2}{2}} \Gamma_p\left(\frac{p}{2}\right)} \\
&\times |\mathbf{H}|^{-\frac{2p+1}{2}} e^{-\frac{1}{2} \operatorname{Tr}\left[\left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right)\left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right)^T \mathbf{H}^{-1}\right]} \\
&= \prod_{i=1}^{m} \frac{1}{k} \sum_{j=1}^{k} \frac{2^{\frac{p^2}{2}} \Gamma_p\left(\frac{p}{2}\right) |\mathbf{H}|^p}{(2\pi)^{\frac{p}{2}} \left| \left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right) \left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right)^T \right|^{\frac{p}{2}}} \\
&\times W^{-1}\left(\mathbf{H} \left| \left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right) \left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right)^T, p \right.\right) \\
&\propto \prod_{i=1}^{m} \sum_{j=1}^{k} \frac{|\mathbf{H}|^p}{|\mathbf{\Psi}_{ij}|^{\frac{p}{2}}} W^{-1}\left(\mathbf{H} | \mathbf{\Psi}_{ij}, p\right) \\
&\propto |\mathbf{H}|^{mp} \prod_{i=1}^{m} \sum_{j=1}^{k} w_{ij} W^{-1}\left(\mathbf{H} | \mathbf{\Psi}_{ij}, p\right)
\end{aligned} \tag{B.1}
$$

where $\mathbf{\Psi_{ij}} = \left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right) \left(\mathbf{x}_i^{(te)} - \mathbf{x}_j^{(tr)}\right)^T$ is a $p \times p$ matrix and $w_{ij} = \frac{|\mathbf{\Psi_{ij}}|^{-\frac{p}{2}}}{\sum_{j=1}^{k} |\mathbf{\Psi_{ij}}|^{-\frac{p}{2}}}$. The Kullback-Leiber divergence of a inverse Wishart distribution $W^{-1}\left(\mathbf{H} | \mathbf{\Psi}, \nu\right)$ from some density function $g\left(\mathbf{H}\right)$ is given by

$$
\begin{aligned}
&\delta\left(\mathbf{\Psi}, \nu\right) \\
&= \int_{\mathbf{H}>0} g\left(\mathbf{H}\right) \ln \frac{g\left(\mathbf{H}\right)}{W^{-1}\left(\mathbf{H} | \mathbf{\Psi}, \nu\right)} d\mathbf{H} \\
&= \int_{\mathbf{H}>0} g\left(\mathbf{H}\right) \ln g\left(\mathbf{H}\right) d\mathbf{H} - \int_{\mathbf{H}>0} g\left(\mathbf{H}\right) \ln W^{-1}\left(\mathbf{H} | \mathbf{\Psi}, \nu\right) d\mathbf{H} \\
&= \int_{\mathbf{H}>0} g\left(\mathbf{H}\right) \ln g\left(\mathbf{H}\right) d\mathbf{H} - \int_{\mathbf{H}>0} g\left(\mathbf{H}\right) \ln \left(\frac{|\mathbf{\Psi}|^{\frac{\nu}{2}}}{2^{\frac{\nu p}{2}} \Gamma_p\left(\frac{\nu}{2}\right)} |\mathbf{H}|^{-\frac{\nu+p+1}{2}} e^{-\frac{1}{2} \operatorname{Tr}\left(\mathbf{\Psi} \mathbf{H}^{-1}\right)} \right) d\mathbf{H} \\
&= \int_{\mathbf{H}>0} g\left(\mathbf{H}\right) \ln g\left(\mathbf{H}\right) d\mathbf{H} - \frac{\nu}{2} \int_{\mathbf{H}>0} g\left(\mathbf{H}\right) \ln |\mathbf{\Psi}| \, d\mathbf{H} + \int_{\mathbf{H}>0} \frac{\nu p}{2} g\left(\mathbf{H}\right) \ln 2 d\mathbf{H} \\
&+ \int_{\mathbf{H}>0} g\left(\mathbf{H}\right) \ln \Gamma_p\left(\frac{\nu}{2}\right) d\mathbf{H} + \frac{\nu+p+1}{2} \int_{\mathbf{H}>0} g\left(\mathbf{H}\right) \ln |\mathbf{H}| \, d\mathbf{H} + \frac{1}{2} \int_{\mathbf{H}>0} g\left(\mathbf{H}\right) \operatorname{Tr}\left(\mathbf{\Psi} \mathbf{H}^{-1}\right) d\mathbf{H} \\
&= c - \frac{\nu}{2} \ln |\mathbf{\Psi}| + \frac{\nu p}{2} \ln 2 + \ln \Gamma_p\left(\frac{\nu}{2}\right) + \frac{\nu+p+1}{2} \mathbb{E}\left[\ln |\mathbf{H}|\right] + \frac{1}{2} \operatorname{Tr}\left(\mathbf{\Psi} \mathbb{E}\left[\mathbf{H}^{-1}\right]\right)
\end{aligned}
$$

