Prediction Methods for Astronomical Data Observed with Measurement Error

by

James Patrick Long

A dissertation submitted in partial satisfaction of the requirements for the degree of Doctor of Philosophy

in

Statistics

in the

Graduate Division

of the

University of California, Berkeley

Committee in charge:

Professor John A. Rice, Co-chair
Associate Professor Noureddine El Karoui, Co-chair
Professor Bin Yu
Associate Professor Joshua S. Bloom

Fall 2013
Prediction Methods for Astronomical Data Observed with Measurement Error

Copyright 2013
by
James Patrick Long
Abstract

Prediction Methods for Astronomical Data Observed with Measurement Error

by

James Patrick Long

Doctor of Philosophy in Statistics
University of California, Berkeley

Professor John A. Rice, Co-chair
Associate Professor Noureddine El Karoui, Co-chair

We study prediction when features are observed with measurement error. The research is motivated by classification challenges in astronomy.

In Chapter 1 we introduce the periodic variable star classification problem. Periodic variable stars are periodic functions which belong to a particular physical class. These functions are often sparsely sampled, which introduces measurement error when attempting to estimate period, amplitude, and other function features. We discuss how measurement error can impact performance of periodic variable star classifiers. We introduce two general strategies, noisification and denoisification, for addressing measurement error in prediction problems.

In Chapter 2 we study density estimation with Berkson error. In this problem, one observes a sample $X_1, \ldots, X_n \sim f_X$ and seeks to estimate $f_Y$, the convolution of $f_X$ with a known error distribution. We derive asymptotic results for the behavior of the mean integrated squared error for kernel density estimates of $f_Y$. The presence of error generally increases convergence rates of estimators and optimal smoothing parameters. We briefly discuss some potential applications for this work, including classification tasks involving measurement error.

In Chapter 3 we study prediction of a continuous response for an observation with measurement error in its features. Using Nadaraya Watson type estimators we derive limit theorems for convergence of the mean squared error as a function of the smoothing parameters.

In Chapter 4 we study the effects of measurement error on classifier performance using data from the Optical Gravitational Lensing Experiment (OGLE) and the Hipparcos satellite. We illustrate some challenges in constructing statistical classifiers when the training data is collected by one astronomical survey and the unlabeled data is collected by a different survey.
We use noisification to construct classifiers that are robust to some sources of measurement error and training–unlabeled data set differences.
This thesis is dedicated to my fiancée, Mika.
Contents

1 Introduction
  1.1 Periodic Variable Stars
  1.2 Classification of Periodic Variables
  1.3 Measurement Error
  1.4 Strategies for Addressing Measurement Error
  1.5 Summary of Chapters

2 Density Estimation with Berkson Error
  2.1 Introduction
  2.2 Problem Setup
  2.3 Error Term has Density
  2.4 Error Term has Support Restricted to Subspace
  2.5 Conclusions
  2.6 Proofs and Lemmas
  2.7 Technical Notes

3 Nadaraya–Watson for Prediction with Feature Measurement Error
  3.1 Introduction
  3.2 Problem Setup
  3.3 Expansion for Bias and Variance
  3.4 Asymptotic Approximations to Bias and Variance
  3.5 Selection of Optimal Bandwidth
  3.6 Asymptotic MSE May Be Invariant to Smoothing: Example
  3.7 Conclusions
  3.8 Theorems and Lemmas

4 Empirical Study of Noisification
  4.1 Introduction
  4.2 Overview of Classification of Periodic Variables
  4.3 Feature Distributions and Survey Systematics
Acknowledgments

This thesis would not have been possible without the help of many individuals over the past five years. First and foremost, I would like to thank my advisors, John Rice and Noureddine El Karoui, who have provided generous amounts of mentorship, encouragement, and time.

I would like to thank Bin Yu for serving as an informal mentor when I was a Graduate Student Instructor. She taught me much about applied statistics and how to conduct research.

Joshua Bloom and Joseph Richards provided generous quantities of help and advice, and played a major role in formulating many of the questions studied in this work.

I would like to thank my parents, Ann and Tom Long. They were my first teachers and have provided many years of love, support, advice, and encouragement.

Throughout my years at Berkeley, I have been supported by a National Science Foundation (NSF) VIGRE grant (0636667), a Cyber-Enabled Discovery and Innovation NSF grant (0941742), and a fellowship from Citadel LLC. I would like to thank the NSF and Citadel for their generous support.
Chapter 1
Introduction

In this thesis we study classification of periodic variable stars and statistical challenges that arise when performing density estimation, regression, and classification on data observed with measurement error. We begin by introducing the astronomical context of the periodic variable star classification problem in Sections 1.1 and 1.2. We demonstrate how the sparse, irregular sampling of periodic variable stars introduces measurement error into derived features in Section 1.3. In Section 1.4 we propose two general strategies: noisification and denoisification for constructing estimators in measurement error problems. In Section 1.5 we briefly summarize the contents of each chapter in this work. Notation and necessary background material is reviewed in each chapter so that they may be read individually.

1.1 Periodic Variable Stars

The size of modern astronomical surveys prohibits direct human involvement in many prediction tasks. As a result, astronomers are using statistical and machine–learning methods for prediction. One important challenge is the classification of periodic variable stars. Variable stars are stars and stellar systems that vary in the amount of light they emit over time. Periodic variable stars are variables which exhibit periodic variation in the amount of light they emit.

Most periodic variable star data is collected by photometric survey telescopes that scan the sky, taking many images over the course of months or years. Images are processed by a reduction pipeline which detects sources (usually stars or galaxies) and performs photometry i.e., estimation of the intensity of light being emitted by the source at the time the image was taken. Telescopes measure the flux, energy per area-time, of a source in some bandpass range of wavelength. The flux $f$ of a source is then converted into apparent magnitude $m$
CHAPTER 1. INTRODUCTION

Figure 1.1: Light curve of a Beta Lyrae star from the *Hipparcos* survey (HIP ID 100253). Note the irregular time sampling. The *Hipparcos* telescope often takes several measurements almost simultaneously followed by long gaps. Each magnitude observation has a measurement error represented by the vertical intervals.

by

\[ m = -2.5 \log(f) + C \]

where \( C \) is some constant. Note that brighter objects have lower apparent magnitudes and vice versa. See Ryden and Peterson [2010] Section 13.2 for a description of how star brightness is measured. The reduction pipeline also estimates uncertainty, \( \sigma \), on each brightness measurement. The uncertainty is caused by variation in photons emitted by the source, background light, and CCD detector noise. In this thesis we study high signal-to-noise (i.e., bright) sources that can be clearly differentiated from random background fluctuations in the images.

Over the course of a survey, many images are taken of the same region of the sky, resulting in many magnitude and magnitude error measurements for each source. Thus for source \( i \), the reduction pipeline produces \( r_i \) 3-tuples \( \{(t_{ij}, m_{ij}, \sigma_{ij})\}_{j=1}^{r_i} \) where \( m_{ij} \) and \( \sigma_{ij} \) are the magnitude and magnitude error estimates for source \( i \) at time \( t_{ij} \). See Udalski et al. [2008a] for a description of a reduction pipeline for the Optical Gravitational Lensing Experiment III (OGLE-III) survey and the website http://ogledb.astrouw.edu.pl/~ogle/CVS/ for the resulting photometry. Recent surveys include *Hipparcos* which concluded operation in 1993 and OGLE which completed phase III in 2009 [Perryman et al., 1997b, Udalski et al., 2008a]. Upcoming surveys include Gaia, launching late 2013, and the Large Synoptic Survey Telescope, starting around 2020 [Eyer et al., 2010, Sweeney, 2006].

For a particular source, astronomers often study magnitude as a function of time. This
Figure 1.2: Folded light curve (see text for definition) of Beta Lyrae in Figure 1.1. Structure useful for class separation is clearly visible in folded light curves. For example the two dips of different depth suggest this is an eclipsing binary with each star emitting a different intensity of light.

function is known as the light curve. Figure 1.1 presents a light curve of a variable star from the Hipparcos survey. We note several important features of this light curve that are common to the data we analyze. The x-axis is time in days with day 0 assigned to be the first day of observation. On the y-axis is apparent magnitude. The black intervals around the dot represent uncertainty, $\sigma_{ij}$ in the magnitude measurements. The level of uncertainty varies from measurement to measurement as well as systematically across surveys.

The light curve is sampled at irregular intervals, sometimes with large gaps. The sampling times of the source are known as the cadence.\footnote{Note that other authors define cadence somewhat differently. For example, Nemiroff [2007] defines cadence as “the average frequency of return [by the telescope] to image the same field.”} Large gaps may be due to the source being behind the sun for several months of the year and as a result not visible by the telescope. Weather conditions also contribute to irregular sampling. Clouds or rain may make it impossible to observe on a given night. Finally, for space telescopes, orbital stability of the instrument dictates where the telescope may be pointed at any given time. Astronomical surveys often observe many stars with approximately the same cadence. Thus we may speak of the survey as having a cadence or particular cadence properties.

Figure 1.1 is a periodic variable of the class Beta Lyrae. Beta Lyrae stars are eclipsing binaries; the change in brightness is caused by each star in the system periodically blocking light from the other star. Determining the class of a particular variable star is typically far easier after estimating the star’s period and plotting magnitude versus phase. This is known
Figure 1.3: Light curve of a folded RR Lyrae AB (HIP ID 101356). RR Lyrae have a linear decline for a large fraction of their period followed by a sudden spike in brightness. The folded light curve shape of RR Lyrae stars differ markedly from eclipsing binaries, such as in Figure 1.2.

as the *folded* light curve. Phase is computed by replacing the times with (time modulo period) / period. The period may be estimated using methods such as the Lomb-Scargle periodogram [Lomb, 1976] or the Nadaraya–Watson estimator [Hall, 2008].

Figure 1.2 shows the folded light curve of the Beta Lyrae in Figure 1.1. The dips in brightness occur at the eclipses. Beta Lyrae is one of several classes of eclipsing binary sources. According to Sterken and Jaschek [1996] (Section 6.2), the Beta Lyrae class is defined in terms of light curve shape, rather than underlying astrophysical cause of variation. In the class Beta Lyrae, the light curve varies continuously between eclipses. Another class of eclipsing binaries, Beta Persei (or Algol), has constant brightness between eclipses. See Sterken and Jaschek [1996] Chapter 6 for eclipsing binary class definitions. In Figure 1.3 we plot the folded light curve of a RR Lyrae AB star. RR Lyrae are characterized by a sharp linear increase in light at one part of their period followed by a slow linear decline. For the RR Lyrae AB in Figure 1.3, the sharp linear increase occurs around 0.6 in phase space. RR Lyrae AB are intrinsic variables meaning that they vary due to pulsational processes occurring within the star itself, unlike eclipsing classes of periodic variables [Sterken and Jaschek, 1996].

There is no universally recognized taxonomy of variable star classes. Rather, different authors construct somewhat different taxonomies depending on scientific goals. Recent works on periodic variable classification have used 26 [Dubath et al., 2011], 25 [Richards et al., 2011], and 35 [Debosscher et al., 2007] classes. Class structure is roughly hierarchical and
CHAPTER 1. INTRODUCTION

classification taxonomies often differ in how far to subdivide classes. For example, an individual particularly interested in studying RR Lyrae stars would likely want a classifier that separated distinct RR Lyrae subtypes (AB, C, D, and E) while other individuals may only be interested in having a general RR Lyrae label. See Richards et al. [2011] Figure 2 for an example of a hierarchical class structure used in a recent work. See Sterken and Jaschek [1996] for background on class definitions, the astrophysical mechanisms causing brightness variation in variable stars, and discussion of class taxonomy.

1.2 Classification of Periodic Variables

The size of modern periodic variable star data sets dictates that much of the classification work be automated. For example, the 4-year Hipparcos mission ending in 1993 catalogued 2712 periodic variable stars, making manual classification of every source possible [Dubath et al., 2011]. More recently, the OGLE-III survey, completed in 2009, cataloged 167,251 variable sources, making manual classification difficult.² Upcoming surveys such as Gaia and the Large Synoptic Survey Telescope are projected to collect in the millions or billions of periodic variable sources, making direct human involvement in every classification decision impossible [Eyer and Cuypers, 2000, Borne et al., 2007]. The necessity of automated classification has generated much interest in developing statistical and machine learning tools for periodic variables (e.g. Eyer and Blake [2005], Debosscher et al. [2007], Richards et al. [2011], Dubath et al. [2011], Sarro and Debosscher J. [2008], Debosscher et al. [2009], Eyer et al. [2008]).

Often the data used to construct the classifier, the training data, and the data to which the classifier will be applied, the unlabeled data, have systematically different cadences or magnitude errors. Alternatively the best observed (most number of measurements per light curve) sources from a survey may be manually classified and then used as training data to construct a classifier on the poorly observed sources from the same survey. See Section 4.1 for a discussion of training and unlabeled data sets used in recent works on variable star classification.

When the training data and the unlabeled data are not identically distributed, statistical classifiers may perform poorly. Cross–validated error rates measured using the training data may not provide an accurate estimate of error for the unlabeled data. Tuning parameters optimized for the training data may not be optimal for the unlabeled data. Perhaps most

²There are 167,251 sources in the OGLE-III Online Catalog of Variable Stars (http://ogledb.astrouw.edu.pl/~ogle/CVS/) as of April 1, 2013. The vast majority of these are periodic variables. Members of the OGLE collaboration classified these sources using a mixture of period search algorithms, manual cuts on features, and visual inspection. For example, see Soszynski et al. [2009b] for classification of RR Lyrae in OGLE-III.
importantly, a classifier constructed on the training data may use features that do not separate classes in the unlabeled data, causing the classifier to have a high error rate. In the remainder of this section we formalize the statistical model for training and unlabeled set differences and discuss feature extraction for periodic variable star classification.

1.2.1 Problem Formalization

Let \( (G, Z) \) be a joint distribution over the set of periodic functions \( G \) and integers (representing classes) \( Z = \{1, \ldots, K\} \). There are \( n \) training observation, drawn independently from this distribution. In other words

\[
(g_i, Z_i) \sim (G, Z)
\]

independently for \( \{i \in 1, \ldots n\} \). We do not observe \( g_i \) directly, but noisy samples from this function. Specifically let \( (C_T, E_T) \) be a distribution across cadences and magnitude errors (\( T \) refers to training). For observation \( i \) we draw \( \{(t_{ij}, \sigma_{ij})\}_{j=1}^{r_i} \sim (C_T, E_T) \). The \( t_{ij} \) are times at which \( g_i \) is observed and \( \sigma_{ij} \) is the standard error on the magnitude measurement at time \( t_{ij} \). Specifically

\[
m_{ij} = g_i(t_{ij}) + e_{ij}
\]

where \( e_{ij} \sim N(0, \sigma_{ij}^2) \) independent across \( i, j \).\(^3\) Define \( l_i = \{(t_{ij}, m_{ij}, \sigma_{ij})\}_{j=1}^{r_i} \). The observed data is \( (l_i, Z_i) \) for \( i \in \{1, \ldots, n\} \).

Denote the unlabeled observations \( l_i^* \) for \( i \in \{1, \ldots, N\} \). Our aim is to predict the unobserved classes \( \{Z_i^*\}_{i=1}^{N} \) for this data. The underlying functions and classes from which these measurements are drawn is assumed to be the same as for the training data, \( (G, Z) \). In addition, Equation (1.1) relates the various components of each \( l_i^* \). However the cadence and magnitude distribution are drawn from \( (C_U, E_U) \) which is different from the training data. Thus the training and unlabeled data are not identically distributed. In Chapter 4 we introduce specific data sets to demonstrate how the cadence and amount of magnitude noise varies between training and unlabeled data. We note several assumptions made by the problem formalization:

1. The probability distribution \( (G, Z) \) is the same for training and unlabeled data.

2. The reported standard errors \( \sigma_{ij} \) are correct and the error is normal and independent across different times.

3. The sampling times and errors \( \{(t_{ij}, \sigma_{ij})\}_{j=1}^{r_i} \) are independent of the underlying function \( g_i \).

\(^3\)The normal distribution is often a good approximation to the true error distribution when the signal-to-noise (brightness of the source relative to background brightness of the image) is high.
4. The functional space $G$ we are drawing from includes only periodic functions.

The validity of these assumptions and any subsequent impact on classifier performance varies from data set to data set. We discuss the validity of the assumptions, the impact on classifier performance and extensions to this model in Chapter 4.

### 1.2.2 Feature Extraction

A popular classification strategy for periodic variable stars involves feature extraction. In this framework, $p$ functions of the training observations are computed. These functions, known as features, are chosen to separate classes. For example, with periodic variable stars amplitude and period are known to be useful. See Richards et al. [2011] and Dubath et al. [2011] for studies of what features are useful for separating which classes. Define the map from the set of light curves, denoted $\ell$, to features as $\mathcal{X}: \ell \to \mathbb{R}^p$. One can compute features for each training light curve ($\mathcal{X}(l_i) = X_i$) and then obtain a representation of the training data as $\{(X_i, Z_i)\}_{i=1}^n$ where $X_i \in \mathbb{R}^p$.

Many classification methods (e.g., Random Forests, Support Vector Machines, Neural Nets, Linear Discriminant Analysis) accept data of the form $\{(X_i, Z_i)\}_{i=1}^n$. These methods will construct a classifier $\hat{C}: \mathbb{R}^p \to \{1, \ldots, K\}$. The goal is for this classifier to have a low error rate when applied to the features of the unlabeled data, specifically $\hat{C}(\mathcal{X}(l_i^*)) = Z_i^*$ with high probability. Note that the distribution of features is a function of the cadence and magnitude error. Thus when cadence and magnitude error distributions differ between training and unlabeled data sets, feature distributions may differ as well.

### 1.3 Measurement Error

We now show how cadence and magnitude distributions introduce measurement error into derived features. This motivates study of measurement error models in Chapters 2 and 3. Consider the continuous periodic function in Figure 1.4. It is meant to resemble a star belonging to the class RR Lyrae. The function has a period of about 0.72 days and peak-to-peak amplitude about 0.6 magnitudes. Period and amplitude are very useful features for separating classes, so nearly any feature extraction approach will estimate these two quantities from the set of times, magnitudes, and magnitude error measurements recorded by the telescope (i.e., estimate period and amplitude using $l_i$).

We study how accurately one can estimate period and amplitude from $l_i$ by sampling this function different numbers of times. Specifically we take the cadence and magnitude error
from a light curve observed by the Optical Gravitational Lensing Experiment (OGLE).\textsuperscript{4} We truncate the cadence at the first $s = 20, 40, 60,$ and $80$ measurements. For each $s$ we sample from the function in Figure 1.4 $50$ times, randomizing the time of the initial observation (uniform across the period of the function) and draws from the magnitude errors. We compute features for each of these simulated light curves. Figure 1.5 contains four plots (one for each $s$) with the estimated features. Additionally the true features are marked. For 20 measurements there is considerable scatter in the amplitude. The period estimate is either very close to correct or completely wrong. At 40 measurements the period is generally estimated correctly, but there is still scatter in the amplitude measure. This scatter goes down, but does not disappear, for the functions observed 60 and 80 times.

In this example, the cadence and magnitude error sampling of the function induce measurement error into the features. The more poorly sampled the light curve, the greater the measurement error. This observation motivates our study of measurement error models for density and function estimation in Chapters 2 and 3. In these Chapters we focus on the case where there is error only in the features of the unlabeled observations. This is approximately true for the periodic variable star problem when the training light curves are well observed ($> 100$ measurements) and the unlabeled light curves are poorly observed.

Note that the distribution of cadences and magnitude errors for the unlabeled data $(C_U, E_U)$ may produce light curves of quite varying quality. For example some draws from this distribution may sample the function 20 times while others may sample the function 80 times. As shown in Figure 1.5, this produces different levels of error in the derived features. Thus we may want to adapt any classifier not simply to the general $(C_U, E_U)$ distribution, but to a particular light curve cadence and set of magnitude errors. We discuss this concept

\textsuperscript{4}Cadence taken from OGLE-LMC-CEP-2233
Figure 1.5: Feature error distributions for light curves sampled different numbers of times. The measurement error in the period and amplitude features decreases as the number of measurements per function increases.

further in the following section when introducing methods for addressing feature error. In Chapter 4 we experiment with developing classifiers for subsets of the unlabeled data that have similar cadences and magnitude errors and thus similar feature error distributions.

1.4 Strategies for Addressing Measurement Error

We now describe two general strategies, \textit{noisification} and \textit{denoisification}, for constructing variable star classifiers when the training and unlabeled data have different cadence and magnitude error distributions. These methodologies are not closely bound to the particular problem of variable star classification and may be applied to many problems where there is
CHAPTER 1. INTRODUCTION

Figure 1.6: Data from hypothetical classification problem. The black dot is an unlabeled observation. The cross hairs represent measurement error on the features. The red line classifier, constructed without using the measurement error, will assign the unlabeled observation to class 1. The error distribution suggests that the unlabeled observation actually belongs to class 2.

measurement error in the features of the unlabeled observations.

1.4.1 Noisification

Noisification matches features distributions by altering the training data features to match the distribution of features in the unlabeled data. This is called noisification because in most examples we consider the features of the unlabeled data have more measurement error (i.e., noise) than the training data features. Thus altering the training feature distribution to match the distribution in the unlabeled data involves adding noise, or noisifying, the training data. We now describe noisification for a simple classification problem with measurement error. The purpose of this example is to clarify the definition of noisification and motivate specific research questions.

Consider a classification problem with training data features and class \( \{(X_i, Z_i)\}_{i=1}^n \). Let

\[
Y = X + \epsilon
\]
be features observed with measurement error for an observation we wish to classify. $X$ are the unknown true features for this observation which follow the same distribution as the training features. The measurement error has distribution $F$, which is known. $\epsilon$ and $X$ are independent. $Z$ is the unknown class for this observation.

Consider constructing a classifier $\hat{C}$ on the training data and applying it directly to $Y$ in order to predict $Z$. Figure 1.6 illustrates this procedure. Here the training data belong to one of two classes (class 1 is orange plus and class 2 is blue triangle). The red vertical line represents a classifier $\hat{C}$ that might be learned by a method such as Classification and Regression Trees (CART, see Breiman et al. [1984] for a description of this classifier). Any observation with a Feature 1 value greater than 1 is assigned to class 1. Otherwise the observation is assigned to class 2. The black dot represents the noisy features $Y$ of an unlabeled observation. The cross hairs represent uncertainty on the features. If one uses the red line classifier, this observation is assigned to class 1. However the shape of the measurement error strongly suggests that the observation is actually from class 2.

With noisification one adds the measurement error to each of the training observations and then constructs a classifier on this data. This measurement error should match the error
present in the unlabeled observation. Specifically,

1. Draw $\epsilon_i \sim F_\epsilon$ for $i \in \{1, \ldots, n\}$
2. $Y_i \leftarrow X_i + \epsilon_i$
3. Construct classifier on $\{(Y_i, Z_i)\}_{i=1}^n$.

Note that the data on which we are constructing a classifier $\{(Y_i, Z_i)\}_{i=1}^n$ is now identically distributed with the observation we want to classify $(Y, Z)$. The result of noisification is illustrated in Figure 1.7. Here noise has been added to the features for the training observations in the shape of the cross hairs. Notice there is much more spread on Feature 1 here than in Figure 1.6. There is slightly more spread on Feature 2. However since there was little noise on Feature 2, this feature has remained relatively unchanged by noisification. The red line is a classifier that might be constructed on the noisified data plotted here. Notice the classifier now splits the data on Feature 2. A classifier is unlikely to make a vertical cut on Feature 1 on the noisified data because this feature no longer separates the classes. Essentially by adding noise to the training data, a classifier is able to determine which features are useful for separating classes in the noisy feature space. Notice that this classifier will classify the black dot as belonging to class 2.

In this example, the measurement error $\epsilon$ is additive and independent of $X$. For variable star classification, the feature error distribution is induced by the sampling of the light curve and as a result is unlikely to be well approximated by independent, additive error. This makes sampling from the feature error distribution far more difficult. The noisification process for light curves is described in detail in Section 4.4. The procedure involves resampling the training light curves at the cadence, magnitude error distribution $(C_U, E_U)$. We note several aspects of noisification that will be discussed in later sections and chapters:

1. Noisification may be applied to regression problems where features are observed with measurement error. We study regression with feature measurement error in Chapter 3.
2. For the variable star classification problem, each unlabeled light curve will have its own feature measurement error based on the particular cadence and magnitude errors at which it is observed. If all cadence and magnitude error draws from $(C_U, E_U)$ are similar, then the feature errors may have similar distributions and it may make sense to noisify all the training data to the “average” measurement error in the unlabeled data and then construct a single classifier on this noisified data. In other cases $(C_U, E_U)$ might produce very different sets of cadences and magnitude errors which will induce very different measurement error properties in each unlabeled light curve. Here it may
make sense to subdivide the unlabeled data into sets of light curves that have similar cadences and levels of measurement error. Then a separate noisified classifier can be constructed for each of these sets. These issues are discussed using real and simulated data in Chapter 4.

3. In principle, the noisification procedure can be repeated many times and a new classifier can be constructed on each noisified training set. The resulting classifiers could be averaged together or voted. Intuitively, repeated noisification may reduce the variance of the classifier, possibly improving performance. We study this in Sections 4.4 and 4.5 for the periodic variable star classification problem. Note that many repetitions of the procedure may become computationally expensive.

1.4.2 Denoisification

Denoisification matches feature distributions between training and unlabeled data by probabilistically inferring the true features for an unlabeled observation. This strategy is motivated by the following argument. As before let $Y$ be the noisy features of an unlabeled observation. Our goal is to construct a classifier for $Y$. A closely related problem is estimation of $p(z|y)$, the conditional probability of an observation with noisy features $y$ belonging to class $z$. The Bayes classifier with 0-1 loss is $\arg\max_k p(z = k|y)$, so knowing $p(z|y)$ allows one to construct the optimal classifier. Let $p(z|x)$ be the conditional probability of class given true features. Let $p(x|y)$ be the conditional probability density of true features given some features $y$ that are observed with measurement error. Under the assumption that $p(z|x, y) = p(z|x)$ (i.e. given the noise–free features $x$, the noisy features $y$ contain no additional information about the class $z$) we have,

$$p(z|y) = \int p(z, x|y)dx$$
$$= \int p(z|x, y)p(x|y)dx$$
$$= \int p(z|x)p(x|y)dx.$$ 

The first equality is by the law of total probability and the second equality is by Bayes theorem. We call

$$p(z|y) = \int p(z|x)p(x|y)dx \quad (1.2)$$

the denoisification identity. This identity suggests that one may estimate $p(z|y)$ using estimates of $p(z|x)$ and $p(x|y)$. $p(z|x)$ may be estimated via a probabilistic classifier constructed on the training data. We note that in the case where the density $p(x|y)$ does not exist,
Figure 1.8: Illustration of denoisification. The red line is a classifier constructed on the training data (see Figure 1.6). This classifier is combined with an estimate of $p(x|y)$ (represented by the grey region) to predict the class of the unlabeled observation (black dot). Equation 1.2 shows how to combine a probabilistic classifier and an estimate of $p(x|y)$ to produce a class prediction for the unlabeled observation.

the identity $p(z|y) = \int p(z|x)df_{x|y}$ still holds where $F_{x|y}$ is the distribution function for $x$ conditioned on $y$.

Figure 1.8 illustrates denoisification on the data from the last section. The red line is the CART classifier learned on the training data (it is the same classifier as in Figure 1.6). We could convert this into a probabilistic classifier by letting $\hat{p}(z = k|x)$ be the proportion of training observations in the same partition as $x$ that belong to class $k$.\(^5\)

For the present example, if observation $x$ has Feature 1 greater than 1, then $\hat{p}(z = \text{class 1}|x) = 1$ and $\hat{p}(z = \text{class 2}|x) = 0$. If observation $x$ has Feature 1 less than 1 then $\hat{p}(z = \text{class 1}|x) = 0$ and $\hat{p}(z = \text{class 2}|x) = 1$. The grey region represents a level set of an estimate of $p(x|y)$. The grey region is offset from the observed point because the density of the training features is high to the left of the feature estimates for the unlabeled observation. Using estimates of $p(z|x)$ and $p(x|y)$ here we would conclude that the unlabeled observation

\(^5\)This method for using CART as a probabilistic classifier is discussed in Breiman et al. [1984] Section 4.6 and implemented in popular versions of CART such as the R package \texttt{rpart} [Therneau et al., 2012].
belongs to class 2. We note several aspects of denoisification that will be discussed in later sections and chapters:

1. A potential advantage to denoisification over noisification is that only one classifier must be constructed. For certain problems this could make denoisification faster to implement.

2. We discuss an application of denoisification to variable star classification in Chapter 4. Denoisification did not perform as well as noisification. In part this was due to difficulty in estimating $p(x|y)$, which is not required for noisification.

3. With denoisification one constructs a single classifier on the training data. Any tuning parameters within the classifier are fixed in the sense that they do not depend on the error distribution of the unlabeled data, $F_e$. It may be the case that different tuning parameters are optimal for different error distributions.

Specifically let $\tilde{p}_H(z|x)$ be a probabilistic classifier that depends on tuning parameters $H$ (and implicitly depends on the training data). The tuning parameters are generally chosen to minimize the risk associated with some loss function $\ell$. In other words the optimal tuning parameters for the training data and loss function $\ell$ are

$$H^* = \arg\min_H \mathbb{E}_{(X,Z)}[\ell(\tilde{p}_H(z|X), (X,Z))]$$

In practice $H^*$ may be estimated through a method such as cross validation. Note that this procedure does not depend on $p(x|y)$. Using $H^*$ (and temporarily assuming that the density $p(x|y)$ is known), Equation 1.2 suggests estimating $p(z|y)$ using

$$\tilde{p}_{H^*}(z|y) = \int \tilde{p}_{H^*}(z|x)p(x|y)dx.$$ 

However it may be the case that

$$H^* \neq \arg\min_H \mathbb{E}_{(Y,Z)}[\ell(\int \tilde{p}_H(z|x)p(x|Y)dx, (Y,Z))]$$

In Chapter 2 we study this issue for density estimation. We present theoretical results for kernel density estimators showing that optimal tuning parameters do depend on the form of the error. Using a fixed set of tuning parameters can result in a suboptimal convergence rate for the density estimate.
1.5 Summary of Chapters

The remainder of this work is divided into three chapters. While the chapters reference each other, they are self-contained and may be read individually. In Chapter 2 we develop asymptotic approximations for choosing the bandwidth matrix for kernel density estimators. We study density estimation when a sample $X_1, \ldots, X_n \sim f_X$ is observed without measurement error but one seeks to estimate the density of $f_X$ convolved with some known error distribution. We discuss some potential applications to quasar target selection. We highlight the phenomenon that the error distribution smooths the density estimate, effectively reducing the amount of smoothing one needs to introduce using the bandwidth matrix.

In Chapter 3 we study measurement error in the regression case using a Nadaraya–Watson type estimator. We develop asymptotic approximations to the mean squared error as a function of the bandwidth. We express the optimal bandwidth as a solution to a quadratic program. We present an example where the optimal bandwidth matrix is 0 and discuss some statistical interpretations.

Finally in Chapter 4 we develop a noisification procedure for classifying periodic variable stars with sparsely observed functions. We illustrate the relationship between feature distributions and the cadence and magnitude error distribution for several real data sets. Noisification improves classifier performance on several simulated and real world data sets. We find differences in training and unlabeled data sets that are not addressed by noisification as well, such as different astronomical surveys drawing from different populations of variable stars. We provide recommendations for constructing variable star classifiers for upcoming surveys.
Chapter 2

Density Estimation with Berkson Error

2.1 Introduction

We consider density estimation for a variable subject to measurement error. Most work on errors–in–variables with density estimation has focused on deconvolution problems in which a contaminated sample is used to estimate an uncontaminated density. However there are situations where the sample is noise free and one is interested in the convolution of the sample density with some error term.

For example, Delaigle [2007] considers estimating NO$_2$ exposure in children using known kitchen and bedroom concentrations. The exposure level in children is modeled as a function of kitchen and bedroom concentrations plus some random error.

In another example, Bovy et al. [2011] considers classification of quasars and stars. Here, each object belongs to the class quasar or star. For each object a telescope has recorded a vector of flux ratios. Using a set of objects of known class, the authors estimate the density of the flux ratios for quasars and for stars. These density estimates are used to construct a likelihood ratio classifier. For objects of unknown class, there is often measurement error in the flux ratios. The authors convolve the flux ratio density estimates with the error density and then compute a likelihood ratio using the error convolved densities. While the ultimate goal of this procedure is accurate classification, an intermediate step is accurate estimation of the error convolved flux ratio densities for stars and quasars.$^1$

We now formalize the density estimation problem. Say we observe independent $\{X_i\}_{i=1}^n \sim f_X$. We use this training data to estimate the density, denoted $f_Y$, of $Y = X + \epsilon$. Here $\epsilon$ is a random variable of known distribution, $X \sim f_X$, and $\epsilon$ and $X$ are independent. All random

$^1$See Section 2 (Equations 1, 2, and 3) and Section 5 of Bovy et al. [2011] for more information.
variables are in $\mathbb{R}^p$. In the literature, $\epsilon$ is known as Berkson error and was introduced in a regression context by Berkson [1950]. It differs from the more familiar “classical” error where one observes a sample contaminated with some error, and one wants to infer properties of the uncontaminated sample. See Carroll et al. [2006] Chapter 1 for a detailed discussion of the differences between classical and Berkson error.

Somewhat counter intuitively, estimation of $f_Y$ is made easier by the fact that the sample is from $f_X$ instead of $f_Y$. In fact it is straightforward to show that if $\epsilon$ has a square–integrable density $f_\epsilon$ and $X$ has a bounded density, the estimator

$$\tilde{f}_Y(y) = \frac{1}{n} \sum_{i=1}^{n} f_\epsilon(y - X_i)$$ (2.1)

is unbiased with a mean integrated squared error (MISE) that converges to 0 at rate $n$ (see Delaigle [2007]). This contrasts with standard kernel density estimation where samples are drawn directly from $Y$ and the MISE is of order $n^{-4/(4+p)}$.

$\tilde{f}_Y$ resembles the standard kernel density estimator for $f_X$, but with the kernel $K_h$ replaced by $f_\epsilon$. This suggests that intuition from standard kernel density estimation theory may be applied to the Berkson error setting. For example, if $n$ is small and $f_\epsilon$ is concentrated close to 0, we may expect $\tilde{f}_Y$ to have high variance as happens with a kernel of small bandwidth (relative to $n$). This suggests that error rates may be improved by considering the use of kernels. In this work we study how to improve estimation of $f_Y$ using kernels.

We find that when the Berkson error satisfies certain smoothness conditions, the optimal bandwidth converges to 0 at rate $\sqrt{n}$ and provides an order $n^{-2}$ reduction in MISE over $\tilde{f}_Y$ in Equation (2.1). We show that it is a poor idea to estimate the density of $X$ using $X_1, \ldots, X_n$ and then convolve this estimate with the density of $\epsilon$ to obtain an estimate of $f_Y$. This procedure results in an asymptotic MISE of order $n^{-4/(4+p)}$.

In the case where $\epsilon$ has a smooth density, one can obtain an order $n^{-1}$ MISE without smoothing. However in certain cases $Y$ may be a noisy version of $X$ along certain dimensions, but noise free along others. For example if $p = 2$, $\epsilon$ could be normal along the first coordinate and 0 with probability 1 along the second coordinate. In this case $\tilde{f}_Y$ in Equation (2.1) is not defined because $\epsilon$ does not have a density. Here kernels are necessary in order to obtain a finite MISE.

We find that in $p$ dimensional problems where $\epsilon$ has a density on a $d_1$ dimensional subspace and is 0 with probability 1 on the remaining $d_2$ dimensions ($d_1 + d_2 = p$), optimal bandwidth sizes over the $d_2$ dimensions uncontaminated by error are order $n^{-1/(4+d_2)}$ and the order of the MISE is $n^{-4/(4+d_2)}$. Bandwidths over the $d_1$ dimensions contaminated with sufficiently

\[\text{The } n^{-4/(4+p)} \text{ order for the MISE requires regularity conditions on } f_Y. \text{ For example on page 95 in Section 4.3, Wand and Jones [1995] assumes each entry of the Hessian of } f_Y \text{ is piecewise continuous and square integrable. See page 100 of Wand and Jones [1995] for the MISE convergence rate.}\]
smooth Berkson error may be set to 0. Here presence of Berkson error lowers the order of the optimal amount of smoothing needed to minimize the MISE. As in the case where \( \epsilon \) has a smooth density, optimizing the bandwidth for estimating \( f_X \) and then convolving this estimate with \( dF_\epsilon \) results in an estimator with a suboptimal convergence rate.

In Section 2.2 we introduce notation and define the problem. In Section 2.3 we present a theorem for the case when the error is sufficiently smooth in all directions. In Section 2.4 we deal with some degenerate cases, such as when \( \epsilon = 0 \) with probability 1 along certain coordinates. In Section 2.5 we summarize our findings and suggest directions for future research. Proofs of all theorems are given in Section 2.6 and some technical issues are addressed in Section 2.7.

### 2.2 Problem Setup

We observe independent random variables \( X_1, \ldots, X_n \sim f_X \). We aim to estimate, \( f_Y \), the density of

\[
Y = X + \epsilon.
\]

Here \( X \sim f_X, \epsilon \) has distribution \( F_\epsilon \), and \( X \) and \( \epsilon \) are independent. All random variables are in \( \mathbb{R}^p \). In all that follows let \( \hat{f}_V \) represent the characteristic function of the random variable \( V \) and let \( \tilde{f} \) represent an estimator of \( f \).

#### 2.2.1 Construction of Estimator for \( f_Y \)

Let \( K \) be a mean 0 density function called the kernel, and \( \hat{K} \) its characteristic function. Let

\[
\Sigma_K = \int x x^T K(x) dx.
\]

Let \( H = H_n \) be a sequence of positive semidefinite \( p \times p \) matrices called the bandwidth.

\[
\tilde{f}_X(\omega) = \frac{1}{n} \sum_{j=1}^{n} e^{i \omega^T X_j}
\]

is an estimate of \( \hat{f}_X \). Consider estimating \( \hat{f}_Y \) using

\[
\tilde{f}_Y(\omega) = \hat{K}(H \omega) \hat{f}_\epsilon(\omega) \tilde{f}_X(\omega). \tag{2.2}
\]

Note that \( \tilde{f}_Y(\omega) \) is a characteristic function because it is the product of characteristic functions. Assuming \( \tilde{f}_Y(\omega) \in L_1 \), we may estimate \( f_Y \) using

\[
\tilde{f}_Y(y) \equiv \frac{1}{(2\pi)^p} \int e^{-i \omega^T y} \tilde{f}_Y(\omega) d\omega. \tag{2.3}
\]
The assumption that \( \tilde{f}_Y(\omega) \in L_1 \) implies \( \tilde{f}_Y \) is a bounded density (see Theorem 3.3 in Durrett [2005]). Throughout this work, we only consider kernels \( K \) and bandwidths \( H \) that guarantee \( \tilde{f}_Y(\omega) \in L_1 \).

### 2.2.2 \( \tilde{f}_Y \) as a Kernel Density Estimator

\( \tilde{f}_Y \) in Equation (2.3) is a generalization of the standard kernel density estimator. To see this, note that \( \tilde{f}_Y(\omega) \) is the product of three characteristic functions. We can thus write \( \tilde{f}_Y \) as a convolution. In the case that \( H \succ 0 \), by defining \( K_H(\cdot) \equiv \frac{1}{\det(H)} K(H^{-1} \cdot) \), we have

\[
\tilde{f}_Y(y) = \int \frac{1}{n} \sum_{i=1}^{n} K_H(y - X_i - \epsilon)dF_\epsilon.
\] (2.4)

Here \( \tilde{f}_Y \) resembles a kernel density estimator convolved with \( dF_\epsilon \). In fact when there is no error, i.e. \( P(\epsilon = 0) = 1 \), then \( f_Y = f_X \) and

\[
\tilde{f}_Y(y) = \frac{1}{n} \sum_{i=1}^{n} K_H(y - X_i).
\] (2.5)

It is important to note that the representations of \( \tilde{f}_Y \) in Equations (2.4) and (2.5) are only valid when \( H \succ 0 \). However \( \tilde{f}_Y \), as defined in Equation (2.3), is a well defined estimator for \( f_Y \) as long as \( \tilde{K}(H\omega)f_\epsilon(\omega)f_X(\omega) \in L_1 \). This includes cases where \( H \not\succ 0 \). For example, in the case where \( H = 0 \) and \( \epsilon \) has a density \( f_\epsilon \in L_1 \),

\[
\tilde{f}_Y(y) = \frac{1}{n} \sum_{i=1}^{n} f_\epsilon(y - X_i).
\]

Thus in the case where \( \epsilon \) has a density, \( \tilde{f}_Y \) in Equation (2.3) is a generalization of the kernel-free estimator studied by Delaigle [2007].

### 2.2.3 MISE and Selection of Bandwidth

We evaluate the estimator \( \tilde{f}_Y \) defined by Equation (2.3) using mean integrated squared error (MISE). Let \( \mathcal{P}_n \) be the product measure on \((X_1, \ldots, X_n)\). Define

\[
\text{MISE}(H) \equiv E_{\mathcal{P}_n} \int \left( \tilde{f}_Y(y) - f_Y(y) \right)^2 dy.
\]

We study how to choose the bandwidth \( H \) to minimize the MISE. Specifically we seek

\[
H_{\text{opt}} = \arg\min_{\{H: H \geq 0, \tilde{f}_Y \in L_1\}} \text{MISE}(H).
\]
Unfortunately the MISE expression is complicated and exact calculation of $H_{opt}$ is not possible. However we can form asymptotic approximations to the MISE and study the rate at which $||H_{opt}||_\infty \to 0$ as $n \to \infty$. This will help in choosing $H$ that are approximately optimal and provide an understanding of the improvement in MISE that a well chosen $H$ obtains over simpler methods. For example, in the case where $\epsilon$ has a density in $L_1$ with respect to Lebesgue measure, we compare $H_{opt}$ to $H = 0$ and the resulting asymptotic MISE. We also compare $H_{opt}$ to determining an optimal bandwidth for estimating $f_X$ and then convolving this estimate with $f_\epsilon$ to obtain an estimate of $f_Y$.

The asymptotic results in this work are divided into two sections based on the structure of $\epsilon$. In Section 2.3 we consider the case where $\epsilon$ has a density that is sufficiently smooth. In Section 2.4 we consider the case where $\epsilon$ has support restricted to a subspace of $\mathbb{R}^p$. For the purposes of forming asymptotic expansions, we represent the MISE in terms of characteristic functions.

**Theorem 2.1.** Assume $\hat{f}_Y \in L_1$ and $\tilde{f}_Y \in L_1$. Then

$$(2\pi)^p \text{MISE}(H) = \int |1 - \hat{K}(H\omega)|^2 d\mu(\omega) + \frac{1}{n} \int |\hat{K}(H\omega)|^2 d\nu(\omega)$$

(2.6)

where

$$d\mu(\omega) = |\hat{f}_\epsilon(\omega)|^2 |\hat{f}_X(\omega)|^2 d\omega,$$

$$d\nu(\omega) = |\hat{f}_\epsilon(\omega)|^2 (1 - |\hat{f}_X(\omega)|^2) d\omega$$

are positive measures.

See Subsection 2.6.1 p.29 for a proof. The representation of the MISE in Equation (2.6) closely resembles that of Tsybakov [2009] Theorem 1.4. In Equation (2.6), $\int |1 - \hat{K}(H\omega)|^2 d\mu(\omega)$ is the integrated squared bias of $\tilde{f}_Y$ and $\frac{1}{n} \int |\hat{K}(H\omega)|^2 d\nu(\omega)$ is the integrated variance of $\tilde{f}_Y$. Notice that for fixed $H$, the variance decreases at rate $n^{-1}$ while the bias is constant.

Throughout this work, we require assumptions on the kernel $K$ and the bandwidth matrix $H$.

**Assumptions A.**

\[ K \text{ is a symmetric density} \] (2.7)

\[ \hat{K} \text{ is four times continuously differentiable} \] (2.8)

\[ H = H_n \succeq 0 \text{ (sequence is positive semidefinite)} \] (2.9)

\[ ||H||_\infty \to 0 \] (2.10)
Since we choose the kernel and bandwidth matrix, these assumptions can always be satisfied in practice. Common kernel choices such as the standard normal and uniform on \([-1, 1]^{p}\) satisfy Assumptions 2.7 and 2.8. For the case where \(\epsilon\) has a smooth density (Section 2.3), these assumptions on the kernel and bandwidth are sufficient. For the case where \(\epsilon\) has support restricted to a subspace (Section 2.4), additional assumptions on the kernel and bandwidth are needed (see Assumptions C).

### 2.3 Error Term has Density

We begin by considering the case where \(\epsilon\) has a density. We require the following assumptions.

**Assumptions B.**

\[
\int ||\omega||^8 \hat{f}_\epsilon(\omega)^2 d\omega < \infty \tag{2.11}
\]

\[
\int |\hat{f}_\epsilon(\omega)| d\omega < \infty \tag{2.12}
\]

Assumptions 2.11 and 2.12 are satisfied as long as the error term has a density that is smooth, such as multivariate normal or Student’s t (see Sutradhar [1986] for the characteristic function of the multivariate Student’s t). 

**Theorem 2.2.** Under Assumptions A and B and with the notation of Theorem 2.1

\[
(2\pi)^p MISE(H) = \frac{1}{n} \int d\nu(\omega) \\
+ \left( \frac{1}{4} \int (\omega^T H^T \Sigma_K H \omega)^2 d\mu(\omega) - \frac{1}{n} \int (\omega^T H^T \Sigma_K H \omega) d\nu(\omega) \right) \left(1 + O(||H||_2^2)\right). \tag{2.13}
\]

See Subsection 2.6.2 p.30 for a proof. The term

\[
\frac{1}{n} \int d\nu(\omega) = \frac{1}{n} \left( \int |\hat{f}_\epsilon(\omega)|^2 d\omega - \int |\hat{f}_\epsilon(\omega)|^2 |\hat{f}_X(\omega)|^2 d\omega \right)
\]

is the MISE when \(H = 0\). Note that this term is always positive. It is the variance of the unbiased estimator

\[
\frac{1}{n} \sum_{i=1}^{n} f_\epsilon(y - X_i).
\]

The \(\frac{1}{4} \int (\omega^T H^T \Sigma_K H \omega)^2 d\mu(\omega)\) term is the bias caused by using a kernel with bandwidth \(H\) while \(-n^{-1} \int (\omega^T H^T \Sigma_K H \omega) d\nu(\omega)\) is the corresponding reduction in variance.
CHAPTER 2. DENSITY ESTIMATION WITH BERKSON ERROR

While the full bandwidth matrix offers the most flexibility and greatest potential for reduction in MISE, this expression is difficult to optimize (see Subsection 2.7.1). We specialize to two cases: 1) diagonal bandwidth matrix with \( \Sigma_K = \text{Id}_p \) (\( \text{Id}_p \) is the \( p \times p \) identity matrix) and 2) scalar bandwidth.

2.3.1 Diagonal Bandwidth and \( \Sigma_K = \text{Id}_p \)

By restricting our kernel to have \( \Sigma_K = \text{Id}_p \) and the bandwidth matrix to be diagonal we achieve considerable simplification of the MISE. Let \( h_i = H_{ii} \) and \( h_S = (h_{11}, \ldots, h_{pp}) \). The MISE becomes

\[
(2\pi)^p \text{MISE}(h_S) = \frac{1}{n} \int d\nu(\omega) + \left( h_S^T B h_S - \frac{1}{n} h_S^T V \right) \left( 1 + ||h_S||_\infty \right),
\]

where

\[
B_{i,j} = \frac{1}{4} \int \omega_i^2 \omega_j^2 d\mu(\omega), \quad V_i = \int \omega_i^2 d\nu(\omega).
\]

We seek the \( h_S \) which minimizes the larger order terms in the MISE expression. In other words we seek

\[
h_S^* = \arg\min_{h_S \geq 0} \left( h_S^T B h_S - \frac{1}{n} h_S^T V \right).
\] (2.14)

\( B \) is positive definite so the expression is strictly convex and there is a unique solution. Enforcing the domain restriction \( h_S \geq 0 \) (elementwise) is necessary. Even in simple cases the unconstrained optimum \( \frac{1}{2n} B^{-1} V \) may have elements less than 0. We present an example in Subsection 2.7.2 where \( \epsilon, X, \) and \( K \) are all bivariate, independent normals and the unconstrained optimum in Equation (2.14) has negative elements. In this example, both \( f_Y \) and \( \tilde{f}_Y \) are product densities, yet the optimal bandwidth for direction \( j \), i.e. \( h_{S,j}^* \), depends on the distribution of variables along coordinates other than \( j \).

2.3.2 Scalar Bandwidth

A second possibility is to use scalar bandwidths. We reparameterize the bandwidth \( H = h \text{Id}_p \). Here the general MISE expression in Equation (2.13) becomes

\[
(2\pi)^p \text{MISE}(h) \quad = \quad \frac{1}{n} \int d\nu(\omega) + \left( \frac{h^4}{4} \int (\omega^T \Sigma_K \omega)^2 d\mu(\omega) - \frac{h^2}{n} \int (\omega^T \Sigma_K \omega) d\nu(\omega) \right) \left( 1 + O(h^2) \right).
\] (2.15)
It is simple to optimize the main terms in this MISE expression. Specifically,

\[ h^* = \arg\min_{h \geq 0} \left( \frac{h^4}{4} \int (\omega^T \Sigma K \omega)^2 d\mu(\omega) - \frac{h^2}{n} \int (\omega^T \Sigma K \omega) d\nu(\omega) \right) \]

\[ = \sqrt{\frac{2}{n} \int (\omega^T \Sigma K \omega) d\nu(\omega)} \frac{1}{\sqrt{n} \int (\omega^T \Sigma K \omega)^2 d\mu(\omega)}. \]

\( h^* \) converges to 0 at rate \( n^{1/2} \). Note that this rate does not depend on the dimensionality of the problem, i.e. the rate is not a function of \( p \). Using, \( h^* \) the MISE is

\[ (2\pi)^p \text{MISE}(h^*) = \frac{1}{n} \int d\nu(\omega) - \frac{1}{n^2} \left( \frac{\int (\omega^T \Sigma K \omega) d\nu(\omega)}{\int (\omega^T \Sigma K \omega)^2 d\mu(\omega)} \right)^2 + O(n^{-3}). \]

We now compare \( h^* \) and \( \text{MISE}(h^*) \) with two other approaches to choosing the bandwidth.

**Approach 1:** Choose \( h \) to minimize MISE in estimating \( f_X \). This approach is motivated by the fact that \( f_Y(y) = \int f_X(y - \epsilon) dF_\epsilon \). Thus one might use a kernel density estimator to construct \( \tilde{f}_X \) and then convolve \( \tilde{f}_X \) with \( dF_\epsilon \) in order to estimate \( f_Y \). In this case, under certain regularity conditions on \( f_X \), the bandwidth is order \( n^{-1/(4+p)} \) (e.g. see Wand and Jones [1995] page 100). Specifically, say

\[ h' = D(n)n^{-1/(4+p)}, \]

where \( D : Z^+ \to \mathbb{R}^+ \) such that \( \limsup_n D(n) < \infty \) and \( \liminf_n D(n) > 0 \). The MISE for estimating \( f_Y \) using \( h' \) (obtained from Equation (2.15)) is

\[ (2\pi)^p \text{MISE}(h') = \frac{1}{n} \int d\nu(\omega) + \left( \frac{D(n)^4 n^{-4/(4+p)}}{4} \int (\omega^T \Sigma K \omega)^2 d\mu(\omega) \right) \]

\[ - D(n)^2 n^{-(6+p)/(4+p)} \int (\omega^T \Sigma K \omega) d\nu(\omega) \left( 1 + O(n^{2/(4+p)}) \right) \]

\[ = \left( \frac{D(n)^4 n^{-4/(4+p)}}{4} \int (\omega^T \Sigma K \omega)^2 d\mu(\omega) \right) (1 + o(1)). \]

The \( n^{-4/(4+p)} \) order for the MISE when using \( h' \) is strictly worse than the \( n^{-1} \) order that can be achieved by optimizing the bandwidth specifically for the error distribution, i.e. using \( h^* \). Essentially using \( h' \) oversmooths \( \tilde{f}_Y \). The first order term in MISE(\( h' \)) is caused entirely by bias.

**Approach 2:** Choose \( h = 0 \). Here we have

\[ (2\pi)^p \text{MISE}(0) = \frac{1}{n} \int d\nu(\omega) = \frac{1}{n} \left( \int |\hat{f}_\epsilon(\omega)|^2 d\omega - \int |\hat{f}_\epsilon(\omega)|^2 |\tilde{f}_X(\omega)|^2 d\omega \right). \]
Asymptotically, this approach is better than Approach 1 since $\text{MISE}(0)$ is order $n^{-1}$. The ratio of optimal smoothing to no smoothing is

$$\frac{\text{MISE}(h^*)}{\text{MISE}(0)} = 1 - \frac{1}{n} \frac{\left( \int \omega^T \Sigma_K \omega d\nu(\omega) \right)^2}{\left( \int (\omega^T \Sigma_K \omega)^2 d\mu(\omega) \right) \left( \int d\nu(\omega) \right)} + O(n^{-2}).$$

### 2.4 Error Term has Support Restricted to Subspace

Here we consider cases where $\epsilon$ has no error in certain directions. This occurs when $Y$ is measured perfectly for some features but with errors for others. Unlike the previous case, here one must smooth along directions where $\epsilon$ is 0 with probability 1, otherwise the MISE is infinite. Roughly, the proofs assume i) the marginal distribution of $\epsilon$ on the first $d_1$ coordinates has a density that is sufficiently smooth and ii) $P(\epsilon_{(d_1+1,\ldots,p)} = 0) = 1$. It is obvious that the coordinates along which $\epsilon$ has no error are arbitrary. A simple argument in Subsection 2.4.3 allows one to construct density estimates when $\epsilon$ has support restricted to any subspace.

We begin with notation and assumptions specific to this section. Let $A = \{1, \ldots, d_1\}$ for some $d_1 \geq 1$. Let $d_1 + d_2 = p$. We define $\omega = (\omega_A, \omega_{AC})$ where $\omega_A = (\omega_1, \ldots, \omega_{d_1})$ and $\omega_{AC} = (\omega_{d_1+1}, \ldots, \omega_p)$. Further we define the bandwidth matrix in terms of blocks. Let

$$H = \begin{bmatrix} H_{11} & H_{12} \\ H_{12}^T & H_{22} \end{bmatrix},$$

where $H_{11} \in \mathbb{R}^{d_1 \times d_1}$, $H_{12} \in \mathbb{R}^{d_1 \times d_2}$, $H_{22} \in \mathbb{R}^{d_2 \times d_2}$. Note that since $H = H_n$ is a sequence, we may refer to sequences of these block elements, e.g. $H_{n,22}$. 
CHAPTER 2. DENSITY ESTIMATION WITH BERKSON ERROR

Assumptions C.

\[ \hat{f}_\epsilon(\omega') = \hat{f}_\epsilon(\omega) \text{ whenever } \omega'_A = \omega_A \]  
(2.16)

\[ \int |\hat{f}_\epsilon(\omega_A, 0)| d\omega_A < \infty \]  
(2.17)

\[ \int ||\omega||^8 d\mu(\omega) < \infty \]  
(2.18)

\[ \int |\hat{f}_X(\omega)||\hat{f}_\epsilon(\omega)| d\omega < \infty \]  
(2.19)

\[ \hat{K}(\omega) \leq \frac{C_0}{(1 + ||\omega||)^{d_2+\Delta}} \text{ for some } C_0, \Delta > 0 \]  
(2.20)

\[ H_{22} = H_{n,22} \text{ is a sequence of positive definite matrices} \]  
(2.21)

\[ \frac{1}{n \det(H_{22})} \to 0 \]  
(2.22)

\[ ||H_{12}H_{22}^{-1}||_\infty \text{ is uniformly bounded above} \]  
(2.23)

Assumption 2.16 on the characteristic function of \( \epsilon \) requires that the error is 0 with probability 1 on the \( A^C = \{d_1 + 1, \ldots, p\} \) coordinates. To interpret Assumption 2.17, note that

\[ \hat{f}_\epsilon(\omega_A, 0) = \int e^{i(\omega_A, 0)T \epsilon} dF_\epsilon = \int e^{i\omega_A^T \epsilon_A} dF_{\epsilon_A}. \]

So \( \hat{f}_\epsilon(\omega_A, 0) \) in Assumption 2.17 is the characteristic function of \( \epsilon_A = \epsilon_{1,\ldots,d_1} \). The assumption is satisfied if \( \epsilon_{1,\ldots,d_1} \) has a standard normal or Student’s t density. Assumptions 2.18 and 2.19 are satisfied if \( \int ||\omega||^8 |\hat{f}_X(\omega)|^2 d\omega < \infty \) and \( \int |\hat{f}_X(\omega)| d\omega < \infty \). We can interpret these assumptions as requiring some smoothness on \( f_X \).