where c is some constant. The minimum of the Kullback-Leiber divergence is found by differentiating with respect to the respective parameters:

$$\frac{\partial \delta\left(\mathbf{\Psi}, \nu\right)}{\partial \nu} = -\frac{1}{2}\ln\left|\mathbf{\Psi}\right| + \frac{p}{2}\ln 2 + \frac{1}{2}\psi_p\left(\frac{\nu}{2}\right) + \frac{1}{2}\mathbb{E}\left[\ln\left|\mathbf{H}\right|\right] = 0 \tag{B.2}$$

$$\frac{\partial \delta\left(\mathbf{\Psi}, \nu\right)}{\partial \mathbf{\Psi}} = -\frac{\nu}{2}\left(\mathbf{\Psi}^{-1}\right)^T + \frac{1}{2}\mathbb{E}\left[\mathbf{H}^{-1}\right]^T = 0 \tag{B.3}$$

Solving (B.2) and (B.3) it is seen that the Kullback-Leiber divergence is minimized if and only if

$$\mathbb{E}\left[\ln\left|\mathbf{H}\right|\right] = \ln\left|\mathbf{\Psi}\right| - p\ln 2 - \psi_p\left(\frac{\nu}{2}\right), \qquad \mathbb{E}\left[\mathbf{H}^{-1}\right]^T = \nu\left(\mathbf{\Psi}^{-1}\right)^T \tag{B.4}$$

where $\psi_p\left(\alpha\right) = \frac{\partial \ln \Gamma_p(\alpha)}{\partial \alpha}$ is the multivariate digamma function.
Let $g\left(\mathbf{H}\right) = \sum_j \rho_j W^{-1}\left(\mathbf{H}|\mathbf{\Psi_{ij}}, p\right)$ be a mixture of inverse Wishart distributions with weights ρ_j. The best approximation to this mixture of inverse Wishart distributions by a single inverse Wishart distribution $W^{-1}\left(\mathbf{H}|\mathbf{\Psi}, \nu\right)$ is obtained by matching the expected values of $\ln|\mathbf{H}|$ and $\mathbf{H}^{-1}$:

$$\mathbb{E}_g\left[\mathbf{H}^{-1}\right] = \sum_j \rho_j p \mathbf{\Psi_j}^{-1} \tag{B.5}$$

$$\begin{aligned}\mathbb{E}_g\left[\ln\left|\mathbf{H}\right|\right] &= \sum_j \rho_j \left\{\ln\left|\mathbf{\Psi_j}\right| - p\ln 2 - \psi_p\left(\frac{p}{2}\right)\right\} \\ &= \sum_j \rho_j \ln\left|\mathbf{\Psi_j}\right| - p\ln 2 - \psi_p\left(\frac{p}{2}\right)\end{aligned} \tag{B.6}$$

Equating the equations in Equation B.4 to the equations in Equation B.6 and Equation B.5 respectively we obtain:

$$\nu\mathbf{\Psi}^{-1} = \sum_j \rho_j p \mathbf{\Psi_j}^{-1}$$

$$\mathbf{\Psi} = \frac{\nu}{p}\left(\sum_j \rho_j \mathbf{\Psi_j}^{-1}\right)^{-1}$$

and

$$\ln|\boldsymbol{\Psi}| - p\ln 2 - \psi_p\left(\frac{\nu}{2}\right) = \sum_j \rho_j \ln|\boldsymbol{\Psi_j}| - p\ln 2 - \psi_p\left(\frac{p}{2}\right)$$