In Assumption 2.20 we require the kernel to have a characteristic function that decays as \( 1/||\omega||^{d_2+\Delta} \) in the tails for some \( \Delta > 0 \) (recall \( d_2 \) is the number of dimensions on which there is no error). This is satisfied for the multivariate standard normal kernel. Note that it is not satisfied for the uniform kernel. We also require \( H_{22} \) to be positive definite (Assumption 2.21) and converge to 0 not too fast (Assumption 2.22). Note that \( H_{22} \) is the block of the bandwidth matrix on the coordinates where there is no error. Assumptions 2.21 and 2.22 are identical to assumptions used for the entire bandwidth matrix in certain works on multivariate density estimation without error (e.g. see the definition of bandwidth in Equation 4.1 on page 91 and Assumption (ii) on page 95 of Wand and Jones [1995]).
Theorem 2.3. Under Assumptions A and C and with the notation of Theorem 2.1

\[(2\pi)^p MISE(H) \]
\[= \left( \frac{1}{4} \int (\omega^T H^T \Sigma K H \omega)^2 d\mu(\omega) \right) + \frac{1}{n \det(H_{22})} \int |\hat{f}_\epsilon(\omega_A, 0)|^2 d\omega_A \int |\hat{K}(H_1 H_2^{-1} \omega_{AC}, \omega_{AC})|^2 d\omega_{AC} (1 + o(1)). \]

See Subsection 2.6.3 p.33 for a proof. This MISE expression has the same squared bias as the case where \(\epsilon\) has a density (see Equation 2.13). The variance contains \(\det(H_{22})\) in the denominator, unlike in the error free kernel density estimation case where there is \(\det(H)\) in the denominator of the variance term. This reflects the fact that along the \(d_2\) dimensions where there is no error, shrinking the bandwidth to 0 causes the variance to become arbitrarily large. However along the \(A = \{1, \ldots, d_1\}\) coordinates where there is error, no smoothing is required. Specifically if \(H_{11} = 0\) and \(H_{12} = 0\), the MISE is still finite.

2.4.1 Normal Kernel

When the kernel is standard normal (\(\Sigma_K = I_{d_p}\)), we can simplify the MISE expression in Equation (2.24).

Corollary 2.1. Assume \(K\) is standard normal. Let \(S = H^T H\). Define block notation for \(S\),

\[S = \begin{bmatrix} S_{11} & S_{12} \\ S_{12}^T & S_{22} \end{bmatrix}.\]

Under these conditions Equation (2.24) becomes

\[(2\pi)^p MISE(H) = \left( \frac{1}{4} \int (\omega^T S \omega)^2 d\mu(\omega) \right) + \frac{\pi^{d_2/2}}{n \det(S_{22})^{1/2}} \int |\hat{f}_\epsilon(\omega_A, 0)|^2 d\omega_A (1 + o(1)). \] (2.25)

See Subsection 2.6.4 for a proof. The MISE in Equation (2.25) is potentially easier to optimize than the MISE in Equation (2.24) because the smoothing parameters are completely decoupled from the kernel.

2.4.2 Optimization Using a Scalar Bandwidth

Perhaps the simplest optimization strategy is to use a scalar bandwidth over directions where there is no error and a bandwidth of 0 over directions where there is error. Specifically one could set \(H_{11} = 0\), \(H_{12} = 0\), and \(H_{22} = h I_{d_2}\). Define \(\Sigma_{K,22} = \int x_{AC} x_{AC}^T K(x_{AC}) dx_{AC}\) (in
other words, $\Sigma_{K,22}$ is the lower right $d_2 \times d_2$ block of $\Sigma_K$). In this case the main terms in the MISE of Equation (2.24) reduce to
\[
\frac{h^4}{4} \int (\omega_{AC}^T \Sigma_{K,22} \omega_{AC})^2 d\mu(\omega) + \frac{1}{nh^{d_2}} \int |\hat{f}_c(\omega_A, 0)|^2 d\omega_A \int |\hat{K}(0, \omega_{AC})|^2 d\omega_{AC}.
\] (2.26)

The minimizer of this expression is
\[
h^* = \left( \frac{d_2 \int |\hat{f}_c(\omega_A, 0)|^2 d\omega_A \int |\hat{K}(0, \omega_{AC})|^2 d\omega_{AC}}{n \int (\omega_{AC}^T \Sigma_{K,22} \omega_{AC})^2 d\mu(\omega)} \right)^{1/(d_2+4)}.
\] (2.27)

Note that the optimal smoothing on the $A^C = \{d_1 + 1, \ldots, p\}$ coordinates depends on the error $\hat{f}_c(\omega_A, 0)$, which only exists on the $A = \{1, \ldots, d_1\}$ coordinates. The numerator in this expression is known. One could estimate the integral in the denominator using plug-in methods.

As was the case in Section 2.3, this result shows that it is suboptimal to choose $h$ to optimize estimation of $f_X$ and then convolve this estimate with $dF$. For estimation of $f_X$ the optimal bandwidth order for $h$ is $n^{-1/(4+p)}$ (see e.g. Section 4.2 page 100 of Wand and Jones [1995]). Equation (2.26) shows that this will create squared bias of order $n^{-4/(4+p)}$ and variance of order $n^{-(p+4-d_2)/(p+4)}$. The resulting MISE is order $n^{-4/(4+p)}$, strictly worse than the $n^{-4/(4+d_2)}$ order one can achieve by using Equation (2.26) to directly optimize $h$ for estimation of $f_Y$. Effectively, optimizing $h$ for $f_X$ results in oversmoothing of the $f_Y$ density estimate because the error term $\epsilon$ acts as a smoother itself.

Finally note that in Subsection 2.3.2 we considered the case where $H = 0$. This is not possible here because with $H = 0$, the MISE is not defined.

### 2.4.3 Error Density Generalization

If there is measurement error on a subspace not aligned with the first $d_1$ axis, one can transform the data so that Theorem 2.3 applies. For some matrix $P$ with $\det(P) = 1$ define $\epsilon' = Pe$, $Y' = PY$ and $X'_i = PX_i$. Say Assumptions C hold for the transformed data. We can view $P$ as a bijective map between density estimates for $Y'$ and $Y$. Specifically, one can use Equation (2.27) to determine the bandwidth for a density estimate $\tilde{f}_{Y'}$ of $f_{Y'}$. Then one can estimate $f_Y$ using $\tilde{f}_{Y'}(\cdot) = \tilde{f}_{Y'}(P \cdot)$. $\tilde{f}_{Y'}$ and $\tilde{f}_{Y'}$ have the same MISE:
\[
\int_y (\tilde{f}_{Y'}(y) - f_{Y'}(y))^2 dy = \int_u (\tilde{f}_{Y'}(Pu) - f_{Y'}(Pu))^2 du = \int_u (\tilde{f}_{Y'}(u) - f_Y(u))^2 du.
\]
2.5 Conclusions

We constructed asymptotic approximations to the MISE for kernel density estimators when there is measurement error in features. Estimators for the \(f_Y\) density converge faster than estimators for the error free density \(f_X\). Optimal bandwidth matrices \(H\) converge to 0 faster as well.

As we highlighted in Sections 2.3 and 2.4, one consequence of this finding is the suboptimality of constructing a density estimator for \(f_X\) and then convolving this estimate with \(dF_\epsilon\). This generally leads to oversmoothing and produces estimators with suboptimal convergence rates.

The current results are rather dichotomous in the sense that bandwidth sizes are only dependent on the presence or absence of Berkson error (by direction), not the covariance of Berkson error, denoted \(\Sigma_\epsilon\). It is possible that more refined statements could be made by considering settings where \(n \to \infty\) and \(\Sigma_\epsilon \to 0\) together at some rate. An approach similar to this was taken by Delaigle [2008] in the context of classical measurement error and the deconvolution problem.

An additional limitation of the present study is that bandwidths depend on the structure of \(\epsilon\). In certain cases we may need density estimates for observations \(Y_1, \ldots, Y_m\), each of which has its own error density. As bandwidths are often chosen by computationally intensive methods such as cross validation, it may be difficult to determine optimal bandwidth for each \(Y_j\).

One natural extension to this work is to study the case where the training data \(X_1, \ldots, X_n\) is observed with error. A second extension is to study the asymptotics of smoothing parameters in kernel methods for regression data with measurement error. This is the subject of Chapter 3.

2.6 Proofs and Lemmas

2.6.1 Proof of Theorem 2.1

We must show

\[
(2\pi)^p \text{MISE}(H) = \int |1 - \hat{K}(H\omega)|^2 d\mu(\omega) + \frac{1}{n} \int |\hat{K}(H\omega)|^2 d\nu(\omega)
\]

where

\[
d\mu(\omega) = |\hat{f}_\epsilon(\omega)|^2 |\hat{f}_X(\omega)|^2 d\omega,
\]

\[
d\nu(\omega) = |\hat{f}_\epsilon(\omega)|^2 (1 - |\hat{f}_X(\omega)|^2) d\omega.
\]
Substituting for \(d\mu(\omega)\) and \(d\nu(\omega)\), it suffices to show that
\[
(2\pi)^p \text{MISE}(H) = \int |\hat{f}_\epsilon(\omega)|^2 \left( |1 - \hat{K}(H\omega)|^2 |\hat{f}_X(\omega)|^2 + \frac{1}{n} |\hat{K}(H\omega)|^2 (1 - |\hat{f}_X(\omega)|^2) \right) d\omega.
\]
(2.28)

\(\hat{f}_Y, \hat{f}_\epsilon \in L_1\) by assumption. They are in \(L_2\) because they are characteristic functions and thus bounded. Under these conditions, the Plancherel theorem (see Theorem 1.8.8 on page 57 in Ushakov [1999]) states
\[
\int (f_Y(y) - \tilde{f}_Y(y))^2 dy = \frac{1}{(2\pi)^p} \int |\hat{f}_Y(\omega) - \hat{\tilde{f}}_Y(\omega)|^2 d\omega.
\]
(2.29)

Let \(P_n\) be the product measure on \((X_1, \ldots, X_n)\). Using the definition of MISE\((H)\), Equation (2.29), and the facts \(\hat{f}_Y(\omega) = \hat{f}_X(\omega)\hat{f}_\epsilon(\omega)\) and \(\hat{\tilde{f}}_Y(\omega) = \hat{K}(H\omega)\hat{f}_\epsilon(\omega)\hat{f}_X(\omega)\), we have
\[
\text{MISE}(H) = E_{P_n} \int (f_Y(y) - \tilde{f}_Y(y))^2 dy
\]
\[
= \frac{1}{(2\pi)^p} E_{P_n} \int |\hat{f}_Y(\omega) - \hat{\tilde{f}}_Y(\omega)|^2 d\omega
\]
\[
= \frac{1}{(2\pi)^p} E_{P_n} \int |\hat{K}(H\omega)\hat{f}_\epsilon(\omega)\hat{f}_X(\omega) - \hat{f}_X(\omega)\hat{f}_\epsilon(\omega)|^2 d\omega
\]
\[
= \frac{1}{(2\pi)^p} E_{P_n} \int |\hat{f}_\epsilon(\omega)|^2 |\hat{f}_X(\omega)\hat{K}(H\omega) - \hat{f}_X(\omega)|^2 d\omega.
\]

Note that the integrand is a non-negative function, so we move the expectation inside the integral using Fubini’s Theorem. We have
\[
(2\pi)^p \text{MISE}(H) = \int |\hat{f}_\epsilon(\omega)|^2 E_{P_n} |\hat{f}_X(\omega)\hat{K}(H\omega) - \hat{f}_X(\omega)|^2 d\omega.
\]

Noting that it is sufficient to show Equation (2.28) holds, all that is left is to show is
\[
E_{P_n} |\hat{f}_X(\omega)\hat{K}(H\omega) - \hat{f}_X(\omega)|^2 = |1 - \hat{K}(H\omega)|^2 |\hat{f}_X(\omega)|^2 + \frac{1}{n} |\hat{K}(H\omega)|^2 (1 - |\hat{f}_X(\omega)|^2).
\]

This identity is shown in the proof of Theorem 1.4 on page 22 in Tsybakov [2009].

\[\square\]

### 2.6.2 Proof of Theorem 2.2

Recall that we are working under Assumptions A and B. This proof is divided into three parts. In **Part 1** we show \(\hat{f}_Y, \hat{f}_\epsilon \in L_1\), which satisfies the conditions for Theorem 2.1 and implies
\[
(2\pi)^p \text{MISE}(H) = \int |1 - \hat{K}(H\omega)|^2 d\mu(\omega) + \frac{1}{n} \int |\hat{K}(H\omega)|^2 d\nu(\omega).
\]
(2.30)
In **Part 2** we expand the first term of the right hand side of Equation (2.30) to show

\[
\int |1 - \hat{K}(H\omega)|^2 d\mu(\omega) = \left( \frac{1}{4} \int (\omega^T \Sigma_K H\omega)^2 d\mu(\omega) \right) (1 + O(||H||^2_\infty)). \tag{2.31}
\]

In **Part 3** we expand the second term of the right hand side of Equation (2.30) to show

\[
\frac{1}{n} \int |\hat{K}(H\omega)|^2 d\nu(\omega) = \frac{1}{n} \int d\nu(\omega) - \left( \frac{1}{n} \int (\omega^T \Sigma_K H\omega) d\nu(\omega) \right) (1 + O(||H||^2_\infty)). \tag{2.32}
\]

Summing Equations (2.31) and (2.32) we have the result

\[
(2\pi)^p \text{MISE}(H) = \frac{1}{n} \int d\nu(\omega) + \left( \frac{1}{4} \int (\omega^T \Sigma_K H\omega)^2 d\mu(\omega) \right) - \frac{1}{n} \int (\omega^T \Sigma_K H\omega) d\nu(\omega) \tag{2.35}
\]

\[
(1 + O(||H||^2_\infty)). \tag{2.36}
\]

**Part 1:** \(\hat{f}_Y, \widetilde{\hat{f}}_Y \in L_1\)

Note that since the modulus of a characteristic function is bounded by 1

\[
|\hat{f}_Y(\omega)| = |\hat{f}_X(\omega)\hat{f}_\epsilon(\omega)| \leq |\hat{f}_\epsilon(\omega)|,
\]

\[
|\widetilde{\hat{f}}_Y(\omega)| = |\hat{K}(H\omega)\hat{f}_\epsilon(\omega)\widetilde{\hat{f}}_X(\omega)| \leq |\hat{f}_\epsilon(\omega)|.
\]

\(\hat{f}_\epsilon \in L_1\) by Assumption (2.12), implying \(\hat{f}_Y, \widetilde{\hat{f}}_Y \in L_1\).

**Part 2: Bias**

By Lemma 2.1 on p.36 there exists \(R\) satisfying

\[
|R(\omega)| \leq C||\omega||^4_\infty \tag{2.33}
\]

such that

\[
\hat{K}(\omega) = 1 - \frac{\omega^T \Sigma_K \omega}{2} + R(\omega). \tag{2.34}
\]

Note that the kernel \(K\) is symmetric so \(\hat{K}\) and \(R\) are real valued functions.

\[
\int |1 - \hat{K}(H\omega)|^2 d\mu(\omega) = \int \left| \frac{\omega^T \Sigma_K H\omega}{2} - R(H\omega) \right|^2 d\mu(\omega)
\]

\[
= \frac{1}{4} \int (\omega^T \Sigma_K H\omega)^2 d\mu(\omega)
- \int R(H\omega)(\omega^T \Sigma_K H\omega) d\mu(\omega)
+ \int R(H\omega)^2 d\mu(\omega). \tag{2.35}
\]
CHAPTER 2. DENSITY ESTIMATION WITH BERKSON ERROR

We have split the integrals formally. We now show that Expressions (2.35) and (2.36) are $O(||H||^6_\infty)$ by bounding their integrands. Using the bound $R(\omega) \leq C||\omega||^4_\infty$ (Equation (2.33)), for some $E$ we have

$$|R(H\omega)(\omega^TH^T\Sigma_KH\omega)| \leq C||H\omega||^4_\infty||\omega^TH^T\Sigma_KH\omega||_\infty \leq E||H||^6_\infty||\omega||^6_\infty,$$

$$|R(H\omega)^2| \leq C^2||H\omega||^8_\infty \leq E||H||^8_\infty||\omega||^8_\infty.$$

Using the definition of $d\mu(\omega)$ and the fact $\int ||\omega||^8_\infty|\hat{f}_\epsilon(\omega)|d\omega < \infty$ (Assumption (2.11)) we have

$$\int ||\omega||^8_\infty d\mu(\omega) = \int ||\omega||^8_\infty|\hat{f}_X(\omega)|^2|\hat{f}_\epsilon(\omega)|^2d\omega \leq \int ||\omega||^8_\infty|\hat{f}_\epsilon(\omega)|^2d\omega < \infty.$$

So Expressions (2.35) and (2.36) are $O(||H||^6)$ and $O(||H||^8)$ respectively. Thus

$$\int |1 - \hat{K}(H\omega)|^2d\mu(\omega) = \left(\frac{1}{4}\int (\omega^TH^T\Sigma_KH\omega)^2d\mu(\omega)\right) (1 + O(||H||^2_\infty)).$$

**Part 3: Variance** Using the expansion of $\hat{K}$ in Equation (2.34) we have

$$\frac{1}{n} \int |\hat{K}(H\omega)|^2d\nu(\omega) = \frac{1}{n} \int \left|1 - \frac{\omega^TH^T\Sigma_KH\omega}{2} + R(H\omega)\right|^2 d\nu(\omega).$$

Expanding the right hand side we have

$$\frac{1}{n} \int \left|1 - \frac{\omega^TH^T\Sigma_KH\omega}{2} + R(H\omega)\right|^2 d\nu(\omega) = \frac{1}{n} \left(\int d\nu(\omega) - \int (\omega^TH^T\Sigma_KH\omega)d\nu(\omega) + \frac{1}{4} \int (\omega^TH^T\Sigma_KH\omega)^2d\nu(\omega) - \int R(H\omega)(\omega^TH^T\Sigma_KH\omega)d\nu(\omega) + 2 \int R(H\omega)d\nu(\omega) + \int R^2(H\omega)d\nu(\omega)\right).$$

We have split the integral formally. Using the bound $R(\omega) \leq C||\omega||^4_\infty$ (Equation (2.33)) we bound the integrands of Expressions (2.39), (2.40), (2.41), and (2.42). For some $F$ we have

$$|\omega^TH^T\Sigma_KH\omega|^2 \leq F||\omega||^4_\infty||H||^4_\infty,$$

$$|R(H\omega)(\omega^TH^T\Sigma_KH\omega)| \leq F||\omega||^6_\infty||H||^6_\infty,$$

$$|R(H\omega)| \leq F||\omega||^8_\infty||H||^8_\infty,$$

$$|R^2(H\omega)| \leq F||\omega||^8_\infty||H||^8_\infty.$$
CHAPTER 2. DENSITY ESTIMATION WITH BERKSON ERROR

Note that by the definition of $d\nu(\omega)$ and the fact $\int ||\omega||^8 ||f_\epsilon(\omega)||^2d\omega < \infty$ (Assumption (2.11)) we have

$$\int ||\omega||^8 d\nu(\omega) = \int ||\omega||^8 |\hat{f}_\epsilon(\omega)|^2d\omega - \int ||\omega||^8 |\hat{f}_\epsilon(\omega)||\hat{f}_X(\omega)|^2d\omega < \infty.$$  

So Expressions (2.39), (2.40), (2.41), and (2.42) are all integrable and $O(||H||^4)$. Thus

$$\frac{1}{n}\int |\hat{K}(H\omega)|^2 d\nu(\omega) = \frac{1}{n}\int d\nu(\omega) - \left(\frac{1}{n}\int (\omega^T H^T \Sigma K H\omega) d\nu(\omega)\right) (1 + O(||H||^2)).$$

2.6.3 Proof of Theorem 2.3

Recall that we are working under Assumptions A and C. This proof is divided into three parts. In Part 1 we prove $\tilde{f}_Y, \hat{f}_Y \in L_1$. So the conditions of Theorem 2.1 are satisfied and we have

$$(2\pi)^p \text{MISE}(H) = \int |1 - \hat{K}(H\omega)|^2 d\mu(\omega) + \frac{1}{n}\int |\hat{K}(H\omega)|^2 d\nu(\omega).$$  

In Part 2 we expand the first term on the right hand side of Equation (2.43) and conclude

$$\int |1 - \hat{K}(H\omega)|^2 d\mu(\omega) = \left(\frac{1}{4}\int (\omega^T H^T \Sigma K H\omega)^2 d\mu(\omega)\right) (1 + O(||H||^2)).$$

In Part 3 we expand the second term of the right hand side of Equation (2.43) and conclude

$$\frac{1}{n}\int |\hat{K}(H\omega)|^2 d\nu(\omega)$$

$$= \frac{1}{n \det(H_{22})} \int |\hat{f}_\epsilon(\omega_A, 0)|^2 d\omega_A \int |\hat{K}(H_{12} H_{22}^{-1} \omega_{AC}, \omega_{AC})|^2 d\omega_{AC} (1 + o(1)).$$

Summing Equations (2.44) and (2.45) we have the result

$$(2\pi)^p \text{MISE}(H)$$

$$= \left(\frac{1}{4}\int (\omega^T H^T \Sigma K H\omega)^2 d\mu(\omega)\right) + \frac{1}{n \det(H_{22})} \int |\hat{f}_\epsilon(\omega_A, 0)|^2 d\omega_A \int |\hat{K}(H_{12} H_{22}^{-1} \omega_{AC}, \omega_{AC})|^2 d\omega_{AC} (1 + o(1)).$$

Part 1: $\hat{f}_Y, \tilde{f}_Y \in L_1$

Note $\hat{f}_Y(\omega) = \tilde{f}_\epsilon(\omega)\hat{f}_X(\omega) \in L_1$ by Assumption 2.19. Now we show $\tilde{f}_Y \in L_1$. Note

$$|\tilde{f}_Y(\omega)| = |\hat{f}_X(\omega)\hat{K}(H\omega)\hat{f}_\epsilon(\omega)| \leq |\hat{K}(H\omega)\hat{f}_\epsilon(\omega)|.$$
CHAPTER 2. DENSITY ESTIMATION WITH BERKSON ERROR

By Lemma 2.3 (with \( j = 1 \)) on p.41 \( \hat{K}(\omega) \hat{f}_\epsilon(\omega) \in L_1 \).

**Part 2: Bias**

Using Assumption 2.18, \( \int ||\omega||^8_\infty d\mu(\omega) < \infty \), for showing existence of integrals, the bias derivation holds the same as in the proof of Theorem 2.2 p.22. So

\[
\int |1 - \hat{K}(\omega)|^2 d\mu(\omega) = \left( \frac{1}{4} \int (\omega^T H^T \Sigma K H \omega)^2 d\mu(\omega) \right) (1 + O(||H||^2_\infty)).
\]

**Part 3: Variance**

Recall

\[
H = \begin{bmatrix} H_{11} & H_{12} \\ H_{12}^T & H_{22} \end{bmatrix},
\]

where \( H_{11} \in \mathbb{R}^{d_1 \times d_1}, H_{12} \in \mathbb{R}^{d_1 \times d_2}, H_{22} \in \mathbb{R}^{d_2 \times d_2} \). Define

\[
J = \int |\hat{f}_\epsilon(\omega_A,0)|^2 d\omega_A \int |\hat{K}(H_{12} H_{22}^{-1} \omega_A, \omega_A)\omega_A)|^2 d\omega_A.
\]

We seek to show

\[
\int |\hat{K}(\omega)|^2 d\nu(\omega) = \frac{1}{\det(H_{22})} J(1 + o(1)). \tag{2.46}
\]

We begin by establishing some preliminary equalities. By Assumption (2.18) \( \int ||\omega||^8_\infty d\mu(\omega) = \int ||\omega||^8_\infty |\hat{f}_\epsilon(\omega)|^2 |\hat{f}_X(\omega)|^2 d\omega < \infty \). Hence \( \int |\hat{f}_\epsilon(\omega)|^2 |\hat{f}_X(\omega)|^2 d\omega < \infty \). This implies

\[
\int |\hat{f}_\epsilon(\omega)|^2 |\hat{f}_X(\omega)|^2 |\hat{K}(\omega)|^2 d\omega = O(1). \tag{2.47}
\]

Let

\[
H^* = \begin{bmatrix} I_{d_1} & 0 \\ 0 & H_{22} \end{bmatrix}.
\]

Define

\[
I = \int |\hat{f}_\epsilon(\omega_A,0)|^2 |\hat{K}(H H^* - 1 \omega)|^2 d\omega.
\]

By Lemma 2.3 (with \( j = 2 \)) p.41 and Lemma 2.2 p.37, we have

\[
\int |\hat{f}_\epsilon(\omega)|^2 |\hat{K}(\omega)|^2 d\omega = \frac{1}{\det(H_{22})} I, \tag{2.48}
\]

\[
I = J(1 + o(1)). \tag{2.49}
\]
We now show Equation (2.46) holds. By the definition of \(d\nu(\omega)\) and Equations (2.47), (2.48), and (2.49), we have

\[
\int |\hat{K}(H\omega)|^2d\nu(\omega) = \int |\hat{f}_c(\omega)|^2|\hat{K}(H\omega)|^2d\omega - \int |\hat{f}_c(\omega)|^2|\hat{f}_X(\omega)|^2|\hat{K}(H\omega)|^2d\omega.
\]

\[
= \int |\hat{f}_c(\omega)|^2|\hat{K}(H\omega)|^2d\omega + O(1)
\]

\[
= \frac{1}{\det(H_{22})}I + O(1)
\]

\[
= \frac{1}{\det(H_{22})}J(1 + o(1)) + O(1).
\]

Noting that \(J\) is uniformly bounded above and below by Lemma 2.5 on p.43, we have

\[
\frac{1}{\det(H_{22})}J(1 + o(1)) + O(1) = \frac{1}{\det(H_{22})}J(1 + o(1)) + \frac{1}{\det(H_{22})}Jo(1)
\]

\[
= \frac{1}{\det(H_{22})}J(1 + o(1)).
\]

\[
\square
\]

2.6.4 Proof of Corollary 2.1

We must show

\[
\frac{1}{4}\int (\omega^TH^T\Sigma_KH\omega)^2d\mu(\omega) + \frac{1}{n\det(H_{22})}\int |\hat{f}_c(\omega_A, 0)|^2d\omega_A\int |\hat{K}(H_{12}H_{22}^{-1}\omega_{AC}, \omega_{AC})|^2d\omega_{AC}
\]

\[
= \frac{1}{4}\int (\omega^TS\omega)^2d\mu(\omega) + \frac{\pi^{d_2/2}}{n\det(S_{22})^{1/2}}\int |\hat{f}_c(\omega_A, 0)|^2d\omega_A.
\]

Since \(\Sigma_K = Id_{p}\) and \(S = HTH\) the bias components on both sides are equal. Now we show the variance components are equal. Under the assumption that \(K\) is standard normal, we have

\[
\frac{1}{n\det(H_{22})}\int |\hat{f}_c(\omega_A, 0)|^2d\omega_A\int |\hat{K}(H_{12}H_{22}^{-1}\omega_{AC}, \omega_{AC})|^2d\omega_{AC}
\]

\[
= \frac{1}{n\det(H_{22})}\int |\hat{f}_c(\omega_A, 0)|^2d\omega_A \int e^{-\omega_{AC}^T[H_{12}H_{22}^{-1}]^T[H_{12}H_{22}^{-1}]\omega_{AC}}d\omega_{AC}
\]

\[
= \frac{1}{n\det(H_{22})}\int |\hat{f}_c(\omega_A, 0)|^2d\omega_A \frac{\pi^{d_2/2}}{\det(Id_{d_2} + H_{22}^{-1}HTH_{12}^{-1}H_{22}^{-1})^{1/2}}
\]

\[
= \frac{\pi^{d_2/2}}{n\det(H_{22}^2 + H_{12}^T H_{12})^{1/2}}\int |\hat{f}_c(\omega_A, 0)|^2d\omega_A. \tag{2.50}
\]
We write $S$ in block form, i.e.

$$S = \begin{bmatrix}
S_{11} & S_{12} \\
S_{T12} & S_{22}
\end{bmatrix},$$

where $S_{11} \in \mathbb{R}^{d_1 \times d_1}$, $S_{12} \in \mathbb{R}^{d_1 \times d_2}$, $S_{22} \in \mathbb{R}^{d_2 \times d_2}$. Since $S = H^T H$, $S_{22} = H_{22}^2 + H_{12}^T H_{12}$. Thus Expression (2.50) becomes

$$\frac{\pi^{d_2/2}}{n \det(S_{22})^{1/2}} \int |\hat{f}_x(\omega_A, 0)|^2 d\omega_A.$$ 

2.6.5 Lemmas

**Lemma 2.1.** Under Assumptions A, $K$ is a symmetric density function in $\mathbb{R}^p$ with a characteristic function $\hat{K}$ that is four times continuously differentiable. Let $\Sigma_K$ be the variance of $K$. We Taylor expand $\hat{K}$ around 0, obtaining

$$\hat{K}(\omega) = 1 - \frac{\omega^T \Sigma_K \omega}{2} + R(\omega).$$

There exists $C$ such that for any $\omega$

$$R(\omega) \leq C||\omega||_\infty^4.$$ 

**Proof.** We bound the remainder term $R(\omega)$ by considering two cases.

1. $\{\omega : ||\omega||_\infty \leq 1\}$: Since $\hat{K}$ is four times continuously differentiable, there exists $D$ such that for any $\{j : \sum_{k=1}^p j_k = 4\}, \forall ||\omega||_\infty \leq 1$

$$\frac{\partial^4 \hat{K}}{\partial \omega_1^{j_1} \cdots \partial \omega_p^{j_p}}(\omega) < D. \quad (2.51)$$

Using the mean value form of the Taylor remainder we have (see e.g. Theorem 7.1 in Edwards Jr [1973] on page 131)

$$R(\omega) = \sum_{\{j : \sum_{k=1}^p j_k = 4\}} \frac{\partial^4 \hat{K}}{\partial \omega_1^{j_1} \cdots \partial \omega_p^{j_p}}(\xi) \prod_{k=1}^p \omega_k^{j_k}.$$ 

for some $\xi = t\omega$ for $t \in [0, 1]$. Using Equation (2.51) and noting $\prod_{k=1}^p \omega_k^{j_k} \leq ||\omega||_\infty^4$, for some $C$ we have

$$|R(\omega)| \leq C||\omega||_\infty^4.$$
2. \{\omega : \|\omega\|_{\infty} > 1\}: Note that for some \(D, \frac{\omega^T \Sigma K \omega}{2} \leq D \|\omega\|_{\infty}^2\). Also note that on the set \(\|\omega\|_{\infty} > 1\) we have \(\|\omega\|_{\infty}^2 \leq \|\omega\|_{4,\infty}^4\). We have

\[
|R(\omega)| = \left| \hat{K}(\omega) - 1 + \frac{\omega^T \Sigma K \omega}{2} \right|
\leq |\hat{K}(\omega)| + |1 + \frac{\omega^T \Sigma K \omega}{2}|
\leq 2 + \frac{\omega^T \Sigma K \omega}{2}
\leq 2 + D \|\omega\|_{\infty}^2
\leq 2\|\omega\|_{\infty}^2 + D \|\omega\|_{\infty}^2
\leq (2 + D) \|\omega\|_{\infty}^4
\]

Lemma 2.2. Under Assumption A and C

\[I = J(1 + o(1)).\]

Proof. Recall

\[
I = \int |\tilde{f}(\omega, 0)|^2 |\hat{K}(HH^{-1})|^2 d\omega,
\]

\[
J = \int |\tilde{f}(\omega, 0)|^2 d\omega \int |\hat{K}(H_{12}H_{22}^{-1}\omega_A, \omega_{AC})|^2 d\omega_{AC}.
\]

By Lemma 2.5 on p.43, \(\exists 0 < m \leq M < \infty\) such that uniform in \(H, m \leq J \leq M\). Thus

\[I = J(1 + o(1)) \iff I = J + o(1).\]

So it is sufficient to show that \(I = J + o(1)\). We show this in three steps: 1) construct \(I_1\) such that \(I = I_1 + o(1)\), 2) construct \(J_1\) such that \(J = J_1 + o(1)\), 3) show that \(I_1 = J_1 + o(1)\). Using 1), 3), and 2) we have

\[I = I_1 + o(1) = J_1 + o(1) = J + o(1).\]

1. \(I = I_1 + o(1)\): Define

\[D_n = \{\omega_A : \|\omega_A\|_{\infty} \leq \log \|H\|_{\infty}^{-1}\}\].

Write \(I = I_1 + I_2\) where

\[
I_1 = \int_{\omega_A \in D_n} \int_{\omega_{AC}} |\tilde{f}(\omega, 0)|^2 |\hat{K}(HH^{-1})|^2 d\omega_{AC} d\omega_A,
\]

\[
I_2 = \int_{\omega_A \notin D_n} \int_{\omega_{AC}} |\tilde{f}(\omega, 0)|^2 |\hat{K}(HH^{-1})|^2 d\omega_{AC} d\omega_A.
\]
We have
\[ I_2 \leq \int_{\omega_A \notin D_n} |\hat{f}_\epsilon(\omega_A, 0)|^2 \left( \sup_{\omega_A} \int \hat{K}(HH^{*-1}\omega)^2 \, d\omega_A \right) \, d\omega_A. \]

By Lemma 2.4 on p.42 \( \sup_{\omega_A} \int |\hat{K}(HH^{*-1}\omega)|^2 \, d\omega_A < C \) for some \( C \) uniformly in \( H \). Thus
\[ I_2 \leq C \int_{\omega_A \notin D_n} |\hat{f}_\epsilon(\omega_A, 0)|^2 \, d\omega_A \]
\[ = C \int_{\omega_A} |\hat{f}_\epsilon(\omega_A, 0)|^2 \mathbb{1}_{D_n^c} \, d\omega_A. \]

Since \( ||H||_\infty \to 0 \), given any \( \omega_A \in \mathbb{R}^{d_1} \), for sufficiently large \( N \), \( \omega_A \in D_n \forall n > N \). Therefore \( |\hat{f}_\epsilon(\omega_A, 0)|^2 \mathbb{1}_{D_n^c} \to 0 \) pointwise. \( \int |\hat{f}_\epsilon(\omega_A, 0)|^2 \, d\omega_A < \infty \) by Assumption 2.17. By the dominated convergence theorem \( I_2 \to 0 \) and
\[ I = I_1 + o(1). \]

2. \( J = J_1 + o(1) \): Write \( J = J_1 + J_2 \) where
\[ J_1 = \int_{\omega_A \in D_n} |\hat{f}_\epsilon(\omega_A, 0)|^2 \, d\omega_A \int \hat{K}(H_{12}H_{22}^{-1}\omega_{AC}, \omega_{AC})^2 \, d\omega_{AC}, \]
\[ J_2 = \int_{\omega_A \notin D_n} |\hat{f}_\epsilon(\omega_A, 0)|^2 \, d\omega_A \int \hat{K}(H_{12}H_{22}^{-1}\omega_{AC}, \omega_{AC})^2 \, d\omega_{AC}. \]

In the previous part we showed \( \int_{\omega_A \notin D_n} |\hat{f}_\epsilon(\omega_A, 0)|^2 \, d\omega_A \to 0 \). By Lemma 2.5
\[ \int |\hat{K}(H_{12}H_{22}^{-1}\omega_{AC}, \omega_{AC})|^2 \, d\omega_{AC} \]
is bounded above uniformly in \( H \). Therefore \( J_2 \to 0 \) and
\[ J = J_1 + o(1). \]

3. \( I_1 = J_1 + o(1) \): We show \( I_1 - J_1 = o(1) \). Recalling the definitions of \( I_1 \) and \( J_1 \), we must show
\[ \int_{\omega_A \in D_n} \int_{\omega_{AC}} |\hat{f}_\epsilon(\omega_A, 0)|^2 \left( |\hat{K}(HH^{*-1}\omega)|^2 - |\hat{K}(H_{12}H_{22}^{-1}\omega_{AC}, \omega_{AC})|^2 \right) \, d\omega_{AC} \, d\omega_A = o(1). \]

Note
\[ HH^{*-1}\omega = \begin{bmatrix} H_{11} & H_{12}H_{22}^{-1} \\ H_{12}^T & Id_{d_2} \end{bmatrix} \omega = \begin{bmatrix} H_{11} \\ H_{12}^T \end{bmatrix} \omega_A + \begin{bmatrix} H_{12}H_{22}^{-1} \\ Id_{d_2} \end{bmatrix} \omega_{AC}, \]
\[ H_{12}H_{22}^{-1}\omega_{AC}, \omega_{AC} = \begin{bmatrix} H_{12}H_{22}^{-1} \\ Id_{d_2} \end{bmatrix} \omega_{AC}. \]
Define
\[ t_A = \begin{bmatrix} H_{11} \\ H_{12}^T \end{bmatrix} \omega_A, \]
\[ t_{AC} = \begin{bmatrix} H_{12}H_{22}^{-1} \\ Id_{d_2} \end{bmatrix} \omega_{AC}. \]

We rewrite the left hand side of Equation (2.52) using this notation and bound it using the fact
\[ \int \int f(x)g(x,y)dx dy \leq \left( \int f(x)dx \right) \left( \sup_x \int g(x,y)dy \right). \]
We have
\[ \left| \int \omega_{AC} \in D_n \int \omega_A \in R_{d_2} C_2 o \left( \frac{1}{2} + ||\omega_{AC}||_2 \right)^2 \left( d \omega_{AC} + \Delta \right) \right| < \gamma. \]
c) Recalling that \( D_n \equiv \{ \omega_A : ||\omega_A||_\infty \leq \log(||H||_\infty^{-1}) \} \), for some \( C \) we have

\[
\lim_{n \to \infty} \sup_{\omega_A \in D_n} ||t_A||_2 = \lim_{n \to \infty} \sup_{\omega_A \in D_n} \left\| \begin{bmatrix} H_{11} \\ H_{12}^T \end{bmatrix} \omega_A \right\|_2 \\
\leq \lim_{n \to \infty} C ||H||_\infty \sup_{\omega_A \in D_n} ||\omega_A||_\infty \\
\leq \lim_{n \to \infty} C ||H||_\infty \log(\lambda^{1/2}) = 0.
\]

By Equation (2.57) there exists \( N \) such that \( \forall n > N \sup_{\omega_A \in D_n} ||t_A||_2 < \alpha \). We now show Equation (2.53) holds.

\[
\sup_{\omega_A \in D_n} \left\| \int_L |\hat{K}(t_A + t_{AC})|^2 - |\hat{K}(t_{AC})|^2 d\omega_{AC} \right\| \leq \sup_{\omega_A \in D_n} \left\| \int_L |\hat{K}(t_A + t_{AC})|^2 - |\hat{K}(t_{AC})|^2 d\omega_{AC} \right\| \\
+ \sup_{\omega_A \in D_n} \left\| \int_{L^C} |\hat{K}(t_A + t_{AC})|^2 - |\hat{K}(t_{AC})|^2 d\omega_{AC} \right\|
\]

By Equation (2.56) and the construction of \( \tau \) in Equation (2.55) we have

\[
\sup_{\omega_A \in D_n} \left\| \int_L |\hat{K}(t_A + t_{AC})|^2 - |\hat{K}(t_{AC})|^2 d\omega_{AC} \right\| \leq \tau L(L) < \gamma/2.
\]

Thus all that is left to show is

\[
\sup_{\omega_A \in D_n} \left\| \int_{L^C} |\hat{K}(t_A + t_{AC})|^2 - |\hat{K}(t_{AC})|^2 d\omega_{AC} \right\| < \gamma/2.
\]

We bound each function in the the integrand pointwise. Noting that \( \hat{K}(\omega) \leq \frac{C_0}{(1+\lambda)^{d_2+\Delta}} \)

(Assumption 2.20), \( ||t_{AC}||_2 = || \begin{bmatrix} H_{12}^{-1} \\ I_{d_2} \end{bmatrix} \omega_{AC} ||_2 \geq ||\omega_{AC}||_2 \), and \( ||t_A||_2 \leq \alpha \leq 1/2 \), we have

\[
|\hat{K}(t_A + t_{AC})| \leq \frac{C_0}{(1+\lambda)^{d_2+\Delta}} \\
\leq \frac{C_0}{(1+||t_{AC}||_2 - ||t_A||_2)^{d_2+\Delta}} \\
\leq \frac{C_0}{(1/2 + ||t_{AC}||_2)^{d_2+\Delta}} \\
\leq \frac{C_0}{(1/2 + ||\omega_{AC}||_2)^{d_2+\Delta}}.
\]
Again using \( ||t_{AC}||_2 \geq ||\omega_{AC}||_2 \) we have
\[
|\hat{K}(t_{AC})| \leq \frac{C_0}{(1 + ||t_{AC}||_2)^{d_2 + \Delta}} \leq \frac{C_0}{(1/2 + ||\omega_{AC}||_2)^{d_2 + \Delta}}.
\]

Using the bound in Equation (2.54), we have
\[
\sup_{\omega_A \in D_n} \left| \int_{L^C} |\hat{K}(t_A + t_{AC})|^2 - |\hat{K}(t_{AC})|^2 d\omega_{AC} \right|
\leq \sup_{\omega_A \in D_n} \max \left( \int_{L^C} |\hat{K}(t_A + t_{AC})|^2 d\omega_{AC}, \int_{L^C} |\hat{K}(t_{AC})|^2 d\omega_{AC} \right)
\leq \sup_{\omega_A \in D_n} \int_{L^C} \max \left( |\hat{K}(t_A + t_{AC})|^2, |\hat{K}(t_{AC})|^2 \right) d\omega_{AC}
\leq \sup_{\omega_A \in D_n} \int_{L^C} \frac{C_0^2}{(1/2 + ||\omega_{AC}||_2)^{2(d_2 + \Delta)}} d\omega_{AC}
\leq \gamma/2.
\]

**Lemma 2.3.** Let \( j = 1 \) or 2. Let
\[
H^* = \begin{bmatrix} I_{d_1} & 0 \\ 0 & H_{22} \end{bmatrix}.
\]

Under Assumptions C
\[
\int |\hat{f}_{\epsilon}^{\prime} (\omega)|^j |\hat{K}(H\omega)|^2 d\omega = \frac{1}{\det(H_{22})} \int |\hat{f}_{\epsilon}^{\prime} (\omega_A,0)|^j |\hat{K}(HH^{*-1}\omega)|^2 d\omega < \infty. \tag{2.58}
\]

**Proof.** By changing variables using \( u = H^*\omega \) and noting \( \hat{f}_{\epsilon}(\omega_A,\omega_{AC}) = \hat{f}_{\epsilon}(\omega_A,0)\forall \omega_A,\omega_{AC} \) (Assumption 2.16) we have
\[
\int |\hat{f}_{\epsilon}(\omega)|^j |\hat{K}(H\omega)|^2 d\omega = \frac{1}{\det(H^*)} \int |\hat{f}_{\epsilon}(H^{*-1}u)|^j |\hat{K}(HH^{*-1}u)|^2 du
\]
\[
= \frac{1}{\det(H_{22})} \int |\hat{f}_{\epsilon}(u_A,H_{22}^{-1}u_{AC})|^j |\hat{K}(HH^{*-1}u)|^2 du
\]
\[
= \frac{1}{\det(H_{22})} \int |\hat{f}_{\epsilon}(u_A,0)|^j |\hat{K}(HH^{*-1}u)|^2 du
\]
\[
= \frac{1}{\det(H_{22})} \int |\hat{f}_{\epsilon}(\omega_A,0)|^j |\hat{K}(HH^{*-1}\omega)|^2 d\omega.
\]
We have established the equality in Equation (2.58). Now it is sufficient to show the integral is bounded. Using the fact \( \int \int g(x) f(x, y) dy \, dx \leq \int g(x) \, dx \sup_x \int f(x, y) \, dy \) we have

\[
\int |\hat{f}_e(\omega_A, 0)|^j |\hat{K}(HH^{s-1}_c, 0)|^j \, d\omega_A \leq \left( \int |\hat{f}_e(\omega_A, 0)|^j \, d\omega_A \right) \left( \sup_{\omega_A} \int |\hat{K}(HH^{s-1}_c)|^j \, d\omega_A \right).
\]

\( \int |\hat{f}_e(\omega_A, 0)|^j \, d\omega_A \) is bounded by Assumption 2.17. \( \sup_{\omega_A} \int |\hat{K}(HH^{s-1}_c)|^j \, d\omega_A \) is bounded by Lemma 2.4.

**Lemma 2.4.** Let \( j = 1 \) or 2. Under Assumptions C

\[
\sup_{\omega_A} \int |\hat{K}(HH^{s-1}_c)|^j \, d\omega_A
\]

is bounded uniformly in \( H \).

**Proof.** In the following we express the argument to \( \hat{K} \) as

\[
HH^{s-1}_c = \begin{bmatrix} H_{11} & H_{12} H_{22}^{-1} \end{bmatrix} \omega = \begin{bmatrix} H_{11} \\ H_{12}^T \\ Id_{d_2} \end{bmatrix} \omega_A + \begin{bmatrix} H_{12} H_{22}^{-1} \end{bmatrix} \omega_{AC}.
\]

We have

\[
\sup_{\omega_A} \int |\hat{K}(HH^{s-1}_c)|^j \, d\omega_A = \sup_{\omega_A} \int |\hat{K} \left( \begin{bmatrix} H_{11} \\ H_{12}^T \\ Id_{d_2} \end{bmatrix} \omega_A + \begin{bmatrix} H_{12} H_{22}^{-1} \end{bmatrix} \omega_{AC} \right) |^j \, d\omega_A
\]

\[
\leq \sup_y \int \left| \hat{K} \left( y + \begin{bmatrix} H_{12} H_{22}^{-1} \\ Id_{d_2} \end{bmatrix} \omega_{AC} \right) \right|^j \, d\omega_A.
\]

We find an upper bound for the integral that does not depend on \( y \) or \( H \). Let

\[
B = \begin{bmatrix} H_{12} H_{22}^{-1} \\ Id_{d_2} \end{bmatrix}
\]

and \( c \) be the vector such that \( Bc \) is the projection of \( y \) onto the columnspace of \( B \). Using the definition of \( B \), a change of variables \( \omega_{AC} = s - c \), and Assumption 2.20 on the decay of \( \hat{K} \) we have

\[
\int \left| \hat{K} \left( y + \begin{bmatrix} H_{12} H_{22}^{-1} \\ Id_{d_2} \end{bmatrix} \omega_{AC} \right) \right|^j \, d\omega_{AC} = \int |\hat{K} (y + B \omega_{AC})|^j \, d\omega_{AC}
\]

\[
= \int |\hat{K} (y - Bc + Bs)|^j \, ds
\]

\[
\leq C_0^j \int \frac{1}{(1 + ||y - Bc + Bs||^2)^{(d_2 + \Delta)}} \, ds.
\]
By construction $y - Bc$ is orthogonal to $Bs$. Note that

$$
\| Bs \|_2 = \left\| \begin{bmatrix} H_{12} H_{22}^{-1} \\ 1 \end{bmatrix} s \right\|_2 = \left\| (H_{12} H_{22}^{-1} s, Id_{d_2} s) \right\|_2 \geq \| s \|_2.
$$

Using these two facts we have

$$
C_j^i \int \frac{1}{(1 + \| y - Bc + Bs \|_2)^{j(d_2 + \Delta)}} ds 
\leq C_j^i \int \frac{1}{(1 + \| Bs \|_2)^{j(d_2 + \Delta)}} ds
\leq C_j^i \int \frac{1}{(1 + \| s \|_2)^{j(d_2 + \Delta)}} ds
\leq C_j^i \int \frac{1}{(1 + \| s \|_2)^{(d_2 + \Delta)}} ds.
$$

By Lemma 2.6 p.44 this final integral is bounded. Clearly the bound is uniform in $H$.

**Lemma 2.5.** Under Assumptions C there exists $0 < m \leq M < \infty$ such that,

$$
m \leq J \leq M
$$

uniformly in $H$.

**Proof.** Recall

$$
J = \int |\hat{f}_\epsilon(\omega_A, 0)|^2 d\omega_A \int |\hat{K}(H_{12} H_{22}^{-1} \omega_A c, \omega_A c)|^2 d\omega_A c.
$$

Since $0 < \int |\hat{f}_\epsilon(\omega_A, 0)|^2 d\omega_A < \infty$ (Assumption 2.17) and does not depend on $H$, it is sufficient to show uniform bounds for $\int |\hat{K}(H_{12} H_{22}^{-1} \omega_A c, \omega_A c)|^2 d\omega_A c$. First we show an upper bound. Note

$$
\left\| (H_{12} H_{22}^{-1} \omega_A c, \omega_A c) \right\|_2 \geq \| \omega_A c \|_2.
$$

By Assumption 2.20 on the decay of $\hat{K}$ and Equation (2.59), we have

$$
\int |\hat{K}(H_{12} H_{22}^{-1} \omega_A c, \omega_A c)|^2 d\omega_A c \leq \int \frac{C_j^2}{(1 + \| (H_{12} H_{22}^{-1} \omega_A c, \omega_A c) \|_2)^{2(d_2 + \Delta)}} d\omega_A c
\leq \int \frac{C_j^2}{(1 + \| \omega_A c \|_2)^{2(d_2 + \Delta)}} d\omega_A c.
$$

This integral does not depend on $H$ and is bounded by Lemma 2.6 on p.44.

Now we show a uniform lower bound for $\int |\hat{K}(H_{12} H_{22}^{-1} \omega_A c, \omega_A c)|^2 d\omega_A c$. Since $\hat{K}$ is a characteristic function, it is uniformly continuous everywhere and $\hat{K}(0) = 1$. Thus $\exists t_0, t_1 > 0$
such that \( \forall \|\omega_A\|_\infty < t_0, \|\hat{K}(\omega_A)\| > t_1 \). By Assumption 2.23 \((H_{12}H_{22}^{-1} \text{ is uniformly bounded above})\) there exists \( C \) independent of \( H \) such that

\[
\sup_{\omega_A^C} \left\| H_{12}H_{22}^{-1} \right\| \omega_A^C \|_\infty \leq C \|\omega_A^C\|_\infty.
\]

Note that for \( \omega_A^C \in \{ \omega_A^C : \|\omega_A^C\|_\infty \leq t_0 \} \) we have \( \|H_{12}H_{22}^{-1}\omega_A^C, \omega_A^C\|_\infty \leq t_0 \) and hence \( \|\hat{K}(H_{12}H_{22}^{-1}\omega_A^C, \omega_A^C)\| > t_1^2 \). Thus we have

\[
\int |\hat{K}(H_{12}H_{22}^{-1}\omega_A^C, \omega_A^C)|^2 d\omega_A^C \geq \int_{\{\omega_A^C : \|\omega_A^C\|_\infty \leq t_0\}} |\hat{K}(H_{12}H_{22}^{-1}\omega_A^C, \omega_A^C)|^2 d\omega_A^C
\]

\[
\geq \int_{\{\omega_A^C : \|\omega_A^C\|_\infty \leq t_0\}} t_1^2 d\omega_A^C
\]

\[
> 0.
\]

Noting that \( C, t_0, t_1 \) were chosen independent of \( H \), we have a uniform lower bound that is greater than 0.

\[\square\]

**Lemma 2.6.** Say \( q \in \mathbb{Z}^+ \) and \( c_0, c_1 > 0 \). Then

\[
\int_{s \in \mathbb{R}^q} \frac{1}{(c_0 + \|s\|_2)^{(q+c_1)}} ds < \infty.
\]

**Proof.** Let \( D_q = \{ s \in \mathbb{R}^q : \|s\|_2 \geq 1 \} \) and \( D_q^C = \mathbb{R}^q \setminus D_q \). We have

\[
\int_{s \in \mathbb{R}^q} \frac{1}{(c_0 + \|s\|_2)^{(q+c_1)}} ds = \int_{D_q^C} \frac{1}{(c_0 + \|s\|_2)^{(q+c_1)}} ds + \int_{D_q} \frac{1}{(c_0 + \|s\|_2)^{(q+c_1)}} ds
\]

\[
\leq \int_{D_q^C} \frac{1}{c_0^{(q+c_1)}} ds + \int_{D_q} \frac{1}{\|s\|_2^{(q+c_1)}} ds.
\]

\[
\int_{D_q^C} \frac{1}{c_0^{(q+c_1)}} ds \text{ is bounded because } D_q^C \text{ is a bounded set and the integrand is constant. So it is sufficient to show that}
\]

\[
\int_{D_q} \frac{1}{\|s\|_2^{(q+c_1)}} ds < \infty. \tag{2.60}
\]

Using Theorem 2.1.3 of Muirhead [2009] we transform to polar coordinates \( r, \theta_1, \ldots, \theta_{q-1} \). We have

\[
\int_{D_q} \frac{1}{\|s\|_2^{(q+c_1)}} ds
\]

\[
= \int_{r=1}^{\infty} \int_{\theta_j \in [0, \pi] \forall j \in \{1, \ldots, q-2\}} \int_{\theta_{q-1} \in [0, 2\pi]} r^{q-1} \sin^{q-2}(\theta_1) \sin^{q-3}(\theta_2) \cdots \sin(\theta_{q-2}) d\theta_1 \cdots d\theta_{q-1} dr
\]

\[
= \left( \int_{\theta_j \in [0, \pi] \forall j \in \{1, \ldots, q-2\}} \int_{\theta_{q-1} \in [0, 2\pi]} \prod_{j=1}^{q-2} \sin^{q-j-1}(\theta_j) d\theta_1 \cdots d\theta_{q-1} \right) \left( \int_{1}^{\infty} \frac{1}{r^{1+c_1}} dr \right).
\]
The first integrand is bounded by 1. The domain of integration is bounded, so the integral is bounded. The integral across $r$ is finite because $c_1 > 0$.

\[ \frac{1}{n} \int d\nu(\omega) + \left( \frac{1}{4} \int (\omega^T S \omega)^2 d\mu(\omega) - \frac{1}{n} \int (\omega^T S \omega) d\nu(\omega) \right) \left( 1 + O(||H||_\infty^2) \right) \]

where $S = H^T \Sigma_k H$. Using vec notation and the identity vec$(EFG) = (G^T \otimes E)\text{vec}(F)$ where $\otimes$ denotes Kronecker product (see Equation 5 on page 67 in Henderson and Searle [1979]), we write the optimization program for $S$ as

\[ S^* = \arg\min_{S \succeq 0} \text{vec}(S)^T B \text{vec}(S) - \frac{1}{n} \text{vec}(S)^T V \tag{2.61} \]

where

\[ B = \frac{1}{4} \int (\omega \otimes \omega)(\omega \otimes \omega)^T d\mu(\omega), \]
\[ V = \int (\omega \otimes \omega) d\nu(\omega). \]

It is important to note that $B$ and $V$ cannot be computed from the data because they depend on the unknown function $\hat{f}_X(\omega)$. In practice we could use plug-in estimators to approximate these integrals.