$$\ln\left|\frac{\nu}{p}\left(\sum_j p_j {\boldsymbol{\Psi_j}}^{-1}\right)^{-1}\right| - p\ln 2 - \psi_p\left(\frac{\nu}{2}\right) = \sum_j \rho_j \ln|\boldsymbol{\Psi_j}| - p\ln 2 - \psi_p\left(\frac{p}{2}\right)$$

$$\ln\frac{\nu}{p} - \psi_p\left(\frac{\nu}{2}\right) = -\ln\left|\left(\sum_j \rho_j {\boldsymbol{\Psi_j}}^{-1}\right)^{-1}\right| + \sum_j p_j \ln|\boldsymbol{\Psi_j}| - \psi_p\left(\frac{p}{2}\right)$$

$$\ln\frac{\nu}{p} + \ln\frac{p}{2} - \ln\frac{p}{2} - \psi_p\left(\frac{\nu}{2}\right) = -\ln\left|\left(\sum_j \rho_j {\boldsymbol{\Psi_j}}^{-1}\right)^{-1}\right| + \sum_j p_j \ln|\boldsymbol{\Psi_j}| - \psi_p\left(\frac{p}{2}\right)$$

$$\ln\frac{\nu}{2} - \psi_p\left(\frac{\nu}{2}\right) = \ln\frac{p}{2} - \psi_p\left(\frac{p}{2}\right) + \ln\frac{e^{\sum_j \rho_j \ln|\boldsymbol{\Psi_j}|}}{\left|\left(\sum_j \rho_j {\boldsymbol{\Psi_j}}^{-1}\right)^{-1}\right|} \tag{B.7}$$

(B.8)

From B.7 it is clear that a numerical method is required to determine the optimal degrees of freedom. Instead of using a numerical method to determine the optimal degrees of freedom, the degrees of freedom can be fixed at the common value, p, as suggested by West and Harrison (1997). However, this approximation is not the optimal approximation. The problem of determining the inverse wishart distribution that approximates the mixture of inverse wishart distributions therefore reduce to

$$\begin{aligned}\boldsymbol{\Psi} &= \frac{\nu}{p}\left(\sum_j \rho_j {\boldsymbol{\Psi_j}}^{-1}\right)^{-1} \\ &= \left(\sum_j \rho_j {\boldsymbol{\Psi_j}}^{-1}\right)^{-1}\end{aligned}$$

An approximation the mixture of inverse wishart distributions, $g(\mathbf{H}) = \sum_j \rho_j W^{-1}(\mathbf{H}|\boldsymbol{\Psi_{ij}}, p)$, by a single inverse wishart distribution, $W^{-1}(\mathbf{H}|\boldsymbol{\Psi_{ij}}, p)$ is therefore given by

$$\sum_j \rho_j W^{-1}(\mathbf{H}|\boldsymbol{\Psi_{ij}}, p) \approx W^{-1}\left(\mathbf{H}|\left(\sum_j \rho_j {\boldsymbol{\Psi_j}}^{-1}\right)^{-1}, p\right)$$

Applying this approximation of the mixture of inverse wishart distributions to the likelihood given in B.1, the likelihood becomes