The unconstrained solution to optimization problem (2.61) may not be positive semidefinite, so we cannot omit the $S \succeq 0$ constraint and use a quadratic solver (see Subsection 2.7.2 for an example). Also note that one cannot analytically solve the unconstrained version of optimization problem (2.61) and then check whether the resulting $S^*$ is positive semidefinite. In other words, the following procedure is not valid:

\[ g(\text{vec}(S)) \equiv \text{vec}(S)^T B \text{vec}(S) - \frac{1}{n} \text{vec}(S)^T V, \]
\[ \implies \nabla g(\text{vec}(S)) = 2B\text{vec}(S) - \frac{1}{n}V. \]

Setting the gradient equal to 0 and solving we have

\[ \text{vec}(S^*) = \frac{1}{2n} B^{-1} V. \]
One could then check whether $S^* \succeq 0$. This procedure is not valid because $B$ is not invertible.

To see that $B$ is not invertible, note that the vector $(\omega \otimes \omega)$ has $p^2$ elements, but not $p^2$ unique elements. For example when $p = 2$, $(\omega \otimes \omega) = (\omega_1^4, \omega_1 \omega_2, \omega_1 \omega_2, \omega_2^2)^T$. When the $j$th and $k$th elements of $(\omega \otimes \omega)$ are equal, the $j$th and $k$th rows of $(\omega \otimes \omega)(\omega \otimes \omega)^T$ are equal. Thus at least two rows of $B \equiv \int (\omega \otimes \omega)(\omega \otimes \omega)^T d\mu(\omega)$ are equal, implying that $B$ cannot be inverted.

### 2.7.2 Non Positive Semidefinite Solutions

Theorem 2.2 on p.22 requires that the bandwidth matrix $H$ be positive semidefinite (see Assumption (2.9)). As a result, the optimization problems for determining the optimal $H$ involve positive semidefinite constraints in the general case (see Equation (2.61)) and first orthant constraints in the diagonal bandwidth matrix case (see Equation (2.14)). Since it is generally computationally faster to solve unconstrained optimization problems, it is useful to know if the unconstrained optimum ever violates the constraints. If not, one can simply solve the unconstrained problem. Unfortunately, even in simple cases, the constraints turn out to be necessary. Here we work out an example where $f_X$ and $f_\epsilon$ are bivariate independent normals with $\epsilon$ having small variance along one direction. The kernel is normal with identity covariance. The normality is not essential for this example, but makes the computations simpler.

We begin by showing that the optimal bandwidth matrix is diagonal, implying that optimizing over the full bandwidth matrix and the diagonal matrix are equivalent. We then show that when optimizing over the unconstrained diagonal matrix, the direction in which $\epsilon$ has larger variance yields a “negative squared bandwidth”. Consider:

\[
\begin{align*}
 f_X &\sim N(0, Id_2), \\
 f_\epsilon &\sim N(0, \begin{bmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{bmatrix}), \\
 \Sigma_K &\equiv \int x x^T K(x) dx = Id_2.
\end{align*}
\]

We parameterize the bandwidth matrix using $H = \begin{bmatrix} h_{11} & h_{12} \\ h_{12} & h_{22} \end{bmatrix}$. First consider optimizing
over the entire bandwidth matrix, Equation (2.61). In our case

\[ S \equiv H^T \Sigma_K H = H^T H, \]

\[ B = \int (\omega \otimes \omega)(\omega \otimes \omega)^T d\mu(\omega) = \int \begin{bmatrix} \omega_1^4 & \omega_1^2 \omega_2 & \omega_1^2 \omega_2^2 & \omega_1 \omega_2^3 \\
\omega_1^3 \omega_2 & \omega_1^2 \omega_2^2 & \omega_1^2 \omega_2^3 & \omega_1 \omega_2^4 \\
\omega_1^2 \omega_2^2 & \omega_1 \omega_2^3 & \omega_1^3 \omega_2^2 & \omega_1 \omega_2^4 \\
\omega_1 \omega_2^3 & \omega_1 \omega_2^4 & \omega_1 \omega_2^4 & \omega_1^4 \end{bmatrix} d\mu(\omega), \]

\[ V = \int (\omega \otimes \omega) d\nu(\omega) = \int \begin{bmatrix} \omega_1^2 \\
\omega_1 \omega_2 \\
\omega_2^2 \end{bmatrix} d\nu(\omega). \]

So Equation (2.61) becomes

\[ \text{vec}(H^T H)^T \int \begin{bmatrix} \omega_1^4 & \omega_1^2 \omega_2 & \omega_1^2 \omega_2^2 & \omega_1 \omega_2^3 \\
\omega_1^3 \omega_2 & \omega_1^2 \omega_2^2 & \omega_1^2 \omega_2^3 & \omega_1 \omega_2^4 \\
\omega_1^2 \omega_2^2 & \omega_1 \omega_2^3 & \omega_1^3 \omega_2^2 & \omega_1 \omega_2^4 \\
\omega_1 \omega_2^3 & \omega_1 \omega_2^4 & \omega_1 \omega_2^4 & \omega_1^4 \end{bmatrix} d\mu(\omega) \text{vec}(H^T H) \]

\[ -\frac{1}{n} \text{vec}(H^T H)^T \left( \int \begin{bmatrix} \omega_1^2 \\
\omega_1 \omega_2 \\
\omega_2^2 \end{bmatrix} d\nu(\omega) \right). \]

The integration causes those terms involving odd powers of \( \omega_i \) to be 0 by independence and symmetry of \( d\nu(\omega) \) and \( d\mu(\omega) \). Additionally the center \( \omega_1^2 \omega_2^2 \) terms are moved outside the main expression. We have

\[ \text{vec}(H^T H)^T \int \begin{bmatrix} 0 & 0 & \omega_1^2 \omega_2^2 \\
0 & 0 & 0 \\
\omega_1^2 \omega_2 & 0 & 0 \\
0 & \omega_2 \end{bmatrix} d\mu(\omega) \text{vec}(H^T H) \]

\[ -\frac{1}{n} \text{vec}(H^T H)^T \left( \int \begin{bmatrix} \omega_1^2 \\
0 \\
0 \end{bmatrix} d\nu(\omega) \right) + 4(h_{12}(h_{11} + h_{22}))^2 \int \omega_1^2 \omega_2^2 d\mu(\omega). \]

Since

\[ H^T H = \begin{bmatrix} h_{11}^2 + h_{12}^2 & h_{12}(h_{11} + h_{22}) \\
h_{12}(h_{11} + h_{22}) & h_{12}^2 + h_{22}^2 \end{bmatrix}, \]
minimization of the first two terms depends on \((h_{11}^2 + h_{12}^2, h_{22}^2 + h_{12}^2)\). So by setting \(h_{12} = 0\) we make the third term in the expression 0, without restricting minimization of the first two terms. So for the general “bandwidth” matrix the minimum occurs when the off-diagonal elements are 0.

Now let \(h_S = (h_{11}^2, h_{22}^2)\). We study the diagonal optimization problem (2.14)

\[
\min_{h_S} h_S^T B' h_S - \frac{1}{n} h_S^T V',
\]

where

\[
B'_{i,j} = \frac{1}{4} \int \omega_i^2 \omega_j^2 d\mu(\omega) = \frac{1}{4} \int \omega_i^2 \omega_j^2 |\widehat{f}_X(\omega)|^2 |\widehat{f}_\epsilon(\omega)|^2 d\omega,
\]

\[
V'_i = \int \omega_i^2 d\nu(\omega) = \int \omega_i^2 |\widehat{f}_\epsilon(\omega)|^2 d\omega - \int \omega_i^2 |\widehat{f}_X(\omega)|^2 |\widehat{f}_\epsilon(\omega)|^2 d\omega.
\]

With no restrictions on \(h_S\) the optimum is

\[
h^*_S = \frac{1}{2n} B'^{-1} V'.
\]

We now compute this quantity for the given densities. First compute \(B'\):

\[
4B'_{11} = \int \omega_1^4 |\widehat{f}_{X_1}(\omega_1)|^2 |\widehat{f}_{\epsilon_1}(\omega_1)|^2 d\omega_1 \int |\widehat{f}_{X_2}(\omega_2)|^2 |\widehat{f}_{\epsilon_2}(\omega_2)|^2 d\omega_2
= \left( \frac{3}{4} \right)^3 \left( \frac{\pi}{1 + \sigma_1^2} \right)^5 \left( \frac{\pi}{1 + \sigma_2^2} \right)^5,
\]

\[
4B'_{22} = \int \omega_2^4 |\widehat{f}_{X_2}(\omega_2)|^2 |\widehat{f}_{\epsilon_2}(\omega_2)|^2 d\omega_2 \int |\widehat{f}_{X_1}(\omega_1)|^2 |\widehat{f}_{\epsilon_1}(\omega_1)|^2 d\omega_1
= \left( \frac{3}{4} \right)^3 \left( \frac{\pi}{1 + \sigma_1^2} \right)^5 \left( \frac{\pi}{1 + \sigma_2^2} \right)^5,
\]

\[
4B'_{12} = \int \omega_1^2 |\widehat{f}_{X_1}(\omega_1)|^2 |\widehat{f}_{\epsilon_1}(\omega_1)|^2 d\omega_1 \int \omega_2^2 |\widehat{f}_{X_2}(\omega_2)|^2 |\widehat{f}_{\epsilon_2}(\omega_2)|^2 d\omega_2
= \left( \frac{1}{2} \right)^3 \left( \frac{\pi}{1 + \sigma_1^2} \right)^3 \left( \frac{\pi}{1 + \sigma_2^2} \right)^3.
\]

Since \(B'\) and \(B'^{-1}\) are symmetric, we write only the upper triangle:

\[
B' = \frac{\pi}{16} \begin{bmatrix}
\frac{1}{\sqrt{(1 + \sigma_1^2)^5(1 + \sigma_2^2)^5}} & \frac{1}{\sqrt{(1 + \sigma_1^2)^3(1 + \sigma_2^2)^3}} \\
\frac{1}{\sqrt{(1 + \sigma_2^2)^5(1 + \sigma_1^2)^5}} & \frac{1}{\sqrt{(1 + \sigma_1^2)^3(1 + \sigma_2^2)^3}}
\end{bmatrix}.
\]
Taking the inverse we obtain

\[
B'^{-1} = \frac{2(1 + \sigma_1^2)(1 + \sigma_2^2)^3}{\pi} \left[ \frac{1}{\sqrt{(1+\sigma_2^2)^3(1+\sigma_1^2)}} - \frac{1}{3 \sqrt{(1+\sigma_2^2)^5(1+\sigma_1^2)}} \right]
\]

\[
= \frac{2}{\pi} \left[ 3\sqrt{(1 + \sigma_2^2)(1 + \sigma_1^2)^5} - \sqrt{(1 + \sigma_1^2)^3(1 + \sigma_2^2)^3} \right] \frac{1}{3 \sqrt{(1+\sigma_2^2)^5(1+\sigma_1^2)}} .
\]

For \(V'\) we have

\[
V' = \frac{\pi}{2} \left( \left[ \sigma_1^{-3} \sigma_2^{-1} \right] - \left[ \frac{1}{\sqrt{(1+\sigma_2^2)^3(1+\sigma_1^2)}} \right] \right)
\]

\[
= \frac{\pi}{2 \sigma_2^3} \left( \left[ \begin{array}{c} 0 \\ \sigma_1^{-1} \end{array} \right] + \sigma_2^2 \left[ \sigma_1^{-3} \right] - \sigma_2^3 \left[ \frac{1}{\sqrt{(1+\sigma_2^2)^3(1+\sigma_1^2)}} \right] \right).
\]

So the optimal \(h_S\) is

\[
h_S^* = \frac{1}{2n} B'^{-1} V'
\]

\[
= \frac{1}{2n \sigma_2^2} \left( \left[ -\sigma_1^{-1} \sqrt{(1 + \sigma_1^2)^3(1 + \sigma_2^2)^3} \right] \right.
\]

\[
+ \sigma_2^2 \left[ \frac{3 \sigma_1^{-3} \sqrt{(1 + \sigma_2^2)(1 + \sigma_1^2)^5}}{-\sigma_1^{-3} \sqrt{(1 + \sigma_1^2)^3(1 + \sigma_2^2)^3}} \right] - 2 \sigma_2^3 \left[ \frac{1 + \sigma_1^2}{1 + \sigma_2^2} \right].
\]

For \(\sigma_2\) close to 0 and small relative to \(\sigma_1\) this quantity is approximately

\[
h_S^* \approx \frac{1}{2n \sigma_1 \sigma_2^3} \left( \left[ -\sqrt{(1 + \sigma_1^2)^3} \right] \right) \left( \frac{3}{3 \sqrt{(1 + \sigma_1^2)}} \right).
\]

(2.62)

The unconstrained optimization results in an \(h_S^*\) with negative elements.
Chapter 3

Nadaraya–Watson for Prediction with Feature Measurement Error

3.1 Introduction

We consider prediction of a response when the features of an observation are subject to measurement error. We assume existence of a sample, termed training data, of observations with known responses and error free features. We study how measurement error affects convergence of the optimal bandwidth ($H$) and mean squared error ($MSE$) in a modified version of the Nadaraya–Watson (NW) estimator [Nadaraya, 1964, Watson, 1964].

One application of regression with measurement error is to photometric redshift estimation. In this problem, observations are galaxies. Astronomers seek to predict a redshift, the response, for each galaxy. Associated with each galaxy is a vector of colors, the features. Using a set of galaxies with estimated colors and redshift, Wang et al. [2007] constructed a function mapping colors to redshift using the Nadaraya–Watson estimator with the intention of using this function to predict redshift for galaxies where only the colors are known.

The colors of each galaxy are observed with measurement error. The distribution of the measurement error varies from galaxy to galaxy. For a galaxy of unknown redshift, incorporating this measurement error into the prediction offers the potential to improve accuracy. While Wang et al. [2007] ignored measurement error in colors, other authors such as Ball et al. [2008] have used these uncertainty measurements for photometric redshift estimation. We note that in photometric redshift estimation problems there is often measurement error in training data features. We do not address measurement error in training data in this work.

In Section 3.2 we introduce the problem, define an estimator $\tilde{\mu}_{Z|Y}$, and relate $\tilde{\mu}_{Z|Y}$ to the Nadaraya–Watson estimator. Analysis of the MSE of $\tilde{\mu}_{Z|Y}$ is difficult because it is the
ratio of dependent random variables. In Section 3.3 we present asymptotic expansions of the bias and variance of $\tilde{\mu}_{Z|Y}$ which involve only products of random variables plus higher order terms. This expansion enables us to derive asymptotic approximations to the squared bias and variance of $\tilde{\mu}_{Z|Y}$ at the $n^{-2}$ order in Section 3.4. While the terms in these expansions are complex, the relationship between the bandwidth parameter $H$ and $\text{MSE}$ of $\tilde{\mu}_{Z|Y}$ remains fairly simple. In Section 3.5, we discuss methods for choosing the bandwidth $H$ that minimizes the $\text{MSE}$. In Section 3.6 we present an example where the bandwidth does not affect the $\text{MSE}$ up to the $n^{-2}$ order. We discuss statistical interpretations of this phenomenon. We summarize our results and offer directions for future research in Section 3.7.

3.2 Problem Setup

We observe independent, identically distributed training data $\{(X_i, Z_i)\}_{i=1}^n$ where

$$Z_i = \mu_{Z|X}(X_i) + \delta_i.$$  

Here $X_i \sim f_X$, $X_i \in \mathbb{R}^p$ and $Z_i \in \mathbb{R}$. The $\delta_i \sim f_\delta$ are independent, $\delta_i \perp X_i$, with $E[\delta_i] = 0$ and $E[\delta_i^2] = \sigma_\delta^2$. We observe

$$Y = X + \epsilon,$$

but do not observe $X$ or $\epsilon$. However the density $f_\epsilon$ of $\epsilon$ is known. Denote the distribution function of $\epsilon$ by $F_\epsilon$.

The unobserved response we aim to predict is $Z = \mu_{Z|X}(X) + \delta$. Here $\delta$ and $X$ are independent, $\delta$ is independent identically distributed with training $\delta_i$, and $X$ is independent, identically distributed with training $X_i$. The density of $Y$ is denoted $f_Y$.

We seek to minimize mean squared error (MSE) in estimating $Z$. It is sufficient to construct estimators which minimize MSE in predicting

$$\mu_{Z|Y}(Y) \equiv E[Z|Y].$$

This is a consequence of the following theorem.

**Theorem 3.1.** Let $\tilde{Z}$ be any estimator of $Z$ i.e., $\tilde{Z} = g(\{(X_i, Z_i)\}_{i=1}^n, Y)$ for some deterministic function $g$. Then

$$E[(\tilde{Z} - Z)^2|Y] = E[(\tilde{Z} - \mu_{Z|Y}(Y))^2|Y] + \text{Var}(Z|Y).$$

See Subsection 3.8.1 on page 61 for a proof. Since only $E[(\tilde{Z} - \mu_{Z|Y}(Y))^2|Y]$ on the right hand side depends on $\tilde{Z}$, we focus on estimation of $\mu_{Z|Y}(y)$. The term

$$\text{Var}(Z|Y) = (E[\mu_{Z|X}(X)|Y] - \mu_{Z|Y}(Y))^2 + \sigma_\delta^2$$
is irreducible error in the prediction \( \hat{Z} \). In the case of no measurement error \( \mathbb{P}(\epsilon = 0) = 1 \) this term is \( \sigma^2 \). See Hastie et al. [2009] Section 7.3 page 223 for this decomposition in the error free case.

### 3.2.1 Definition of Estimator \( \hat{\mu}_{Z|Y} \)

We now construct an estimator for \( \mu_{Z|Y}(y) \) that is a weighted average of the responses \( Z_i \). We discuss some intuition for the weighting in Subsection 3.2.2. Let \( K \) be a function \( K : \mathbb{R}^p \rightarrow \mathbb{R} \) and define

\[
K_H(\cdot) = \frac{1}{\det(H)}K(H^{-1} \cdot),
\]

for any \( p \times p \) positive definite matrix \( H \). We refer to \( K \) as the kernel and \( H \) as the bandwidth matrix. Define

\[
\hat{f}_{Y,X_i}(y) = \int K_H(y - X_i - \epsilon)dF_\epsilon.
\]

Note that \( \hat{f}_{Y,X_i}(y) \) is the regularized estimator of \( f_Y(y) \) studied in Chapter 2. In this work \( y \) is fixed, so we often refer to \( \hat{f}_{Y,X_i}(y) \) as \( \hat{f}_{Y,X_i} \). We estimate \( \mu_{Z|Y}(y) \) using

\[
\hat{\mu}_{Z|Y}(y) \equiv \left\{ \frac{\sum_{i=1}^{n} Z_i \hat{f}_{Y,X_i}}{\sum_{i=1}^{n} \hat{f}_{Y,X_i}} : \sum_{i=1}^{n} \hat{f}_{Y,X_i} > 0 \\
0 : otherwise \right. \tag{3.1}
\]

### 3.2.2 \( \hat{\mu}_{Z|Y} \), Nadaraya–Watson, and Kernel Free Estimators

\( \hat{\mu}_{Z|Y} \) is an adaptation of the Nadaraya–Watson (NW) estimator to the setting where there is measurement error in features. To see this, note that when \( \mathbb{P}(\epsilon = 0) = 1 \), \( \mu_{Z|Y} = \mu_{Z|X} \) and \( \hat{\mu}_{Z|Y}(y) \) in Equation (3.1) is

\[
\frac{\sum_{i=1}^{n} Z_i K_H(y - X_i)}{\sum_{i=1}^{n} K_H(y - X_i)},
\]

the standard NW estimator. When there is measurement error in \( Y \), it is possible to construct estimators that do not require a kernel or bandwidth. For example, Carroll et al. [2009] (Equation 2.6) proposed using

\[
\hat{\mu}_{Z|Y}(y) = \left\{ \frac{\sum_{i=1}^{n} Z_i f_\epsilon(y - X_i)}{\sum_{i=1}^{n} f_\epsilon(y - X_i)} : \sum_{i=1}^{n} f_\epsilon(y - X_i) > 0 \\
0 : otherwise \right. \tag{3.2}
\]

as an estimator for \( \mu_{Z|Y} \). \( \hat{\mu}_{Z|Y} \) is a kernel free version of \( \hat{\mu}_{Z|Y} \) in Equation (3.1). To see this, note that fixing the training data \( \{(X_i, Z_i)\}_{i=1}^{n} \) and assuming \( \sum_{i=1}^{n} f_\epsilon(y - X_i) > 0 \),

\[
\lim_{H \rightarrow 0} \hat{\mu}_{Z|Y}(y) = \lim_{H \rightarrow 0} \frac{\sum_{i=1}^{n} Z_i \int K_H(y - X_i - \epsilon)dF_\epsilon}{\sum_{i=1}^{n} \int K_H(y - X_i - \epsilon)dF_\epsilon} = \frac{\sum_{i=1}^{n} Z_i f_\epsilon(y - X_i)}{\sum_{i=1}^{n} f_\epsilon(y - X_i)} = \hat{\mu}_{Z|Y}(y).
\]
3.2.3 Minimization of MSE

We study how to choose $H$ to minimize the MSE of $\tilde{\mu}_{Z|Y}$. Specifically we seek

$$H_{opt} = \arg\min_{H \succ 0} \text{MSE}(\tilde{\mu}_{Z|Y}(y)) = \arg\min_{H \succ 0} \mathbb{E}[(\mu_{Z|Y}(y) - \tilde{\mu}_{Z|Y}(y))^2].$$

(3.3)

While we suppress the dependence of $H_{opt}$ on $n$, we are in fact looking for a sequence of matrices. Direct minimization of the MSE at finite $n$ is not possible, so we perform asymptotic approximations and study the behavior of $H_{opt}$ for large $n$.

In particular, we focus on how the feature error $F_\epsilon$ impacts the MSE and optimal amount of smoothing $H_{opt}$. We highlight the cases where $\epsilon$ is highly concentrated around 0 (see Equation (3.12)) and where the error is spread across the support of $X$ (see Section 3.6).

As we show in Section 3.4, the MSE is invariant to $H$ at the $n^{-1}$ level. We compute terms in the asymptotic expansion up to order $n^{-2}$. Many terms of order $n^{-2}$ do not depend on $H$ and are represented as $Cn^{-2}$ where $C$ is some constant that does not depend on $H$ and whose value may change at each appearance.

3.3 Expansion for Bias and Variance

Analysis of $\tilde{\mu}_{Z|Y}$ is complicated by the fact that it is a ratio of dependent random variables. We now present a theorem that simplifies the asymptotic bias and variance of $\tilde{\mu}_{Z|Y}(y)$. We first make two sets of assumptions. These assumptions are used in all subsequent theorems and lemmas of this work. The first set of assumptions applies to the kernel $K$ and the bandwidth $H$. Since we choose these quantities, these assumptions can always be satisfied in practice.

Assumptions D.

$$K(u) = K(-u) \forall u$$

(3.4)

$K$ has four moments

(3.5)

$$H = H_n \succ 0 \text{ (sequence is positive definite)}$$

(3.6)

$$||H||_\infty = O(n^{-1/2})$$

(3.7)

Common kernels such as the uniform and normal satisfy these conditions. We now make assumptions on the underlying regression function $\mu_{Z|X}$, the moments of $\delta$, and the error density $f_\epsilon$. 
CHAPTER 3. NADARAYA–WATSON FOR PREDICTION WITH FEATURE MEASUREMENT ERROR

Assumptions E.

\[ f_Y(y) > 0 \] \hspace{1cm} (3.8)
\[ \mu_{Z|X} \text{ is uniformly bounded} \] \hspace{1cm} (3.9)
\[ \mathbb{E}[\delta^4] < \infty \] \hspace{1cm} (3.10)

Let \( l \in \{0, 1, 2, 3, 4\} \). Let \( j \in \{0, 1\}^p \) such that \( \sum_{k=1}^{p} j_k = l \). We assume that for all such \( l \) and \( j \)

\[ \frac{\partial^l f}{\partial \epsilon_1^{j_1} \ldots \partial \epsilon_p^{j_p}} \] \hspace{1cm} (3.11)

is uniformly bounded.

Assumption 3.8 requires the density of \( Y \) to be positive at the point \( y \) where we wish to predict the response. This assumption is essentially unavoidable. Assumption 3.9 is satisfied if \( \mu_{Z|X} \) is continuous and supported on a compact set. The existence of four moments for \( \delta \) in Assumption 3.10 permits many interesting error distributions, such as normal and uniform. Assumption 3.11 requires the error density to be smooth. In this work we assume the error density is known, so this assumption could be checked.

**Theorem 3.2.** Under Assumptions D and E we have

\[ \text{Bias}(\tilde{\mu}_{Z|Y}(y)) = \mathbb{E}[\omega_n] + O(n^{-3/2}), \]
\[ \text{Var}(\tilde{\mu}_{Z|Y}(y)) = \text{Var}(\omega_n) + 2\text{Cov}(\omega_n, s_n) + O(n^{-5/2}), \]

where

\[ \omega_n = f_Y^{-1} \Delta_B - f_Y^{-1} \mu_{Z|Y} \Delta_A - f_Y^{-2} \Delta_A \Delta_B + f_Y^{-2} \Delta_A^2 \mu_{Z|Y}, \]
\[ s_n = \Delta_B \Delta_A^2 f_Y^{-3} - \Delta_A^3 f_Y^{-3} \mu_{Z|Y}, \]
\[ \Delta_A = \frac{1}{n} \sum_{i=1}^{n} \tilde{f}_{Y,X_i}(y) - f_Y(y), \]
\[ \Delta_B = \frac{1}{n} \sum_{i=1}^{n} Z_i \tilde{f}_{Y,X_i}(y) - f_Y(y) \mu_{Z|Y}(y). \]

See Subsection 3.8.2 on page 62 for a proof. The terms \( \mathbb{E}[\omega_n] \), \( \text{Var}(\omega_n) \), and \( \text{Cov}(\omega_n, s_n) \) are easier to analyze than \( \tilde{\mu}_{Z|Y} \) directly because they do not involve ratios of dependent random variables.
3.4 Asymptotic Approximations to Bias and Variance

We present two theorems which describe the asymptotic behavior of the squared bias and variance of $\tilde{\mu}_{Z|Y}(y)$. These expansions are based on the results of Theorem 3.2. We seek to use these asymptotic expansions to select the $H$ that minimizes $MSE(\tilde{\mu}_{Z|Y})$ in Equation (3.3). To this end, we focus on terms that involve $H$. Second order terms that do not depend on $H$ are grouped together and denoted by $Cn^{-2}$ where $C$ is some constant that does not depend on $H$ or $n$. The precise value of $C$ may change with each appearance.

For these theorems it is useful to reparameterize the bandwidth matrix. Let $S = H^T \Sigma K H$. $S$ is the variance of $K_H$. The asymptotic squared bias and variance of $\tilde{\mu}_{Z|Y}$ depend on $H$ only through $S$, so we express the squared bias, variance, and MSE as a function of $S$.

Let $I(f)(x)$ and $\nabla(f)(x)$ be the Hessian and gradient of $f$ evaluated at $x$, respectively.