$$\begin{aligned}
&L\left(\mathbf{H}, \mathbf{x}^{(tr)}, \mathbf{x}^{(te)}\right) \\
&\propto |\mathbf{H}|^{mp} \prod_{i=1}^{m} \sum_{j=1}^{k} w_{ij} W^{-1}\left(\mathbf{H}|\boldsymbol{\Psi_{ij}}, p\right) \\
&\approx |\mathbf{H}|^{mp} \prod_{i=1}^{m} W^{-1}\left(\mathbf{H}|\left(\sum_{j}^{k} w_{ij} {\boldsymbol{\Psi_{ij}}}^{-1}\right)^{-1}, p\right) \\
&= |\mathbf{H}|^{mp} \prod_{i=1}^{m} \frac{\left|\left(\sum_{j}^{k} w_{ij} {\boldsymbol{\Psi_{j}}}^{-1}\right)^{-1}\right|^{\frac{p}{2}}}{2^{\frac{p^2}{2}} \Gamma_p\left(\frac{p}{2}\right)} |\mathbf{H}|^{-\frac{2p+1}{2}} e^{-\frac{1}{2}\operatorname{Tr}\left[\left(\sum_{j}^{k} w_{ij} {\boldsymbol{\Psi_{j}}}^{-1}\right)^{-1}\mathbf{H}^{-1}\right]} \\
&\propto |\mathbf{H}|^{mp} |\mathbf{H}|^{-\frac{2mp+m}{2}} e^{-\frac{1}{2}\sum_{i=1}^{m}\operatorname{Tr}\left[\left(\sum_{j}^{k} w_{ij} {\boldsymbol{\Psi_{j}}}^{-1}\right)^{-1}\mathbf{H}^{-1}\right]} \\
&= |\mathbf{H}|^{-\frac{m}{2}} e^{-\frac{1}{2}\operatorname{Tr}\left[\sum_{i=1}^{m}\left(\sum_{j}^{k} w_{ij} {\boldsymbol{\Psi_{j}}}^{-1}\right)^{-1}\mathbf{H}^{-1}\right]} \\
&= |\mathbf{H}|^{-\frac{m}{2}} e^{-\frac{1}{2}\operatorname{Tr}\left(\boldsymbol{\Psi}^{*}\mathbf{H}^{-1}\right)}
\end{aligned} \tag{B.9}$$

where $\boldsymbol{\Psi}^{*} = \sum_{i=1}^{m}\left(\sum_{j}^{k} w_{ij} {\boldsymbol{\Psi_{ij}}}^{-1}\right)^{-1}$

B.0.2 Approximate posterior distribution for the bandwidth matrix

The Jeffreys prior is known to be

$$\pi\left(\mathbf{A}\right) = \left|I\left(\mathbf{A}\right)\right|^{\frac{1}{2}}$$

for some matrix $\mathbf{A}$ and $I\left(\mathbf{A}\right) = -\mathbb{E}\left[\frac{\partial^2}{\partial\mathbf{A}\partial\mathbf{A}}\right]$ the Fisher information matrix. According to Yang and Berger (1994) the determinant of the Fisher information matrix for the covariance matrix $\mathbf{H}$ is

$$I\left(\mathbf{H}\right) \propto |\mathbf{H}|^{-(p+1)}$$

It follows that the Jeffrys prior is

$$\pi\left(\mathbf{H}\right) \propto |\mathbf{H}|^{\frac{-(p+1)}{2}} \tag{B.10}$$

Combining B.9 and B.11 using Bayes' rule, results in the posterior distribution of $\mathbf{H}$, which can be viewed as an inverse wishart distribution.

$$\begin{aligned}
\pi\left(\mathbf{H}|\mathbf{x}^{(tr)}, \mathbf{x}^{(te)}\right) &\propto |\mathbf{H}|^{\frac{-(p+1)}{2}} |\mathbf{H}|^{-\frac{m}{2}} e^{-\frac{1}{2}\operatorname{Tr}\left(\boldsymbol{\Psi}^{*}\mathbf{H}^{-1}\right)} \\
&= |\mathbf{H}|^{-\frac{m+p+1}{2}} e^{-\frac{1}{2}\operatorname{Tr}\left(\boldsymbol{\Psi}^{*}\mathbf{H}^{-1}\right)} \\
&\propto W^{-1}\left(\mathbf{H}|\boldsymbol{\Psi}^{*}, m\right)
\end{aligned}$$

It is important to take into account the restrictions $\boldsymbol{\Psi}^{*} > 0$ and $m > p - 1$.

B.0.3 Approximate posterior predictive distribution

The approximate posterior distribution for the bandwidth can now be used to approximate the posterior predictive distribution

$$
\begin{aligned}
&\pi\left(\mathbf{x}|\mathbf{x}^{(tr)},\mathbf{x}^{(te)}\right)\\
&\approx \frac{1}{k}\sum_{j=1}^{k}\int_{\mathbf{H}>0} MVN\left(\mathbf{x}|\mathbf{x}_j^{(tr)},\mathbf{H}\right) W^{-1}\left(\mathbf{H}|\mathbf{\Psi}^*,m\right)d\mathbf{H}\\
&= \frac{1}{k}\sum_{j=1}^{k}\int_{\mathbf{H}>0} \frac{1}{(2\pi)^{\frac{p}{2}}|\mathbf{H}|^{\frac{1}{2}}} e^{(\mathbf{x}-\mathbf{x}_j^{(tr)})^T\mathbf{H}^{-1}(\mathbf{x}-\mathbf{x}_j^{(tr)})} \frac{|\mathbf{\Psi}^*|^{\frac{m}{2}}}{2^{\frac{mp}{2}}\Gamma_p\left(\frac{m}{2}\right)}|\mathbf{H}|^{-\frac{m+p+1}{2}} e^{-\frac{1}{2}\mathrm{Tr}\left(\mathbf{\Psi}^*\mathbf{H}^{-1}\right)}d\mathbf{H}\\
&= \frac{1}{k}\sum_{j=1}^{k}\frac{|\mathbf{\Psi}^*|^{\frac{m}{2}}\Gamma_p\left(\frac{m+1}{2}\right)}{\pi^{\frac{p}{2}}\Gamma_p\left(\frac{m}{2}\right)\left|\Psi^*+(\mathbf{x}-\mathbf{x}_j^{(tr)})(\mathbf{x}-\mathbf{x}_j^{(tr)})^T\right|^{\frac{m+1}{2}}}\\
&\times\int_{\mathbf{H}>0}\frac{\left|\mathbf{\Psi}^*+(\mathbf{x}-\mathbf{x}_j^{(tr)})(\mathbf{x}-\mathbf{x}_j^{(tr)})^T\right|^{\frac{m+1}{2}}}{2^{\frac{(m+1)p}{2}}\Gamma_p\left(\frac{m+1}{2}\right)}|\mathbf{H}|^{-\frac{m+p+2}{2}}\\
&\times e^{-\frac{1}{2}\mathrm{Tr}\left(\mathbf{\Psi}^*\mathbf{H}^{-1}+(\mathbf{x}-\mathbf{x}_j^{(tr)})(\mathbf{x}-\mathbf{x}_j^{(tr)})^T\mathbf{H}^{-1}\right)}d\mathbf{H}\\
&= \frac{1}{k}\sum_{j=1}^{k}\frac{|\mathbf{\Psi}^*|^{\frac{m}{2}}\Gamma_p\left(\frac{m+1}{2}\right)}{\pi^{\frac{p}{2}}\Gamma_p\left(\frac{m}{2}\right)\left|\mathbf{\Psi}^*+(\mathbf{x}-\mathbf{x}_j^{(tr)})(\mathbf{x}-\mathbf{x}_j^{(tr)})^T\right|^{\frac{m+1}{2}}} \qquad \text{(B.11)}\\
&= \frac{1}{k}\sum_{j=1}^{k}\frac{|\mathbf{\Psi}^*|^{\frac{m}{2}}\Gamma_p\left(\frac{m+1}{2}\right)}{\pi^{\frac{p}{2}}\Gamma_p\left(\frac{m}{2}\right)}|\mathbf{\Psi}^*|^{-\frac{m+1}{2}}\left|\mathbf{I}+\mathbf{\Psi}^{*-1}(\mathbf{x}-\mathbf{x}_j^{(tr)})(\mathbf{x}-\mathbf{x}_j^{(tr)})^T\right|^{-\frac{m+1}{2}}\\
&= \frac{1}{k}\sum_{j=1}^{k}\frac{\Gamma_p\left(\frac{m+1}{2}\right)}{\pi^{\frac{p}{2}}\Gamma_p\left(\frac{m}{2}\right)|\mathbf{\Psi}^*|^{\frac{1}{2}}}\left|\mathbf{I}+\mathbf{\Psi}^{*-1}(\mathbf{x}-\mathbf{x}_j^{(tr)})(\mathbf{x}-\mathbf{x}_j^{(tr)})^T\right|^{-\frac{m+1}{2}}\\
&= \frac{1}{k}\sum_{j=1}^{k}\frac{\Gamma\left(\frac{m+1}{2}\right)}{\pi^{\frac{p}{2}}\Gamma\left(\frac{m+1-p}{2}\right)|\mathbf{\Psi}^*|^{\frac{1}{2}}}\left|\mathbf{I}+\mathbf{\Psi}^{*-1}(\mathbf{x}-\mathbf{x}_j^{(tr)})(\mathbf{x}-\mathbf{x}_j^{(tr)})^T\right|^{-\frac{m+1}{2}} \qquad \text{(B.12)}\\
&= \frac{1}{k}\sum_{j=1}^{k}\frac{\Gamma\left(\frac{m+1}{2}\right)}{\pi^{\frac{p}{2}}\Gamma\left(\frac{m+1-p}{2}\right)(m+1-p)^{\frac{p}{2}}\left|\frac{1}{m+1-p}\mathbf{\Psi}^*\right|^{\frac{1}{2}}}\\
&\times\left[1+\frac{1}{m+1-p}(\mathbf{x}-\mathbf{x}_j^{(tr)})^T\left(\frac{1}{m+1-p}\mathbf{\Psi}^*\right)^{-1}(\mathbf{x}-\mathbf{x}_j^{(tr)})\right]^{-\frac{m+1}{2}} \qquad \text{(B.13)}\\
&= \frac{1}{k}\sum_{j=1}^{k}t_{m-p+1}\left(\mathbf{x},\mathbf{x}_j^{(tr)},\frac{1}{m+1-p}\mathbf{\Psi}^*\right)
\end{aligned}
$$