We define several quantities that do not depend on $H$ or $n$. After defining the terms we show all expectations exist. We will discuss the interpretation of these terms later in this section. Let

$$\nu_y(x) = (\mu_{Z|X}(x) - \mu_{Z|Y}(y))^2 + \sigma^2,$$

$$V = f_Y^{-2} \mathbb{E}[f^2_\epsilon(y - X_1)\nu_y(X_1)],$$

$$A_1 = (2f_Y)^{-1} \left(I(\mu_{Z|Y})f_Y + 2\nabla(\mu_{Z|Y})\nabla^T(f_Y)\right),$$

$$A_2 = -2f_Y^{-2} \mathbb{E}[(\mu_{Z|X}(X_1) - \mu_{Z|Y})f^2_\epsilon(y - X_1)]A_1,$$

$$A_3 = -f_Y^{-3} \mathbb{E}[f^2_\epsilon(y - X_1)\nu_y(X_1)]I(f_Y),$$

$$A_4 = f_Y^{-2} \mathbb{E}[I(f_\epsilon)(y - X_1)f_\epsilon(y - X_1)\nu_y(X_1)].$$

By Assumption 3.9 $\mu_{Z|X}$ is bounded. Hence $\mu_{Z|Y}$ is bounded. Hence $\nu_y$ is bounded. By Assumption 3.11 $f_\epsilon$ is bounded. Thus $f^2_\epsilon(y - X_1)\nu_y(X_1)$ and $(\mu_{Z|X}(X_1) - \mu_{Z|Y})f^2_\epsilon(y - X_1)$ are bounded and the expectations in $V$, $A_2$ and $A_3$ exist. Noting that $I(f_\epsilon)$ is bounded by Assumption 3.11, the expectation in $A_4$ exists.

3.4.1 Bias

Theorem 3.3. Under Assumptions D and E

$$\text{Bias}(\tilde{\mu}_{Z|Y})^2 = \left[tr(SA_1)\right]^2 + \frac{1}{n}tr(SA_2) + Cn^{-2} + O(n^{-5/2}).$$

See Subsection 3.8.3 for a proof of this theorem. Note that the squared bias depends on $n$, unlike in other problems such as kernel density estimation (see e.g., Chapter 2). This is
caused by the fact that the Nadaraya–Watson estimator is not the average of independent, identically distributed random variables. The term \([\text{tr } (SA_1)]^2\) closely resembles the bias for the error free NW estimator. In particular with no error

\[
A_1 = (2f_Y)^{-1} (I(\mu_{Z|Y})f_Y + 2\nabla(\mu_{Z|Y})\nabla^T(f_Y)) \\
= (2f_X)^{-1} (I(\mu_{Z|X})f_X + 2\nabla(\mu_{Z|X})\nabla^T(f_X)).
\]

Thus

\[
[\text{tr } (SA_1)]^2 = \frac{1}{4f_X^2} [\text{tr } (S(I(\mu_{Z|X})f_X + 2\nabla(\mu_{Z|X})\nabla^T(f_X)))]^2.
\]

The one dimensional version of this quantity is the squared bias found by Wand and Jones [1995] (Section 5.4, Equation 5.12 on page 125) in the error free case.

Typically with smoothing methods there is a bias–variance tradeoff where larger values of the smoothing parameter, in this case \(S\), introduce more bias but reduce the variance. While the term \([\text{tr } (SA_1)]^2 \geq 0 \forall S\), the existence of the \(n^{-1}\text{tr } (SA_2)\) term complicates this interpretation because it appears possible that certain values of \(S\) could actually reduce the squared bias. We see later (Equation (3.13)) that the \(n^{-1}\text{tr } (SA_2)\) term has the same form as terms in the asymptotic variance expansion.

### 3.4.2 Variance

**Theorem 3.4.** Under Assumptions D and E

\[
\text{Var}(\tilde{\mu}_{Z|Y}) = \frac{1}{n} V + \frac{1}{n} \text{tr}(S(A_2 + A_3 + A_4)) + Cn^{-2} + O(n^{-5/2}).
\]

See Subsection 3.8.4 for a proof of this theorem. The \(n^{-1}V\) term is the only quantity in either the squared bias or variance of order \(n^{-1}\). We can lower bound the constant:

\[
V = f_Y^{-2}E[\nu_y(X_1)f_x^2(y - X_1)] \\
= f_Y^{-2}E[((\mu_{Z|X}(X_1) - \mu_{Z|Y}(y))^2 + \sigma_\epsilon^2) f_x^2(y - X_1)] \\
\geq f_Y^{-2}\sigma_\epsilon^2 \int f_x^2(y - x)f_X(x)dx.
\]

When the error density \(f_\epsilon\) is highly concentrated around 0, this lower bound is large (assuming \(f_X(y)\) is positive). For example if \(f_\epsilon\) is uniform on \([-c/2, c/2]^p\) and \(f_X\) is constant in \([y - c/2, y + c/2]^p\), then \(\int f_x^2(y - x)f_X(x)dx = f_X(y)\int f_x^2(y - x)dx\) and \(f_Y(y) = f_X(y)\). Thus the lower bound becomes

\[
V \geq f_Y^{-2}\sigma_\epsilon^2 \int f_x^2(y - x)f_X(x)dx \\
= f_Y^{-2}(y)\sigma_\epsilon^2 f_X(y)\int f_x^2(y - x)dx \\
= f_X^{-1}(y)\sigma_\epsilon^2 c^{-p}.
\]

(3.12)
As $c$ gets smaller (i.e., the error term $f_\epsilon$ becomes more concentrated), this lower bound increases.

We can see how highly concentrated error distributions cause high variance by directly examining the kernel free estimator. For the uniform error case described, the kernel free estimator (Equation (3.2)) is

$$\hat{\mu}_{Z|Y}(y) = \begin{cases} \frac{\sum_{i=1}^{n} Z_i}{\sum_{i=1}^{n} 1} & : \sum_{i=1}^{n} 1 \{ ||y - X_i||_{\infty} \leq c/2 \} > 0 \\ 0 & : \text{otherwise} \end{cases}$$

When $c$ is close to 0 (i.e., $f_\epsilon$ is concentrated around 0) this estimator averages the responses, $Z_i$, of only a small fraction of the $n$ observations because $1 \{ ||y - X_i||_{\infty} \leq c/2 \}$ will be 0 for most $X_i$.

We discuss procedures for choosing $H$ to minimize $MSE(\hat{\mu}_{Z|Y})$ in Section 3.5. In Section 3.6 we present an example where $A_1 = A_2 = A_3 = A_4 = 0$ and smoothing does not affect the $MSE$ at the $n^{-2}$ order.

### 3.5 Selection of Optimal Bandwidth

Summing the squared bias and variance from Theorems 3.3 and 3.4 we have

$$MSE(\hat{\mu}_{Z|Y}(y)) = \frac{1}{n} V + \frac{1}{n} [\text{tr}(SA_1)]^2 + \frac{1}{n} \text{tr}(S(2A_2 + A_3 + A_4)) + Cn^{-2} + O(n^{-5/2}), \quad (3.13)$$

where $C$ is some quantity that does not depend on $H$ or $n$. We would like to select a bandwidth matrix $H$ (recall $S = H^T \Sigma_K H$) that minimizes this quantity. We study how to select $H$ from some subset, denoted $\mathcal{H}$, of positive definite matrices that minimizes the lower order terms. We consider three choices for $\mathcal{H}$: all positive definite matrices, all diagonal positive definite matrices, and all matrices that equal the identity scaled by a positive constant. Noting that $n^{-1} V$ and $Cn^{-2}$ do not depend on $H$, and defining

$$A = 2A_2 + A_3 + A_4,$$

it is sufficient to determine

$$H^* = \arg\min_{H \in \mathcal{H}} [\text{tr}(SA_1)]^2 + \frac{1}{n} \text{tr}(SA). \quad (3.14)$$

We note that in Equation (3.14) $[\text{tr}(SA_1)]^2 \geq 0$. Thus when $A = 0$, the objective function is bounded below by 0. In such cases, by choosing an $H \in \mathcal{H}$ with sufficiently small norm, one can make the objective function arbitrarily close to 0. However for all $\mathcal{H}$ we consider, $0 \notin \mathcal{H}$. This is because the theorems in Sections 3.3 and 3.4 assumed $H \succ 0$ (see Assumption 3.6). In Section 3.6 we present an example where $A = 0$ and $A_1 = 0$ and discuss some consequences for the $MSE$ and selection of the bandwidth $H$. 
3.5.1 Full Bandwidth Matrix

Here $H$ is the set of all positive definite matrices. Since $\Sigma_K \succ 0$ and $S = H^T \Sigma_K H$, one can solve

$$S^* = \arg\min_{\{S \succ 0\}} \left[ \frac{1}{n} \text{tr}(SA_1)^2 + \frac{1}{n} \text{tr}(SA) \right]$$

and then let $H^* = S^{1/2} \Sigma_K^{-1/2}$. Note that for any compatible matrices $G, F$, $\text{tr}(G^T F) = \text{vec}(G)^T \text{vec}(F)$ (see Henderson and Searle [1979] p.67 Equation 8). Using vec notation the optimization problem becomes

$$S^* = \arg\min_{\{S \succ 0\}} \text{vec}(S)^T \text{vec}(A_1) \text{vec}(A_1)^T \text{vec}(S) + \frac{1}{n} \text{vec}(S)^T \text{vec}(A).$$

The objective function is quadratic in vec$(S)$. However the constraint that $S \succ 0$ may make the problem difficult to optimize using standard quadratic solvers.

3.5.2 Diagonal Bandwidth Matrix and $\Sigma_K = I_{dp}$

Let $I_{dp}$ be the $p \times p$ identity matrix. We assume the kernel $K$ has covariance matrix $I_{dp}$ and the bandwidth matrix $H$ is diagonal with positive entries. This implies $S = H^T \Sigma_K H = H^T H$ is a diagonal matrix. Define $h_S = \text{diag}(H^T H) \in \mathbb{R}^p$ where diag$(M)$ is a column vector of the diagonal elements of the matrix $M$. Note that for any matrix $B$

$$\text{tr}(SB) = \text{diag}(S)^T \text{diag}(B) = \text{diag}(H^T H)^T \text{diag}(B) = h_S^T \text{diag}(B).$$

Thus, the optimization problem in Equation (3.14) becomes

$$h_S^* = \arg\min_{h_S > 0} h_S^T \text{diag}(A_1) \text{diag}(A_1)^T h_S + \frac{1}{n} h_S^T \text{diag}(A).$$

The domain restriction $h_S > 0$ is understood to be elementwise on the vector $h_S$. The objective function in this optimization problem is quadratic in $h_S$ and the domain is restricted to the first orthant i.e., $h_S > 0$.

3.5.3 Scalar Bandwidth Matrix

The simplest optimization strategy is to restrict $H$ to be the identity times a positive constant. Let $H = hI_p$ where $h \in \mathbb{R}^+$. The optimization problem in Equation (3.14) becomes

$$h^* = \arg\min_{\{h > 0\}} h^4 \text{tr}(\Sigma_K A_1)^2 + \frac{h^2}{n} \text{tr}(\Sigma_K A).$$
Under the assumptions that $\text{tr}(\Sigma K A) < 0$ and $\text{tr}(\Sigma K A_1)^2 \neq 0$ we can solve this optimization problem explicitly and obtain

$$h^* = \sqrt{-\text{tr}(\Sigma K A) \over 2n\text{tr}(\Sigma K A_1)^2}.$$ (3.16)

The optimal amount of smoothing is of order $n^{-1/2}$. This rate matches the bandwidth rate for kernel density estimation with Berkson error (see Chapter 2, specifically Subsection 2.3.2). This rate does not depend on the dimension of the problem, $p$, unlike the Nadaraya–Watson estimator in the standard, error-free case.

The assumption $\text{tr}(\Sigma K A) < 0$ and $\text{tr}(\Sigma K A_1)^2 \neq 0$ are not satisfied in all cases. In the following section we present an example where $A_1 = 0$ and $A = 0$.

### 3.6 Asymptotic MSE May Be Invariant to Smoothing: Example

We present an example where $A_1 = A_2 = A_3 = A_4 = 0$. We then discuss some implications for the choice of the bandwidth matrix $H$ and the $MSE$.

**Example 3.1.** Let $f_X$ and $f_\epsilon$ be any densities such that for some constant $D > 0$

$$f_\epsilon(y - x) f_X(x) = D f_X(x).$$

In other words, $f_\epsilon(y - \cdot)$ equals a positive constant on the support of $f_X(\cdot)$. Then $A_1 = A_2 = A_3 = A_4 = 0$.

**Proof.** We first establish some identities. Note that $f_\epsilon(y - X_1) f_X(X_1) = D f_X(X_1)$ implies

$$\nabla(f_\epsilon)(y - x) f_X(x) = 0,$$

(3.17)

$$\mathcal{I}(f_\epsilon)(y - x) f_X(x) = 0.$$ (3.18)

Noting that the gradient (Hessian) of the convolution of $f_\epsilon$ and $f_X$ is the convolution of the gradient (Hessian) of $f_\epsilon$ and $f_X$ and using Equations (3.17) and (3.18) we have

$$\nabla(f_Y)(y) = \nabla \left( \int f_\epsilon(y - x) f_X(x) dx \right) = \int \nabla(f_\epsilon)(y - x) f_X(x) dx = 0,$$

(3.19)

and

$$\mathcal{I}(f_Y)(y) = \mathcal{I} \left( \int f_\epsilon(y - x) f_X(x) dx \right) = \int \mathcal{I}(f_\epsilon)(y - x) f_X(x) dx = 0.$$ (3.20)
• \(A_1 = (2f_Y)^{-1}(I(\mu_{Z|Y})f_Y + 2\nabla(\mu_{Z|Y})\nabla^T(f_Y))\). By Equation (3.19) we have \(\nabla^T(f_Y) = 0\). Thus in order to show \(A_1 = 0\), it is sufficient to show \(I(\mu_{Z|Y}) = 0\). By Lemma 3.18 on p.87 we have \(\mu_{Z|Y}(y)f_Y(y) = \mathbb{E}[\mu_Z(X_1)f_\epsilon(y - X_1)]\). Thus

\[
I(\mu_{Z|Y})(y) = I(f_Y(y)^{-1}\mathbb{E}[\mu_Z(X_1)f_\epsilon(y - X_1)])
\]

\[
= I\left(\int \mu_{Z|X}(x)f_\epsilon(y - x)\frac{f_X(x)}{f_Y(y)}dx\right)
\]

\[
= \int \mu_{Z|X}(x)I\left(\frac{f_\epsilon(y - x)f_X(x)}{f_Y(y)}\right)dx
\]

By Equation (3.18) \(I(f_\epsilon(y - x)f_X(x)) = 0\) and by Equation (3.20) \(I(f_Y(y)) = 0\). Thus

\[
I\left(\frac{f_\epsilon(y - x)f_X(x)}{f_Y(y)}\right) = 0.
\]

Hence \(I(\mu_{Z|Y})(y) = 0\).

• \(A_2 = -2f_Y^{-2}\mathbb{E}[(\mu_{Z|X}(X_1) - \mu_{Z|Y}(y))f_\epsilon^2(y - X_1)]A_1\). Since \(A_1 = 0\), \(A_2 = 0\).

• \(A_3 = -f_Y^{-3}\mathbb{E}[f_\epsilon^2(y - X_1)\nu_y(X_1)]I(f_Y)(y)\). In Equation (3.20) we showed \(I(f_Y)(y) = 0\). Thus \(A_3 = 0\).

• \(A_4 = f_Y^{-2}\mathbb{E}[I(f_\epsilon)(y - X_1)f_\epsilon(y - X_1)\nu_y(X_1)]\). As shown in Equation 3.18, \(I(f_\epsilon)(y - X_1)f_X(X_1) = 0\). Therefore \(A_4 = 0\).

For this example, the asymptotic expansions for the squared bias and variance (see Theorems 3.3 and 3.4) become

\[
Bias(\tilde{\mu}_{Z|Y})^2 = Cn^{-2} + O(n^{-5/2}),
\]

\[
\text{Var}(\tilde{\mu}_{Z|Y}) = \frac{1}{n}V + Cn^{-2} + O(n^{-5/2}),
\]

where \(C\) is a constant that does not depend on \(H\) or \(n\) and may change value at each appearance. Here the \(MSE\) is invariant to smoothing at the \(n^{-2}\) order. In this case, the present asymptotic expansions at the \(n^{-2}\) order do not provide guidance for selecting the bandwidth \(H\). Note that the example is general in that it does not impose any conditions on the underlying regression function \(\mu_{Z|X}\).

The cause of this phenomenon appears to be that the kernel free estimator already has low variance, implying that there is little benefit to further regularization. In particular, since \(f_\epsilon(y - X_1)f_X(X_1) = Df_X(X_1)\), the kernel free estimator (Equation (3.2)) is

\[
\tilde{\mu}_{Z|Y}(y) = \frac{\sum_{i=1}^{n}Z_i f_\epsilon(y - X_i)}{\sum_{i=1}^{n} f_\epsilon(y - X_i)} = \frac{1}{n} \sum_{i=1}^{n} Z_i.
\]
CHAPTER 3. NADARAYA–WATSON FOR PREDICTION WITH FEATURE MEASUREMENT ERROR

3.7 Conclusions

We have derived approximations to the $MSE$ as a function of $H$ for the Nadaraya–Watson estimator adapted to the errors in features setting. Up to order $n^{-2}$, the $MSE$ depends on the kernel only through the variance of the kernel $S = H^T K_H H$. This is unlike the error free case where the variance is a function of the squared density of the kernel (e.g., see Wand and Jones [1995], Section 5.4, Equation 5.13 on page 125). In the scalar bandwidth case, the optimal rate of convergence for the bandwidth is $n^{-1/2}$.

Our results show that the smoothing matrix that minimizes the $MSE$ is a function of the density of the error in the features, $f_\epsilon$ (see Equation 3.16). In many practical problems, such as redshift estimation discussed in Section 3.1, there are many observations with estimated feature, $Y_1, \ldots, Y_m$, and unknown responses. Thus there will be a different optimal bandwidth for each of these observations. These leads to computational difficulties as the bandwidth is often chosen by computationally expensive methods such as cross-validation. Further understanding of the relationship between the error in the features of an observation, $f_\epsilon$, and the optimal bandwidth, could assist in developing fast procedures for selecting a bandwidth that is optimal, or near optimal, for a particular error distribution.

3.8 Theorems and Lemmas

3.8.1 Proof of Theorem 3.1 (p.51)

We must show

$$E[(\tilde{Z} - Z)^2|Y] = E[(\tilde{Z} - \mu_{Z|Y}(Y))^2|Y] + \text{Var}(Z|Y).$$

Recall that $\tilde{Z} = g(\{(X_i, Z_i)\}_{i=1}^n, Y)$ for some deterministic function $g$, $Z = \mu_{Z|X}(X) + \delta$, and $Y = X + \epsilon$. Thus by construction of the random variables in Section 3.2, $\tilde{Z} | Y \perp \perp Z | Y$. Noting that $E[Z - E[Z|Y]|Y] = 0$ and $\mu_{Z|Y} \equiv E[Z|Y]$, we have

$$E[(\tilde{Z} - Z)^2|Y] = E[((\tilde{Z} - E[Z|Y]) - (Z - E[Z|Y]))^2|Y]$$


$$+ E[(Z - E[Z|Y])^2|Y]$$

$$= E[(\tilde{Z} - \mu_{Z|Y}(Y))^2|Y] + \text{Var}(Z|Y).$$
3.8.2 Proof of Theorem 3.2 (p.54)

We must show

\[ \text{Bias}(\hat{\mu}_{Z|Y}(y)) = \mathbb{E}[\omega_n] + O(n^{-3/2}), \]  

\[ \text{Var}(\hat{\mu}_{Z|Y}(y)) = \text{Var}(\omega_n) + 2\text{Cov}(\omega_n, s_n) + O(n^{-5/2}), \]  

where

\[ \omega_n = f_Y^{-1}\Delta_B - f_Y^{-1}\mu_{Z|Y}\Delta_A - f_Y^{-2}\Delta_A\Delta_B + f_Y^{-2}\Delta_A^2\mu_{Z|Y}, \]

\[ s_n = \Delta_B\Delta_A^2f_Y^{-3} - \Delta_A^3f_Y^{-3}\mu_{Z|Y}, \]

\[ \Delta_A = \frac{1}{n} \sum_{i=1}^{n} \tilde{f}_{Y,X_i}(y) - f_Y(y), \]

\[ \Delta_B = \frac{1}{n} \sum_{i=1}^{n} Z_i \tilde{f}_{Y,X_i}(y) - f_Y(y)\mu_{Z|Y}(y). \]

For notational simplicity, we drop the dependence on \( y \) in \( \hat{\mu}_{Z|Y}(y) \) and all other functions. The initial steps here follow the proof of Theorem 4.1 in Appendix A in Carroll et al. [2009]. The main element of this proof is a Taylor approximation of \( \hat{\mu}_{Z|Y} \) on the set \( \mathcal{E} = \{ |\Delta_A| \leq \frac{1}{2}f_Y \} \).

We note that by Lemma 3.4 on p.75 for any \( k \in \mathbb{Z}^+ \), \( \mathbb{E}[|\Delta_A|^k] = O(n^{-k/2}) \). Thus using Chebyshev’s inequality for any \( k \) we have

\[ \mathbb{P}(\mathcal{E}^C) = \mathbb{P}(|\Delta_A| > \frac{1}{2}f_Y) \leq \frac{2^k\mathbb{E}[|\Delta_A|^k]}{f_Y^k} = O(n^{-k/2}). \]

In other words, we can bound \( \mathbb{P}(\mathcal{E}^C) \) at any power of \( n \). Let

\[ R = \frac{f_Y^{-5}\Delta_A^4}{1 + f_Y^{-1}\Delta_A}. \]

On \( \mathcal{E} \), \( |\Delta_Af_Y^{-1}| \leq 1/2 \). Using the geometric series expansion, we have

\[ \frac{1}{f_Y + \Delta_A} \mathbb{I}_\mathcal{E} = \frac{1}{f_Y(1 + \Delta_Af_Y^{-1})} \mathbb{I}_\mathcal{E} = f_Y^{-1}\left(1 - \Delta_Af_Y^{-1} + \Delta_A^2f_Y^{-2} - \Delta_A^3f_Y^{-3} + \frac{f_Y^{-4}\Delta_A^4}{1 + \Delta_Af_Y^{-1}}\right) \mathbb{I}_\mathcal{E} = (f_Y^{-1} - \Delta_Af_Y^{-2} + \Delta_A^2f_Y^{-3} - \Delta_A^3f_Y^{-4} + R) \mathbb{I}_\mathcal{E}. \]
CHAPTER 3. NADARAYA–WATSON FOR PREDICTION WITH FEATURE MEASUREMENT ERROR

We express $\tilde{\mu}_{Z|Y}$ in terms of $\Delta_A$ and $\Delta_B$ and use the expansion in Equation (3.25) to obtain

$$\tilde{\mu}_{Z|Y} \mathbb{1}_E = (\Delta_B + f_Y \mu_{Z|Y}) \left( \frac{1}{\Delta_A + f_Y} \right) \mathbb{1}_E$$

$$= (\Delta_B + f_Y \mu_{Z|Y}) \left( f_Y^{-1} - \Delta_A f_Y^{-2} + \Delta_A^2 f_Y^{-3} - \Delta_A^3 f_Y^{-4} + R \right) \mathbb{1}_E$$

$$= \left( \mu_{Z|Y} + \frac{f_Y^{-1} \Delta_B - f_Y^{-2} \Delta_A \Delta_B + f_Y^{-3} \Delta_A^2 \mu_{Z|Y}}{\equiv \omega_n} \right) \mathbb{1}_E$$

$$\equiv \nu_n$$

$$+ \Delta_B \Delta_A^2 f_Y^{-3} - \Delta_A^3 f_Y^{-3} \mu_{Z|Y} - \Delta_A^3 \Delta_B f_Y^{-4} + R(\Delta_B + f_Y \mu_{Z|Y}) \mathbb{1}_E$$

(3.26)

To summarize we have

$$\tilde{\mu}_{Z|Y} \mathbb{1}_E = (\mu_{Z|Y} + \omega_n + \nu_n) \mathbb{1}_E.$$  (3.28)

The remainder of the proof is divided into two parts in which we show the bias expansion (Equation (3.21)) and the variance expansion (Equation (3.22)) hold.

1. **Bias:** By Lemma 3.1 on p.68 (with $k = 1$) $E[\tilde{\mu}_{Z|Y}] = E[\tilde{\mu}_{Z|Y} \mathbb{1}_E] + O(P(E^C))$. Using this fact and the expansion of $\tilde{\mu}_{Z|Y} \mathbb{1}_E$ in Equation (3.28), we have

$$E[\tilde{\mu}_{Z|Y}] = E[\tilde{\mu}_{Z|Y} \mathbb{1}_E] + O(P(E^C))$$

$$= E[\mu_{Z|Y} \mathbb{1}_E] + E[\omega_n \mathbb{1}_E] + E[\nu_n \mathbb{1}_E] + O(P(E^C))$$

$$= \mu_{Z|Y} E[\mathbb{1}_E] + E[\omega_n \mathbb{1}_E] + E[\nu_n \mathbb{1}_E] + O(P(E^C))$$

(3.29)

We now show that $E[\nu_n \mathbb{1}_E]$ is of high order. First note that $|\Delta_A f_Y^{-1}| \leq 1/2$ on $E$, thus $(1 + \Delta_A f_Y^{-1})^{-1} \mathbb{1}_E \leq 2$. Also note that by Lemma 3.4 on p.75, for any $l \in \{0, 1, 2\}$, $E[|\Delta_A^l \Delta_B|] = O(n^{-(l+1)/2})$. Using the definitions of $\nu_n$ (Equation (3.27)) and $R$ (Equation (3.24)), we have

$$E[|\nu_n \mathbb{1}_E|] = E[|\Delta_B \Delta_A^2 \Delta_Y^{-3} - \Delta_A^3 f_Y^{-3} \mu_{Z|Y} - \Delta_A^3 \Delta_B f_Y^{-4} + R(\Delta_B + f_Y \mu_{Z|Y})| \mathbb{1}_E]$$

$$\leq E[|\Delta_B \Delta_A^2 \Delta_Y^{-3} - \Delta_A^3 f_Y^{-3} \mu_{Z|Y} + |\Delta_A^3 \Delta_B f_Y^{-4}]]$$

$$+ E[|\frac{f_Y^{-5} \Delta_A^4}{1 + f_Y^{-1} \Delta_A} (\Delta_B + f_Y \mu_{Z|Y})| \mathbb{1}_E]$$

$$\leq E[|2 f_Y^{-5} \Delta_A^4 (\Delta_B + f_Y \mu_{Z|Y})|] + O(n^{-3/2})$$

$$= O(n^{-3/2}).$$  (3.30)

We now remove the $\mathbb{1}_E$ from $E[\omega_n \mathbb{1}_E]$ in Equation (3.29). Note $E[\omega_n \mathbb{1}_E] = E[\omega_n] - E[\omega_n \mathbb{1}_E]$. Using the Cauchy–Schwartz inequality we have $E[\omega_n \mathbb{1}_E] \leq (E[\omega_n^2] P(E^C))^{1/2} = O(n^{-3/2})$. Thus

$$E[\omega_n \mathbb{1}_E] = E[\omega_n] + O(n^{-3/2}).$$  (3.31)
Substituting Equations (3.31) and (3.30) into Equation (3.29) we have
\[
\mathbb{E}[\tilde{\mu}_Z|Y] = \mu_Z|Y + \mathbb{E}[\omega_n] + O(n^{-3/2}).
\]
Thus
\[
\text{Bias}(\tilde{\mu}_Z|Y(y)) = \mathbb{E}[\omega_n] + O(n^{-3/2}).
\]