The posterior predictive distribution can thus be approximated by a mixture of multivariate Student's t distributions, with each distribution having centre $\mathbf{x}_i$. Take note that Equation

B.11 follows from the fact that

$$\begin{aligned}
&\int_{\mathbf{H}>0} \frac{\left|\boldsymbol{\Psi}^{*}+(\mathbf{x}-\mathbf{x}_j^{(tr)})(\mathbf{x}-\mathbf{x}_j^{(tr)})^T\right|^{\frac{m+1}{2}}}{2^{\frac{(m+1)p}{2}}\Gamma_p\left(\frac{m+1}{2}\right)} |\mathbf{H}|^{-\frac{m+p+2}{2}} e^{-\frac{1}{2}\operatorname{Tr}\left(\boldsymbol{\Psi}^{*}\mathbf{H}^{-1}+(\mathbf{x}-\mathbf{x}_j^{(tr)})(\mathbf{x}-\mathbf{x}_j^{(tr)})^T\mathbf{H}^{-1}\right)} d\mathbf{H} \\
&= \int_{\mathbf{H}>0} W^{-1}\left(\mathbf{H}|\boldsymbol{\Psi}^{*}+(\mathbf{x}-\mathbf{x}_j^{(tr)})(\mathbf{x}-\mathbf{x}_j^{(tr)})^T, m+1\right) d\mathbf{H} \\
&= 1
\end{aligned}$$

and Equation B.12 is true since

$$\begin{aligned}
\frac{\Gamma_p\left(\frac{m+1}{2}\right)}{\Gamma_p\left(\frac{m}{2}\right)} &= \frac{\pi^{\frac{p(p-1)}{4}}\prod_{i=1}^{p}\Gamma\left(\frac{m+1}{2}+\frac{1-i}{2}\right)}{\pi^{\frac{p(p-1)}{4}}\prod_{i=1}^{p}\Gamma\left(\frac{m}{2}+\frac{1-i}{2}\right)} \\
&= \frac{\Gamma\left(\frac{m+1}{2}\right)\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{m-1}{2}\right)\ldots\Gamma\left(\frac{m+3-p}{2}\right)\Gamma\left(\frac{m+2-p}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{m-1}{2}\right)\ldots\Gamma\left(\frac{m+3-p}{2}\right)\Gamma\left(\frac{m+2-p}{2}\right)\Gamma\left(\frac{m+1-p}{2}\right)} \\
&= \frac{\Gamma\left(\frac{m+1}{2}\right)}{\Gamma\left(\frac{m+1-p}{2}\right)}
\end{aligned}$$

Finally Equation B.13 can be seen to be true by considering the following identities; Let $\mathbf{A}$ be some $k \times l$, $\mathbf{B}$ some $l \times k$ and $\mathbf{D}$ some $n \times n$ matrix and let c be some constant, then

$$|\mathbf{I}_k + \mathbf{AB}| = |\mathbf{I}_l + \mathbf{BA}|$$

and

$$|c\mathbf{D}| = c^n |\mathbf{D}|$$

References/Bibliography

Bernardo, J.M. (1999). "Model-free objective Bayesian prediction". In: *Rev. Acad. Ciencias de Madrid* 93, pp. 295–302.