2. Variance: From Lemma 3.1 on p.68 we have \( \mathbb{E}[\tilde{\mu}_Z^2|Y] = \mathbb{E}[\tilde{\mu}_Z|Y \mathbb{1}_\mathcal{C}] + O(\mathbb{P}(\mathcal{E}^C)), \mathbb{E}[\tilde{\mu}_Z|Y] = \mathbb{E}[\tilde{\mu}_Z|Y \mathbb{1}_\mathcal{C}] + O(\mathbb{P}(\mathcal{E}^C)), \) and \( \mathbb{E}[\tilde{\mu}_Z|Y] = O(\mathbb{P}(\mathcal{E})) = O(1) \). Using these three equalities, we have
\[
\text{Var}(\tilde{\mu}_Z|Y) = \mathbb{E}[\tilde{\mu}_Z^2|Y] - \mathbb{E}[\tilde{\mu}_Z|Y]^2
= \mathbb{E}[\tilde{\mu}_Z|Y \mathbb{1}_\mathcal{C}] + O(\mathbb{P}(\mathcal{E}^C)) - (\mathbb{E}[\tilde{\mu}_Z|Y \mathbb{1}_\mathcal{C}] + O(\mathbb{P}(\mathcal{E}^C)))^2
= \mathbb{E}[\tilde{\mu}_Z^2|Y \mathbb{1}_\mathcal{C}] - \mathbb{E}[\tilde{\mu}_Z|Y \mathbb{1}_\mathcal{C}]^2 + O(\mathbb{P}(\mathcal{E}^C))
= \text{Var}(\tilde{\mu}_Z|Y \mathbb{1}_\mathcal{C}) + O(\mathbb{P}(\mathcal{E}^C)). \tag{3.32}
\]

We separate \( \nu_n \) into components. Recalling the definition of \( \nu_n \) from Equation (3.27), let
\[
\nu_n = \underbrace{\Delta_B \Delta^2_A f_Y^{-3}}_{\equiv s_n} - \underbrace{\Delta^3_A \Delta B f_Y^{-4} + R(\Delta_B + f_Y \mu_Z|Y)}_{\equiv \Delta_4}.
\tag{3.33}
\]

Using the expansion of \( \tilde{\mu}_Z|Y \mathbb{1}_\mathcal{C} \) in Equation (3.28), building from Equation (3.32), we have
\[
\text{Var}(\tilde{\mu}_Z|Y) = \text{Var}((\mu_Z|Y + \omega_n + \nu_n) \mathbb{1}_\mathcal{C}) + O(\mathbb{P}(\mathcal{E}^C))
= \text{Var}((\mu_Z|Y + \omega_n + s_n + \Delta_4 \mathbb{1}_\mathcal{C}) + O(\mathbb{P}(\mathcal{E}^C))
= \text{Var}(\mu_Z|Y + \omega_n + s_n + \Delta_4 \mathbb{1}_\mathcal{C} - (\mu_Z|Y + \omega_n + s_n) \mathbb{1}_\mathcal{C}^c) + O(\mathbb{P}(\mathcal{E}^C))
= \text{Var}(\omega_n + s_n) + \text{Var}(\Delta_4 \mathbb{1}_\mathcal{C} - (\mu_Z|Y + \omega_n + s_n) \mathbb{1}_\mathcal{C}^c)
+ 2\text{Cov}(\omega_n + s_n, \Delta_4 \mathbb{1}_\mathcal{C} - (\mu_Z|Y + \omega_n + s_n) \mathbb{1}_\mathcal{C}^c) + O(\mathbb{P}(\mathcal{E}^C)). \tag{3.34}
\]

Below (see **Bounds**) we show \( \text{Var}(\Delta_4 \mathbb{1}_\mathcal{C} - (\mu_Z|Y + \omega_n + s_n) \mathbb{1}_\mathcal{C}^c) \) in Equation (3.34) and the covariance term in Equation (3.35) are \( O(n^{-5/2}) \). Using these facts and the fact that \( O(\mathbb{P}(\mathcal{E}^C)) = O(n^{-5/2}) \), we have
\[
\text{Var}(\tilde{\mu}_Z|Y) = \text{Var}(\omega_n + s_n) + O(n^{-5/2})
= \text{Var}(\omega_n) + \text{Var}(s_n) + 2\text{Cov}(\omega_n, s_n) + O(n^{-5/2}).
\]

Noting that the terms in \( s_n \) (see Equation 3.33) are of the form \( \Delta^j_A \Delta^k_B \) for \( j + k \geq 3 \) we have by Lemma 3.4 on p.75 that \( \text{Var}(s_n) \leq \mathbb{E}[s_n^2] = O(n^{-3}) \). Using this bound we have the result
\[
\text{Var}(\tilde{\mu}_Z|Y) = \text{Var}(\omega_n) + 2\text{Cov}(\omega_n, s_n) + O(n^{-5/2}).
\]
Bounds: Now we show \( \text{Var}(\Delta_4 \mathbb{I}_c - (\mu_{Z|Y} + \omega_n + s_n) \mathbb{I}_{c^C}) \) in Equation (3.34) and the covariance term in Equation (3.35) are \( O(n^{-5/2}) \). First note that the terms in \( \Delta_4 \) (see Equation (3.33)) are of the form \( \Delta_A^j \Delta_B^k \) for \( j + k \geq 4 \). Thus by Lemma 3.4 on p.75 \( \mathbb{E}[\Delta_4^2 \mathbb{I}_c] = O(n^{-4}) \). Second note that \( \omega_n \) and \( s_n \) are affine in \( \Delta_B \). Therefore \( (\mu_{Z|Y} + \omega_n + s_n)^4 \) contains powers of \( \Delta_B \) no higher than 4 and thus \( \mathbb{E}[(\mu_{Z|Y} + \omega_n + s_n)^4] = O(1) \) by Lemma 3.4 on p.75. Thus

\[
\text{Var}(\Delta_4 \mathbb{I}_c - (\mu_{Z|Y} + \omega_n + s_n) \mathbb{I}_{c^C}) = O(n^{-4}).
\]

Since \( \text{Var}(W) \leq \mathbb{E}[W^2] \) we have by Equation (3.36)

\[
\text{Var}(\Delta_4 \mathbb{I}_c - (\mu_{Z|Y} + \omega_n + s_n) \mathbb{I}_{c^C}) = O(n^{-4}).
\]

Thus all that is left to show is that the covariance term in Equation (3.35) is \( O(n^{-5/2}) \). Note that terms in \( \omega_n + s_n \) are of the form \( \Delta_A^j \Delta_B^k \) for \( j + k \geq 1 \). Therefore \( \mathbb{E}[(\omega_n + s_n)^2] = O(n^{-1}) \) by Lemma 3.4 on p.75. Using the Cauchy–Schwartz inequality and Equation (3.36) we have

\[
\text{Cov}(\omega_n + s_n, \Delta_4 \mathbb{I}_c - (\mu_{Z|Y} + \omega_n + s_n) \mathbb{I}_{c^C}) \\
\leq \sqrt{\mathbb{E}[(\omega_n + s_n)^2] \mathbb{E}[(\Delta_4 \mathbb{I}_c - (\mu_{Z|Y} + \omega_n + s_n) \mathbb{I}_{c^C})^2]} \\
= O(n^{-1}) O(n^{-4}) \\
= O(n^{-5/2}).
\]

3.8.3 Proof of Theorem 3.3 (p.55)

By Theorem 3.2

\[
\text{Bias}(\tilde{\mu}_{Z|Y}) = \mathbb{E}[(\omega_n)] + O(n^{-3/2}).
\] (3.37)

We now determine \( \mathbb{E}[(\omega_n)] \) using asymptotic expansions. By Lemma 3.7 on p.79 we have

\[
\omega_n = f_Y^{-2} \left( \frac{1}{n} \sum_{i=1}^{n} (Z_i - \mu_{Z|Y}) \tilde{f}_{Y;i} \right) \left( 2 f_Y - \frac{1}{n} \sum_{i=1}^{n} \tilde{f}_{Y;i} \right).
\]
Define $\tau_1 = (Z_1 - \mu_{Z|Y})\tilde{f}_{Y,X_1}$. By symmetry and the fact that $\tau_1 \perp \tilde{f}_{Y,X_i}$ for $i \neq 1$, we have

$$
\mathbb{E}[\omega_n] = f_Y^{-2} \mathbb{E} \left[ \left( \frac{1}{n} \sum_{i=1}^{n} (Z_i - \mu_{Z|Y})\tilde{f}_{Y,X_i} \right) \left( 2f_Y - \frac{1}{n} \sum_{i=1}^{n} \tilde{f}_{Y,X_i} \right) \right]
$$

$$
= f_Y^{-2} \mathbb{E} \left[ (Z_1 - \mu_{Z|Y})\tilde{f}_{Y,X_1} \left( 2f_Y - \frac{1}{n} \sum_{i=1}^{n} \tilde{f}_{Y,X_i} \right) \right]
$$

$$
= \frac{2\mathbb{E}[\tau_1]}{f_Y} - \frac{1}{nf_Y^2} \sum_{i=1}^{n} \mathbb{E} \left[ \tau_1 \tilde{f}_{Y,X_i} \right]
$$

$$
= \frac{2\mathbb{E}[\tau_1]}{f_Y} - \frac{\mathbb{E}[\tau_1]\mathbb{E}[\tilde{f}_{Y,X_1}]}{f_Y^2} + \frac{1}{nf_Y^2} \left( \mathbb{E}[\tau_1]\mathbb{E}[\tilde{f}_{Y,X_1}] - \mathbb{E}[\tau_1\tilde{f}_{Y,X_1}] \right). \quad (3.38)
$$

By Lemma 3.16 on p.85, $\tilde{f}_{Y,X_1}(y) = f_\epsilon(y - X_1) + T(||H||_\infty^2)$ for a random variable $T$ that is bounded uniformly in $H$. Therefore $\mathbb{E}[\tilde{f}_{Y,X_1}] = f_Y + O(||H||_\infty^2) = f_Y + O(n^{-1})$. By Lemma 3.15 on p.84 $\mathbb{E}[\tau_1] = O(||H||_\infty^2) = O(n^{-1})$. Thus

$$
\frac{\mathbb{E}[\tau_1]\mathbb{E}[\tilde{f}_{Y,X_1}]}{f_Y^2} = \frac{\mathbb{E}[\tau_1]}{f_Y} + O(||H||_\infty^2)\mathbb{E}[\tau_1] = \frac{\mathbb{E}[\tau_1]}{f_Y} + O(n^{-2}) \text{ and}
$$

$$
\frac{\mathbb{E}[\tau_1]\mathbb{E}[\tilde{f}_{Y,X_1}]}{nf_Y^2} = O(n^{-2}).
$$

Substituting these equations into (3.38) we have

$$
\mathbb{E}[\omega_n] = \frac{\mathbb{E}[\tau_1]}{f_Y} - \frac{1}{nf_Y^2} \mathbb{E}[\tau_1\tilde{f}_{Y,X_1}] + O(n^{-2}). \quad (3.39)
$$

By Lemma 3.16 on p.85 (with $j = 2$), $\tilde{f}_{Y,X_1}(y) = f_\epsilon^2(y - X_1) + T||H||_\infty^2$ for a random variable $T$ that is bounded uniformly in $H$ and independent of $\delta_1$. Thus we have

$$
\mathbb{E}[\tau_1\tilde{f}_{Y,X_1}] = \mathbb{E}[(Z_1 - \mu_{Z|Y})f_\epsilon^2_{Y,X_1}]
$$

$$
= \mathbb{E}[(\mu_{Z|X}(X_1) + \delta_1 - \mu_{Z|Y})f_\epsilon^2(y - X_1) + T||H||_\infty^2)]
$$

$$
= \mathbb{E}[(\mu_{Z|X}(X_1) - \mu_{Z|Y})f_\epsilon^2(y - X_1)] + O(||H||_\infty^2). \quad (3.40)
$$

Substituting Equation (3.40) into Equation (3.39) and noting that $||H||_\infty^2 = O(n^{-1})$, we have

$$
\mathbb{E}[\omega_n] = \frac{\mathbb{E}[\tau_1]}{f_Y} - \frac{1}{nf_Y^2} \mathbb{E}[(\mu_{Z|X}(X_1) - \mu_{Z|Y})f_\epsilon^2(y - X_1)] + O(n^{-2}).
$$

Using the expansion of $\mathbb{E}[\tau_1]$ in Lemma 3.15 on p.84, we have

$$
\mathbb{E}[\omega_n] = \frac{1}{2f_Y} \text{tr} \left( H^T \Sigma_K H \left( I(\mu_{Z|Y})f_Y + 2\nabla(\mu_{Z|Y})\nabla^T(f_Y) \right) \right)
$$

$$
- \frac{1}{nf_Y^2} \mathbb{E}[(\mu_{Z|X}(X_1) - \mu_{Z|Y})f_\epsilon^2(y - X_1)] + O(n^{-2}). \quad (3.41)
$$
CHAPTER 3. NADARAYA–WATSON FOR PREDICTION WITH FEATURE MEASUREMENT ERROR

Plugging this equation into Equation (3.37) we obtain

$$
\text{Bias}(\mu_{Z|Y}) = \frac{1}{2f_Y} \text{tr} \left( H^T \Sigma_K H \left( \mathcal{I}(\mu_{Z|Y}) f_Y + 2 \nabla(\mu_{Z|Y}) \nabla^T(f_Y) \right) \right) 
- \frac{1}{nf_Y^2} \mathbb{E}[\{\mu_{Z|X}(X_1) - \mu_{Z|Y}\} f^2_\epsilon(y - X_1)] + O(n^{-3/2}).
$$

Denote

$$
A_1 = \frac{1}{2f_Y} \left( \mathcal{I}(\mu_{Z|Y}) f_Y + 2 \nabla(\mu_{Z|Y}) \nabla^T(f_Y) \right),
$$

$$
A_2 = -\frac{2}{f_Y^2} \mathbb{E}[\{\mu_{Z|X}(X_1) - \mu_{Z|Y}\} f^2_\epsilon(y - X_1)] A_1. \tag{3.42}
$$

Squaring both sides and denoting $S = H^T \Sigma_K H$ we obtain

$$
\text{Bias}(\mu_{Z|Y})^2 = [\text{tr} (SA_1)]^2 + \frac{1}{n} \text{tr} (SA_2) + Cn^{-2} + O(n^{-5/2}),
$$

where $C$ is a constant that does not depend on $H$ or $n$.

3.8.4 Proof of Theorem 3.4

Let $C$ be a constant that does not depend on $H$ or $n$. The precise value of $C$ may change at each appearance. From Theorem 3.2 we have

$$
\text{Var}(\mu_{Z|Y}) = \text{Var}(\omega_n) + 2\text{Cov}(\omega_n, s_n) + O(n^{-5/2}).
$$

In Lemma 3.3 on p.74 we show

$$
2\text{Cov}(\omega_n, s_n) = Cn^{-2} + O(n^{-5/2}).
$$

Using Lemma 3.2 on p.69 to expand $\text{Var}(\omega_n)$ we obtain the result

$$
\text{Var}(\mu_{Z|Y}) = \frac{1}{n} V + \frac{1}{n} \text{tr}(S(A_2 + A_3 + A_4)) + \frac{1}{n^2} C + O(n^{-5/2}),
$$

where

$$
S = H^T \Sigma_K H,
$$

$$
\nu_y(x) = (\mu_{Z|X}(x) - \mu_{Z|Y}(y))^2 + \sigma_\delta^2,
$$

$$
V = f_Y^2 \mathbb{E}[f^2_\epsilon(y - X_1)\nu_y(X_1)],
$$

$$
A_2 = -f_Y^{-3} \mathbb{E}[\{\mu_{Z|X}(X_1) - \mu_{Z|Y}\} f^2_\epsilon(y - X_1)] \left( \mathcal{I}(\mu_{Z|Y}) f_Y + 2 \nabla(\mu_{Z|Y}) \nabla^T(f_Y) \right),
$$

$$
A_3 = -f_Y^{-3} \mathbb{E}[f^2_\epsilon(y - X_1)\nu_y(X_1)] \mathcal{I}(f_Y),
$$

$$
A_4 = f_Y^{-2} \mathbb{E}[\mathcal{I}(f_\epsilon)(y - X_1) f_\epsilon(y - X_1)\nu_y(X_1)].
$$

Note that the definition of $A_2$ here is consistent with the definition of $A_2$ in Equation (3.42).
3.8.5 Lemmas

Lemma 3.1. For $k \in \{1, 2\}$ we have

\[ E[\mu_{Z|Y}(y)^k] = E[\mu_{Z|Y}(y)1_{E}] + O(P(E^C)), \]  
(3.43)
\[ E[\mu_{Z|Y}(y)^k1_{E}] = O(P(E)). \]  
(3.44)

Proof. We show that for $A = E, E^C$

\[ E[|\mu_{Z|Y}(y)|_A^k] = O(P(A)). \]  
(3.45)

This immediately establishes Equation (3.44). Since $1_{E}1_{E^C} = 0$ by Equation (3.45), for $A = E^C$, we have

\[ E[|\mu_{Z|Y}(y)|_A^k] = E[(|\mu_{Z|Y}(y)|_E + |\mu_{Z|Y}(y)|_E^C)^k] \]
\[ = E[|\mu_{Z|Y}(y)|_E^k] + E[|\mu_{Z|Y}(y)|_E^C] \]
\[ = E[|\mu_{Z|Y}(y)|_E^k] + O(P(E^C)), \]

which establishes Equation (3.43).

We now show Equation (3.45) holds. Let

\[ c_i = \frac{f_{Y,X_i}(y)}{\sum_{j=1}^{n} f_{Y,X_j}(y)} \geq 0. \]

By convexity of $f(x) = |x|^k$ we have for any $a, b$, $|(a + b)/2|^k \leq (|a|^k + |b|^k)/2$. Thus $|a + b|^k \leq 2^{k-1}(|a|^k + |b|^k)$. Using this inequality we have

\[ E[|\mu_{Z|Y}(y)|_A^k] = E[|1_A \sum_{i=1}^{n} c_i Z_i|^k] \]
\[ = E[|1_A \sum_{i=1}^{n} c_i (\mu_{Z|X}(X_i) + \delta_i)|^k] \]
\[ \leq 2^{k-1}E[|1_A \sum_{i=1}^{n} c_i \mu_{Z|X}(X_i)|^k] + 2^{k-1}E[|1_A \sum_{i=1}^{n} c_i \delta_i|^k]. \]

Since $\mu_{Z|X}$ is uniformly bounded (Assumption 3.9) and $\sum |c_i| = 1$, we have

\[ E[|1_A \sum_{i=1}^{n} c_i \mu_{Z|X}(X_i)|^k] = O(E[1_A]) = O(P(A)). \]
Recalling that \(1_A\) and \(c_i\) are independent of \(\delta_i\) and for \(k \leq 4, \mathbb{E}[|\delta_1|^k] < \infty\) (Assumption 3.10), we have

\[
\mathbb{E}\left[1_A \sum_{i=1}^n |c_i \delta_i|^k\right] = \mathbb{E}\left[1_A \mathbb{E}\left[\sum_{i=1}^n |c_i \delta_i|^k \{X_j\}_{j=1}^n\right]\right] \\
\leq \mathbb{E}\left[1_A \sum_{i=1}^n |c_i| \mathbb{E}\left[|\delta_i|^k \{X_j\}_{j=1}^n\right]\right] \\
= \mathbb{E}\left[1_A \sum_{i=1}^n |c_i|\right] \mathbb{E}[|\delta_1|^k] \\
= \mathbb{P}(A) \mathbb{E}[|\delta_1|^k] \\
= O(\mathbb{P}(A)).
\]

Hence

\[
\mathbb{E}[|\tilde{\mu}_Z|Y_1 A|^k] = O(\mathbb{P}(A)).
\]

Lemma 3.2.

\[
\text{Var}(\omega_n) = \frac{1}{n} V + \frac{1}{n} \text{tr}(S(A_2 + A_3 + A_4)) + \frac{1}{n^2} C + O(n^{-3}),
\]

where

\[
S = H^T \Sigma_R H, \\
\nu_y(x) = (\mu_{Z|X}(x) - \mu_{Z|Y}(y))^2 + \sigma^2, \\
V = f_Y^{-2} \mathbb{E}[f_e^2(y - X_1) \nu_y(X_1)], \\
A_2 = -f_Y^{-3} \mathbb{E}[(\mu_{Z|X}(x_1) - \mu_{Z|Y}(y))^2 (y - X_1)] \left( \mathcal{I}(\mu_{Z|Y}) f_Y + 2 \nabla(\mu_{Z|Y}) \nabla^T f_Y \right), \\
A_3 = -f_Y^{-3} \mathbb{E}[f_e^2(y - X_1) \nu_y(X_1)] \mathcal{I}(f_Y), \\
A_4 = f_Y^{-2} \mathbb{E}[\mathcal{I}(f_e)(y - X_1) f_e(y - X_1) \nu_y(X_1)].
\]

Proof. Using Lemma 3.7 on p.79 for expressing \(\omega_n\) we have

\[
\text{Var}(\omega_n) = \text{Var}\left[ f_Y^{-2} \left( \frac{1}{n} \sum_{i=1}^n (Z_i - \mu_{Z|Y}) \tilde{f}_{Y,X_i} \right) \left( 2f_Y - \frac{1}{n} \sum_{i=1}^n \tilde{f}_{Y,X_i} \right) \right] \\
= \frac{1}{n^4 f_Y^4} \text{Var}\left[ \sum_{i,j} (Z_i - \mu_{Z|Y}) \tilde{f}_{Y,X_i} \left( 2f_Y - \tilde{f}_{Y,X_j} \right) \right].
\]
Let
\[ \tau_i = (Z_i - \mu_Z|Y) \tilde{f}_{Y,X_i}, \]
\[ \gamma_i = 2f_Y - \tilde{f}_{Y,X_i}. \]

Using this notation, we must compute
\[
\frac{1}{n^4 f_Y^4} \text{Var}(\sum_{i,j} \tau_i \gamma_j) = \frac{1}{n^4 f_Y^4} \sum_{i,j,k,l} \text{Cov}(\tau_i \gamma_j, \tau_k \gamma_l). \tag{3.46}
\]

We partition summands on the right hand side of Equation (3.46) into sets. These sets are denoted by 4-tuples, where identical elements in the tuple imply identical indices in the summand. For example, \((i, j, i, k)\) is the set of summands where the the first and third indices are equal, but different from the second and fourth indices. The second and fourth indices are different from each other. We organize these sets based on the number of unique indices. For example, the tuple \((i, j, i, k)\) has 3 unique indices while the tuple \((i, i, j, j)\) has 2 unique indices.

We are looking for terms that are order \(n^{-1}\) and terms that are order \(n^{-2}\) and depend on \(H\). We do not keep track of terms that are order \(n^{-2}\) and do not depend on \(H\). Such terms will be denoted \(n^{-2}C\) where \(C\) is a constant that does not depend on \(H\) or \(n\). The precise value of \(C\) may change with each appearance.

For a given term, say \(\mathbb{E}[\tau_i]\), \(\mathbb{E}[\tau_i](0)\) refers to the 0th order expansion of \(\mathbb{E}[\tau_i]\) while \(\mathbb{E}[\tau_i](H^TH)\) refers to the second order expansion of \(\mathbb{E}[\tau_i]\). The constants in these expansions are computed in the referenced lemmas. In what follows we frequently use the identity
\[
\text{Cov}(\tau_i \gamma_j, \tau_k \gamma_l) = \mathbb{E}[\tau_i \gamma_j \tau_k \gamma_l] - \mathbb{E}[\tau_i \gamma_j] \mathbb{E}[\tau_k \gamma_l].
\]

- **4 unique indices**: \(n(n-1)(n-2)(n-3)\) summands.
  1. \((i, j, k, l)\): \(n(n-1)(n-2)(n-3)\) summands.

The \(\{(X_i, \delta_i)\}_{i=1}^n\) are independent. \(\tau_i\) depends only on \((X_i, \delta_i)\) and \(\gamma_i\) depends only on \(X_i\). Thus by independence
\[
\text{Cov}(\tau_i \gamma_j, \tau_k \gamma_l) = 0.
\]

- **3 unique indices**: \(6n(n-1)(n-2)\) summands.
  1. \((i, i, j, k)\) and \((i, j, k, k)\): \(2n(n-1)(n-2)\) summands.

By independence
\[
\text{Cov}(\tau_i \gamma_i, \tau_j \gamma_k) = 0, \\
\text{Cov}(\tau_i \gamma_j, \tau_k \gamma_k) = 0.
\]
2. \((i, j, i, k): n(n-1)(n-2)\) summands. We have

\[
\text{Cov}(\tau_i \gamma_j, \tau_i \gamma_k) = E[\tau_i^2]E[\gamma_j]E[\gamma_k] - E[\tau_i]E[\gamma_j]E[\gamma_k] = \text{Var}(\tau_1)E[\gamma_1]^2
\]

Using Lemma 3.13 on p.82 for the expansion of \(\text{Var}(\tau_1)\) and Lemma 3.8 on p.80 for the expansion of \(E[\gamma_1]\) we have

\[
\frac{n(n-1)(n-2)}{f_Y^4n^4} \text{Cov}(\tau_i \gamma_j, \tau_i \gamma_k) = \frac{n(n-1)(n-2)}{f_Y^4n^4} \frac{\text{Var}(\tau_1)}{n^4} \frac{E[\gamma_1]^2}{f_Y^4} \times \left( (\text{Var}(\tau_1)(0) + \text{Var}(\tau_1)(H^T H) + O(||H||^4)) \times (E[\gamma_1](0) + E[\gamma_1](H^T H) + O(||H||^4)) \right) = \frac{\text{Var}(\tau_1)(0)E[\gamma_1](0)^2}{n f_Y^4} + \frac{2\text{Var}(\tau_1)(0)E[\gamma_1](H^T H)E[\gamma_1](0)}{n f_Y^4} + \frac{\text{Var}(\tau_1)(H^T H)E[\gamma_1](0)^2}{n f_Y^4} + n^{-2}C + O(n^{-3}).
\]

3. \((i, j, k, j): n(n-1)(n-2)\) summands. We have

\[
\text{Cov}(\tau_i \gamma_j, \tau_k \gamma_j) = E[\gamma_j^2]E[\tau_i]E[\gamma_k] - E[\tau_i]E[\gamma_j]E[\gamma_k]^2 = \text{Var}(\gamma_1)E[\tau_1]^2.
\]

By Lemma 3.15 on p.84 that \(E[\tau_1] = O(||H||^2_{\infty})\) and by Assumption 3.7 that \(||H||_{\infty} = O(n^{-1/2})\) we have \(E[\tau_1] = O(n^{-1})\). Thus

\[
\frac{n(n-1)(n-2)}{n^4f_Y^4} \text{Cov}(\tau_i \gamma_j, \tau_k \gamma_j) = \frac{n(n-1)(n-2)}{n^4f_Y^4} \frac{\text{Var}(\tau_1)}{n^4f_Y^4} \frac{E[\gamma_1]^2}{n^4f_Y^4} = O(n^{-3}).
\]

4. \((i, j, k, i)\) and \((i, j, j, k): 2n(n-1)(n-2)\) summands. Note that by symmetry the summands in the sets \((i, j, k, i)\) and \((i, j, j, k)\) have the same value. We have

\[
\text{Cov}(\tau_i \gamma_j, \tau_k \gamma_i) = E[\gamma_j \gamma_i]E[\tau_i]E[\gamma_k] - E[\tau_i]E[\gamma_j]E[\gamma_k]E[\gamma_i] = E[\tau_1 \gamma_1]E[\tau_1]E[\tau_1] - E[\tau_1]^2E[\gamma_1]^2.
\]

By Lemma 3.15 on p.84 \(E[\tau_1] = O(||H||^2_{\infty})\). Therefore \(E[\tau_1]^2E[\gamma_1]^2 = O(||H||^4_{\infty}) = O(n^{-2})\). Using Lemma 3.10 on p.81 for the expansion of \(E[\tau_1 \gamma_1]\), Lemma 3.8 on
p.80 for the expansion of $\mathbb{E}[\gamma_1]$, and Lemma 3.15 on p.84 for the expansion of $\mathbb{E}[\tau_1]$, we have

$$\frac{2n(n-1)(n-2)}{n^4 f_Y^4} \text{Cov}(\tau_i \gamma_j, \tau_k \gamma_i)$$

$$= \frac{2}{nf_Y^4} \mathbb{E}[\tau_1 \gamma_1] \mathbb{E}[\gamma_1] \mathbb{E}[\tau_1] + O(n^{-3})$$

$$= \frac{2}{nf_Y^4} \mathbb{E}[\tau_1 \gamma_1](0) \mathbb{E}[\gamma_1](0) \mathbb{E}[\tau_1](H^T H) + O(n^{-3}).$$

- **2 unique indices:** $7n(n-1)$ summands. We show that each summand with exactly 2 unique indices is of the form $C + O(n^{-1})$ where $C$ is some constant that does not depend on $H$ or $n$. Since there are $O(n^2)$ summands with exactly 2 unique indices, this shows that after multiplying these summands by $n^{-4}$ (see Equation (3.46)), these terms together are of order $C n^{-2} + O(n^{-3})$ for some $C$ that does not depend on $H$ or $n$.