Bishop, C.M. (2006). *Pattern recognition and machine learning*. Springer.

Castillo, F. et al. (2003). "Real World Applications-A Methodology for Combining Symbolic Regression and Design of Experiments to Improve Empirical Model Building". In: *Lecture Notes in Computer Science* 2724, pp. 1975–1985.

De Lima, M.S., Pereira, J.R.G., and Souza, D.S. (2013). "Bayesian predictive kernel discriminant analysis". In: *Pattern Recognition Letters* 34.16, pp. 2079–2085.

Féraud, R. and Clérot, F. (2002). "A methodology to explain neural network classification". In: *Neural Networks* 15.2, pp. 237–246.

Fienberg, S.E. et al. (2006). "When did Bayesian inference become "Bayesian"?" In: *Bayesian Analysis* 1.1, pp. 1–40.

Giudici, P. and Figini, S. (2009). *Applied Data Mining for Business and Industry*. Wiley.

Gujarati, D.N. and Porter, D.C. (1999). *Essentials of econometrics*. McGraw-Hill Singapore.

Hastie, T., Tibshirani, R., and Friedman, J. (2008). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer, New York.

Holmes, C. and Knorr-Held, L. (2003). *Efficient simulation of Bayesian logistic regression models*. Tech. rep. Discussion papers/Sonderforschungsbereich 386 der Ludwig-Maximilians-Universität München.

Johnson, R. and Wichern, D. (2014). *Applied multivariate statistical analysis*. Vol. 5. 8. Pearson Education Limited.

Kennedy, K., Mac Namee, B., and Delany, S.J. (2012). "Using semi-supervised classifiers for credit scoring". In: *Journal of the Operational Research Society* 64.4, pp. 513–529.

Lending club (2016). *Lending club loan data*. URL: `https://www.kaggle.com/wendykan/lending-club-loan-data/data`.

Lichman, M. (2013). *UCI Machine Learning Repository*. URL: `http://archive.ics.uci.edu/ml`.

Lohr, S.L. (2010). *Sampling: Design and Analysis*. Ed. by Michelle Julet. 2nd. Richard Stratton.

Longadge, R. and Dongre, S. (2013). "Class imbalance problem in data mining review". In: *arXiv preprint arXiv:1305.1707*.

Mahalanobis, C.P. et al. (1936). "On the generalised distance in statistics". In: *Proceedings of the National Institute of Sciences of India*. Vol. 2. 1, pp. 49–55.

Mester, L.J. et al. (1997). "Whats the point of credit scoring?" In: *Business review* 3.Sep/Oct, pp. 3–16.

Murphy, K.P. (2012). *Machine learning: a probabilistic perspective*. MIT press.

Nath, R., Rajagopalan, B., and Ryker, R. (1997). "Determining the saliency of input variables in neural network classifiers". In: *Computers & Operations Research* 24.8, pp. 767–773.

Shannon, C.E. (2001). "A mathematical theory of communication". In: *ACM SIGMOBILE Mobile Computing and Communications Review* 5.1, pp. 3–55.

Van der Walt, C.M. (2014). "Maximum-likelihood kernel density estimation in high-dimensional feature spaces". PhD thesis. North-West University.

Van der Walt, C.M. and Barnard, E. (2013). "Kernel bandwidth estimation for non-parametric density estimation: a comparative study". In: *Proceedings of the Twenty-Fourth Annual Symposium of the Pattern Recognition Association of South Africa*. Johannesburg, South-Africa.

Van der Walt, C.M. and Barnard, E. (2017). "Variable Kernel Density Estimation in High-Dimensional Feature Spaces". In: *Thirty-First AAAI Conference on Artificial Intelligence*.

West, M. and Harrison, J. (1997). *Bayesian Forecasting and Dynamic Models*. Springer-Verlag New York.

Yang, R. and Berger, J.O. (1994). "Estimation of a covariance matrix using the reference prior". In: *The Annals of Statistics*, pp. 1195–1211.

Zekic-Susac, M., Sarlija, N., and Bensic, M. (2004). "Small business credit scoring: a comparison of logistic regression, neural network, and decision tree models". In: *Information Technology Interfaces, 2004. 26th International Conference on*. IEEE, pp. 265–270.