1. $(i, i, j, j)$: $n(n - 1)$ summands
   By independence
   $$\text{Cov}(\tau_i \gamma_i, \tau_j \gamma_j) = 0.$$

2. $(i, j, i, j)$: $n(n - 1)$ summands
   $$\text{Cov}(\tau_i \gamma_j, \tau_i \gamma_j) = \mathbb{E}[\tau_i^2 \mathbb{E}[\gamma_j^2] - \mathbb{E}[\tau_i]^{2} \mathbb{E}[\gamma_j]^{2} = \mathbb{E}[\tau_i^2] \mathbb{E}[\gamma_j^2] - \mathbb{E}[\tau_i] \mathbb{E}[\gamma_j]^{2}.$$  
   By Lemma 3.15 on p.84 $\mathbb{E}[\tau_1] = O(||H||_2^2)$. Thus $\mathbb{E}[\tau_1^2] \mathbb{E}[\gamma_1]^{2} = O(||H||_2^4) = O(n^{-2})$. Using Lemmas 3.14 on p.83 and 3.9 on p.81, we have that $\mathbb{E}[\tau_i^2]$ and $\mathbb{E}[\gamma_i^2]$ are of the form $C + O(n^{-1})$ for some $C$ that does not depend on $H$ or $n$. Thus
   $$\text{Cov}(\tau_i \gamma_j, \tau_i \gamma_j) = C + O(n^{-1}).$$

3. $(i, j, j, i)$: $n(n - 1)$ summands
   $$\text{Cov}(\tau_i \gamma_j, \tau_j \gamma_i) = \mathbb{E}[\tau_i \gamma_j \tau_j \gamma_i] - \mathbb{E}[\tau_i] \mathbb{E}[\gamma_j] \mathbb{E}[\tau_j \gamma_i] = \mathbb{E}[\tau_i \gamma_1]^{2} - \mathbb{E}[\tau_1]^{2} \mathbb{E}[\gamma_1]^{2}.$$  
   By Lemma 3.15 on p.84 $\mathbb{E}[\tau_1] = O(||H||_2^2)$. Thus $\mathbb{E}[\tau_1]^{2} \mathbb{E}[\gamma_1]^{2} = O(n^{-2})$. By Lemma 3.10 on p.81 $\mathbb{E}[\tau_1]^{2} = C + O(n^{-1})$. Thus we have
   $$\text{Cov}(\tau_i \gamma_j, \tau_j \gamma_i) = C + O(n^{-1}).$$
4. \((i, i, i, j)\) and \((i, j, i, i)\): \(2n(n - 1)\) summands. By symmetry summands in the sets 
\((i, i, i, j)\) and \((i, j, i, i)\) have the same value.

\[
\text{Cov}(\tau_i \gamma_i, \tau_i \gamma_j) = \mathbb{E}[\tau_i \gamma_i \tau_i \gamma_j] - \mathbb{E}[\tau_i \gamma_i] \mathbb{E}[\tau_i \gamma_j] = \mathbb{E}[\tau_i^2 \gamma_i] \mathbb{E}[\gamma_i] - \mathbb{E}[\tau_i \gamma_i] \mathbb{E}[\tau_i] \mathbb{E}[\gamma_i]
\]

By Lemma 3.15 on p.84 \(\mathbb{E}[\tau_i] = O(||H||_\infty^2) = O(n^{-1})\). By Lemma 3.11 on p.82 \(\mathbb{E}[\tau_i^2 \gamma_i] = C + O(n^{-1})\). By Lemma 3.8 on p.80 \(\mathbb{E}[\gamma_i] = C + O(n^{-1})\). Thus

\[
\text{Cov}(\tau_i \gamma_i, \tau_i \gamma_j) = C + O(n^{-1}).
\]

5. \((i, i, j, i)\) and \((i, j, j, j)\): \(2n(n - 1)\) summands. By symmetry summands in the sets 
\((i, i, j, i)\) and \((i, j, j, j)\) have the same value.

\[
\text{Cov}(\tau_i \gamma_j, \tau_j \gamma_j) = \mathbb{E}[\tau_i \gamma_j \tau_j \gamma_j] - \mathbb{E}[\tau_i \gamma_j] \mathbb{E}[\tau_j \gamma_j] = (\mathbb{E}[\tau_i \gamma_i^2] - \mathbb{E}[\tau_i \gamma_i] \mathbb{E}[\gamma_i]) \mathbb{E}[\tau_i].
\]

By Lemma 3.15 on p.84 \(\mathbb{E}[\tau_i] = O(||H||_\infty^2) = O(n^{-1})\). Thus

\[
\text{Cov}(\tau_i \gamma_j, \tau_j \gamma_j) = O(n^{-1}).
\]

- **1 unique index**: \(n\) summands

1. \((i, i, i, i)\): \(n\) summands. First note

\[
\text{Cov}(\tau_i \gamma_i, \tau_i \gamma_i) = Var(\tau_1 \gamma_1).
\]

\[
Var(\tau_1 \gamma_1) = O(1)\] by Lemma 3.12 on p.82 so

\[
\frac{n}{n^4 f_Y^4} \text{Cov}(\tau_i \gamma_i, \tau_i \gamma_i) = O(n^{-3}).
\]

Summing all the terms we obtain

\[
\text{Var}(\omega_n) = n^{-4} f_Y^{-4} \left( \text{Var}(\tau_1)(0) \mathbb{E}[\gamma_1](0)^2 + 2 \text{Var}(\tau_1)(0) \mathbb{E}[\gamma_1] \mathbb{E}[H^T H] \mathbb{E}[\gamma_1](0) + \text{Var}(\tau_1)(H^T H) \mathbb{E}[\gamma_1](0)^2 + 2 \mathbb{E}[\tau_1 \gamma_1](0) \mathbb{E}[\gamma_1](0) \mathbb{E}[\tau_1] \mathbb{E}[H^T H] + n^{-2} C + O(n^{-3}) \right).
\]
CHAPTER 3. NADARAYA–WATSON FOR PREDICTION WITH FEATURE MEASUREMENT ERROR

From Lemmas 3.8 on p.80, 3.10 on p.81, 3.13 on p.82, and 3.15 on p.84 we have

\[ \mathbb{E} [\gamma_1] = f_Y - \frac{1}{2} \text{tr} (H^T \Sigma_K H I(f_Y)) + O(||H||^4) , \]
\[ \mathbb{E} [\tau_1 \gamma_1] = -\mathbb{E} [\mu_Z | X(X_1) - \mu_Z | Y] f_\epsilon^2 (y - X_1) + O(||H||^2) , \]
\[ \text{Var} (\tau_1) = \mathbb{E} [f_\epsilon^2 (y - X_1) \nu_y (X_1)] + \text{tr} \left( H^T \Sigma_K H \mathbb{E} [I(f_\epsilon) (y - X_1) f_\epsilon (y - X_1) \nu_y (X_1)] \right) + O(||H||^4) , \]
\[ \mathbb{E} [\tau_1] = \frac{1}{2} \text{tr} (H^T \Sigma_K H (I(\mu_Z | Y) f_Y + 2 \nabla (\mu_Z | Y) \nabla^T (f_Y))) + O(||H||^4) . \]

Substituting these expansions into the \( \text{Var}(\omega_n) \) equation and denoting \( S = H^T \Sigma_K H \) we have

\[ \text{Var}(\omega_n) = n^{-1} f_Y^{-4} \left( \mathbb{E} [f_\epsilon^2 (y - X_1) \nu_y (X_1)] f_Y^2 - \mathbb{E} [f_\epsilon^2 (y - X_1) \nu_y (X_1)] \text{tr} (S I(f_Y)) f_Y + \text{tr} \left( S \mathbb{E} [I(f_\epsilon) (y - X_1) f_\epsilon (y - X_1) \nu_y (X_1)] \right) f_Y^2 \right. \]
\[ \left. - \mathbb{E} [\mu_Z | X(X_1) - \mu_Z | Y] f_\epsilon^2 (y - X_1) \text{tr} \left( S (I(\mu_Z | Y) f_Y + 2 \nabla (\mu_Z | Y) \nabla^T (f_Y)) \right) f_Y \right) + n^{-2} C + O(n^{-3}). \]

Define

\[ V = f_Y^{-2} \mathbb{E} [f_\epsilon^2 (y - X_1) \nu_y (X_1)], \]
\[ A_2 = -f_Y^{-3} \mathbb{E} [\mu_Z | X(X_1) - \mu_Z | Y] f_\epsilon^2 (y - X_1) (I(\mu_Z | Y) f_Y + 2 \nabla (\mu_Z | Y) \nabla^T (f_Y)), \]
\[ A_3 = -f_Y^{-3} \mathbb{E} [f_\epsilon^2 (y - X_1) \nu_y (X_1)] I(f_Y), \]
\[ A_4 = f_Y^{-2} \mathbb{E} [I(f_\epsilon) (y - X_1) f_\epsilon (y - X_1) \nu_y (X_1)]. \]

We have

\[ \text{Var}(\omega_n) = \frac{1}{n} V + \frac{1}{n} \text{tr} (S (A_2 + A_3 + A_4)) + n^{-2} C + O(n^{-3}). \]

\[ \Box \]

**Lemma 3.3.** We have

\[ 2 \text{Cov}(\omega_n, s_n) = C n^{-2} + O(n^{-5/2}) \]

where \( C \) is a constant that does not depend on \( H \) or \( n \).

**Proof.** Recall

\[ \omega_n = f_Y^{-1} \Delta_B - f_Y^{-1} \mu_Z | Y \Delta_A - f_Y^{-2} \Delta_A \Delta_B + f_Y^{-2} \Delta_A^2 \mu_Z | Y, \]
\[ s_n = \Delta_B \Delta_A^2 f_Y^{-3} - \Delta_A^3 f_Y^{-3} \mu_Z | Y. \]
CHAPTER 3. NADARAYA–WATSON FOR PREDICTION WITH FEATURE MEASUREMENT ERROR

Note that each term in $s_n$ is of the form $C \Delta_j A \Delta_l B$ where $j + l = 3$. Thus by Lemma 3.4 on p.75 $\mathbb{E}[|s_n|] = O(n^{-3/2})$. From Equation (3.41) in Subsection 3.2 we have $\mathbb{E}[\omega_n] = O(||H||_\infty^2) + O(n^{-1}) = O(n^{-1})$. Hence $\mathbb{E}[\omega_n] \mathbb{E}[s_n] = O(n^{-5/2})$. Thus
\[
2\text{Cov}(\omega_n, s_n) = 2\mathbb{E}[\omega_n s_n] + O(n^{-5/2}).
\]

Note that every term in the product $\omega_n s_n$ is of the form $C \Delta_j A \Delta_l B$ where $j + l \geq 4$ and $C$ is some constant. By Lemma 3.4 on p.75 we have
\[
2\mathbb{E}[\omega_n s_n] = Cn^{-2} + O(n^{-5/2}),
\]
where $C$ is a constant that does not depend on $H$ or $n$. Thus
\[
2\text{Cov}(\omega_n, s_n) = Cn^{-2} + O(n^{-5/2}).
\]

\[\square\]

Lemma 3.4. For any $j \in \mathbb{Z}^+$ and $l \in \{0, 1, 2, 3, 4\}$
\[
\mathbb{E}[\Delta_A^j \Delta_B^l] = Cn^{-(j+l)/2} + O(n^{-(j+l+1)/2}),
\]
where $C$ is a constant that does not depend on $H$ or $n$.

Proof. By Lemma 3.6 on p.78 we have
\[
\mathbb{E}[\Delta_A^j \Delta_B^l] = \mathbb{E} \left[ \left( \frac{1}{n} \sum_{i=1}^n W_i + T||H||_\infty^2 \right)^j \left( \frac{1}{n} \sum_{i=1}^n (Q_i + ||H||_\infty^2 R_i) + U||H||_\infty^2 \right)^l \right],
\]
where $\{(W_i, Q_i, R_i)\}_{i=1}^n$ are independent identically distributed with $\mathbb{E}[W_1] = \mathbb{E}[Q_1] = \mathbb{E}[R_1] = 0$. $\{(W_i, Q_i)\}_{i=1}^n$ do not depend on $H$. $W_1$, $T$ and $U$ are uniformly bounded in $H$ while $Q_1$ and $R_1$ have 4 moments that are bounded uniformly in $H$. Lemma 3.6 on p.78 has explicit representations for many of these random variables, however the stated properties are sufficient for proving this lemma.

We expand the binomials on the right hand side of Equation (3.47). For some $k \in \{0, \ldots, j\}$ and $m \in \{0, \ldots, l\}$ each term in the product of the binomial expansions is, up to a constant (that depends on $j$, $l$, $k$ and $m$ but not $n$), of the form
\[
\mathbb{E} \left[ \left( \frac{1}{n} \sum_{i=1}^n W_i \right)^{j-k} (T||H||_\infty^2)^k \left( \frac{1}{n} \sum_{i=1}^n (Q_i + ||H||_\infty^2 R_i) \right)^{l-m} (U||H||_\infty^2)^m \right].
\]
By Lemma 3.5 on p.76 when $k = m = 0$, Equation (3.48) is
\[
\mathbb{E} \left[ \left( \frac{1}{n} \sum_{i=1}^{n} W_i \right)^{j} \left( \frac{1}{n} \sum_{i=1}^{n} (Q_{i} + ||H||^2_{\infty} R_{i}) \right)^{l} \right] = C n^{-\frac{j+l}{2}} + O(n^{-\frac{j+l+1}{2}})
\]
for some $C$ that does not depend on $H$ or $n$.

Thus all that is left to show is that when $k \neq 0$ or $m \neq 0$, Expression 3.48 is $O(n^{-\frac{j+l+1}{2}})$.

In this case, Expression 3.48 is bounded by
\[
||H||^{2(k+m)}_{\infty} \max |T|^k \max |U|^m \mathbb{E} \left[ \left( \frac{1}{n} \sum_{i=1}^{n} W_i \right)^{j-k} \left( \frac{1}{n} \sum_{i=1}^{n} (Q_{i} + ||H||^2_{\infty} R_{i}) \right)^{l-m} \right]. \tag{3.49}
\]

Using the facts that $T$ and $U$ are bounded uniformly in $H$, $||H||_{\infty} = O(n^{-1/2})$ (by Assumption 3.7), and Lemma 3.5 on p.76, we bound Expression 3.49 by
\[
O(n^{-\frac{(k+m)}{2}})O(n^{-\frac{(j+l-k-m)}{2}}) = O(n^{-\frac{(j+l+1)}{2}}).
\]

\[\square\]

Lemma 3.5. Let $\{(W_{i}, Q_{i}, R_{i})\}_{i=1}^{n}$ be independent identically distributed with $\mathbb{E}[W_{i}] = \mathbb{E}[Q_{i}] = \mathbb{E}[R_{i}] = 0$. Assume $\{(W_{i}, Q_{i})\}_{i=1}^{n}$ do not depend on $H$. Assume $W_{1}$ is a bounded random variable while $Q_{1}$ and $R_{1}$ have 4 moments that are bounded uniformly in $H$. (Note: These assumptions are satisfied by the random variables constructed in Lemma 3.6 on p.78.) Then for $l \leq 4$ we have
\[
\mathbb{E} \left[ \left( \frac{1}{n} \sum_{i=1}^{n} W_i \right)^{j} \left( \frac{1}{n} \sum_{i=1}^{n} (Q_{i} + ||H||^2_{\infty} R_{i}) \right)^{l} \right] = C n^{-\frac{j+l}{2}} + O(n^{-\frac{j+l+1}{2}}),
\]
where $C$ is some constant that does not depend on $n$ or $H$.

Proof.
\[
\mathbb{E} \left[ \left( \frac{1}{n} \sum_{i=1}^{n} W_i \right)^{j} \left( \frac{1}{n} \sum_{i=1}^{n} (Q_{i} + ||H||^2_{\infty} R_{i}) \right)^{l} \right] = \frac{1}{n^{j+l}} \sum_{k \in \{1, \ldots, n\}^j} \sum_{m \in \{1, \ldots, n\}^l} \mathbb{E} \left[ \prod_{s=1}^{j} W_{k_s} \prod_{r=1}^{l} (Q_{m_r} + ||H||^2_{\infty} R_{m_r}) \right]. \tag{3.50}
\]

We separate the summands on the right side of Equation (3.50) into cases based on the number of unique elements in the the concatenated vector $(k, m) \in \{1, \ldots, n\}^{j+l}$.
CHAPTER 3. NADARAYA–WATSON FOR PREDICTION WITH FEATURE MEASUREMENT ERROR

- The number of unique elements in \((k, m)\) is greater than \((j + l)/2\). Since \((k, m)\) is of length \(j + l\), at least one element must appear only once in \((k, m)\). The random variable associated with this element is independent of all other terms in the product \(\prod W_k \prod (Q_{m_r} + ||H||^2_{\infty} R_{m_r})\). Since \(E[W_1] = E[Q_1] = E[R_1] = 0\), for any such \((k, m)\)

\[
E \left[ \prod_{s=1}^{j} W_k \prod_{r=1}^{l} (Q_{m_r} + ||H||^2_{\infty} R_{m_r}) \right] = 0.
\]

- The number of unique elements in \((k, m)\) is less than \((j + l)/2\). There are \(O(n^{(j+l-1)/2})\) summands of this type. Thus their contribution to the right hand side of Equation (3.50) (after multiplying by \(n^{-j-l}\)) is \(O(n^{-(j+l+1)/2})\).

- The number of unique elements in \((k, m)\) equals \((j + l)/2\). (This case only applies when \(j + l\) is even.) If any element appears 3 times in \((k, m)\), then another element must appear in \((k, m)\) exactly once. These terms will have expectation 0 as shown earlier. Thus it is sufficient to consider summands where every element in \((k, m)\) appears exactly two times.

Note that such summands are a product of terms of the form \(E[W_1^2]\), \(E[W_1(Q_1 + ||H||^2_{\infty} R_1)]\), and \(E[(Q_1 + ||H||^2_{\infty} R_1)^2]\). Since the moments of \(R_1\) are bounded uniformly in \(H\) and \(H = O(n^{-1/2})\) by Assumption 3.7, the summand

\[
E \left[ \prod_{s=1}^{j} W_k \prod_{r=1}^{l} (Q_{m_r} + ||H||^2_{\infty} R_{m_r}) \right] = E \left[ \prod_{s=1}^{j} W_k \prod_{r=1}^{l} Q_{m_r} \right] + O(n^{-1}). \tag{3.51}
\]

Since there are \(O(n^{(j+l)/2})\) summands where every element in \((k, m)\) occurs exactly two times, summing the \(O(n^{-1})\) terms (for each of these summands) in Equation (3.51), contributes (after multiplying by \(n^{-j-l}\)) \(O(n^{-(j+l+1)/2-1})\) to the right hand side of Equation (3.50). Hence it is sufficient to show

\[
\frac{1}{n^{j+l}} \sum_{k \in \{1, \ldots, n\}^j} \sum_{m \in \{1, \ldots, n\}^l} E \left[ \prod_{s=1}^{j} W_k \prod_{r=1}^{l} Q_{m_r} \right] = Cn^{-(j+l)/2} + O(n^{-(j+l+1)/2}), \tag{3.52}
\]

restricting the sums to \((k, m)\) where each element occurs exactly twice.

Partition the summands with exactly \((j + l)/2\) unique indices into sets, with two summands belonging to the same set if they share the same indices. There are \(\binom{n}{(j+l)/2}\) sets of this type. By symmetry, the sum of the summands in each of these sets has the same expectation, say \(C\). This \(C\) does not depend on \(H\) because \(W_i\) and \(Q_i\) do not depend on \(H\). Thus these terms all together sum to

\[
\binom{n}{(j+l)/2} C = Cn^{(j+l)/2} + O(n^{(j+l-1)/2}).
\]
CHAPTER 3. NADARAYA–WATSON FOR PREDICTION WITH FEATURE MEASUREMENT ERROR

After dividing by \( n^{j+l} \) these summands contribute \( Cn^{-(j+l)/2} + O(n^{-(j+l+1)/2}) \) to the right hand side of Equation (3.52).

Lemma 3.6. There exist random variables \( \{(W_i, Q_i, R_i)\}_{i=1}^n \), \( T \), and \( U \) such that

\[
\Delta_A = \frac{1}{n} \sum_{i=1}^n W_i + T \|H\|_\infty^2, \tag{3.53}
\]

\[
\Delta_B = \frac{1}{n} \sum_{i=1}^n (Q_i + \|H\|_\infty^2 R_i) + U \|H\|_\infty^2, \tag{3.54}
\]

where \( \{(W_i, Q_i, R_i)\}_{i=1}^n \) are independent identically distributed with \( \mathbb{E}[W_i] = \mathbb{E}[Q_1] = \mathbb{E}[R_1] = 0 \). \( \{(W_i, Q_i)\}_{i=1}^n \) do not depend on \( H \). \( W_1 \), \( T \) and \( U \) are bounded random variables, uniformly in \( H \), while \( Q_1 \) and \( R_1 \) have 4 moments that are bounded uniformly in \( H \).

Proof. We first show Equation (3.53) holds. Recall

\[
\Delta_A = \frac{1}{n} \sum_{i=1}^n \tilde{f}_{Y,X_i}(y) - f_Y(y).
\]

Let \( T \) be a random variable uniformly bounded in \( H \) that may change distribution at each appearance. By Lemma 3.16 on p.85

\[
\tilde{f}_{Y,X_i}(y) = f_\epsilon(y - X_i) + T \|H\|_\infty^2.
\]

Thus

\[
\Delta_A = \frac{1}{n} \sum_{i=1}^n (f_\epsilon(y - X_i) - f_Y(y)) + T \|H\|_\infty^2.
\]

Note that \( f_\epsilon(y - X_i) \) is bounded (by Assumption 3.11) and \( \mathbb{E}[f_\epsilon(y - X_i)] = f_Y(y) \). Thus the \( W_i \) are bounded with mean 0. We have shown that Equation (3.53) holds.

We now show Equation (3.54) holds. Recall

\[
\Delta_B = \frac{1}{n} \sum_{i=1}^n Z_i \tilde{f}_{Y,X_i}(y) - f_Y(y) \mu_{Z|Y}(y).
\]

By Lemma 3.16 on p.85

\[
\tilde{f}_{Y,X_i}(y) = f_\epsilon(y - X_i) + T_i \|H\|_\infty^2.
\]
where \( \{ T_i \}_{i=1}^{n} \) are i.i.d. random variables, uniformly bounded in \( H \). Recalling that \( Z_i = \mu_{Z|X}(X_i) + \delta_i \), we have

\[
\Delta_B = \frac{1}{n} \sum_{i=1}^{n} (\mu_{Z|X}(X_i) + \delta_i)(f_\epsilon(y - X_i) + T_i ||H||_\infty^2) - \mu_{Z|Y}(y)f_Y(y)
\]

\[
= \frac{1}{n} \sum_{i=1}^{n} \left( \mu_{Z|X}(X_i)f_\epsilon(y - X_i) + \delta_if_\epsilon(y - X_i) - \mu_{Z|Y}(y)f_Y(y) \right) + ||H||_\infty^2 \delta_i T_i \equiv Q_i
\]

\[
+ ||H||_\infty^2 \frac{1}{n} \sum_{i=1}^{n} T_i \mu_{Z|X}(X_i) \equiv U
\]

Note that \( T_i \) and \( \mu_{Z|X}(X_i) \) are bounded. By construction \( \{ T_i \mu_{Z|X}(X_i) \}_{i=1}^{n} \) are independent, identically distributed. Thus the conditions on \( U \) are satisfied.

\( \delta_i \) is independent of \( T_i \) and \( \mathbb{E}[\delta_i] = 0 \). \( \delta_i \) has 4 moments by Assumption 3.10 and \( T_i \) is bounded by construction. Therefore \( R_i \) has mean 0 and 4 moments. Further these moments are bounded uniformly in \( H \) since \( \mathbb{E}[R_i^k] \leq \max |T_i|^k \mathbb{E}[||\delta_i||^k] \) where \( T_i \) is bounded uniformly in \( H \) and \( \delta_i \) does not depend on \( H \).

Now it is sufficient to show that the \( Q_i \) have mean 0 and 4 moments. Note that by assumption \( \mathbb{E}[\delta_i] = 0 \). By construction \( \delta_i \) is independent of \( X_i \). By Lemma 3.18 on p.87 \( \mathbb{E}[\mu_{Z|X}(X_i)f_\epsilon(y - X_i)] = \mu_{Z|Y}(y)f_Y(y) \). Thus we have \( \mathbb{E}[Q_i] = 0 \). Note that \( \mu_{Z|X}(X_i)f_\epsilon(y - X_i) \) and \( f_\epsilon(y - X_i) \) are bounded and \( \delta_i \) has 4 moments by Assumption 3.10. Therefore \( \mathbb{E}[Q_i^4] < \infty \).

**Lemma 3.7.**

\[
\omega_n = f_Y^{-2} \left( \frac{1}{n} \sum_{i=1}^{n} (Z_i - \mu_{Z|Y}) \tilde{f}_{Y,X_i} \right) \left( 2f_Y - \frac{1}{n} \sum_{i=1}^{n} \tilde{f}_{Y,X_i} \right)
\]

**Proof.** Recall the definitions

\[
\Delta_A = \frac{1}{n} \sum_{i=1}^{n} \tilde{f}_{Y,X_i} - f_Y,
\]

\[
\Delta_B = \frac{1}{n} \sum_{i=1}^{n} Z_i \tilde{f}_{Y,X_i} - f_Y \mu_{Z|Y}.
\]
Using the definition of $\omega_n$ we have

\[
\omega_n = f_Y^{-1}\Delta_B - f_Y^{-1}\mu_{Z|Y}\Delta_A - f_Y^{-2}\Delta_A\Delta_B + f_Y^{-2}\Delta_A^2\mu_{Z|Y}
\]

\[
= (f_Y^{-1}\Delta_B - f_Y^{-1}\mu_{Z|Y}\Delta_A)(1 - f_Y^{-1}\Delta_A)
\]

\[
= \left( f_Y^{-1}\left[ \frac{1}{n}\sum_{i=1}^{n} Z_i\tilde{f}_{Y,X_i} - f_Y\mu_{Z|Y} \right] - f_Y^{-1}\mu_{Z|Y} \left[ \frac{1}{n}\sum_{i=1}^{n} \tilde{f}_{Y,X_i} - f_Y \right] \right) (1 - f_Y^{-1}\Delta_A)
\]

\[
= \left[ \frac{1}{n}\sum_{i=1}^{n} (Z_i - \mu_{Z|Y})\tilde{f}_{Y,X_i} \right] f_Y (1 - f_Y^{-1}\Delta_A)
\]

\[
= \left[ \frac{1}{n}\sum_{i=1}^{n} (Z_i - \mu_{Z|Y})\tilde{f}_{Y,X_i} \right] \left( 2 - \frac{1}{n}\sum_{i=1}^{n} \tilde{f}_{Y,X_i} \right)
\]

\[
= f_Y^2 \left( \frac{1}{n}\sum_{i=1}^{n} (Z_i - \mu_{Z|Y})\tilde{f}_{Y,X_i} \right) \left( 2f_Y - \frac{1}{n}\sum_{i=1}^{n} \tilde{f}_{Y,X_i} \right)
\]

Lemma 3.8.

\[
\mathbb{E}[\gamma_1] = f_Y - \frac{1}{2}\text{tr}(H^T\Sigma_K H\mathcal{I}(f_Y)) + O(||H||_\infty^4).
\]

Proof. By Lemma 3.17 on p.85, for some random variable $T$ uniformly bounded in $H$ we have

\[
\tilde{f}_{Y,X_i}(y) = f_i(y - X_1) + \frac{1}{2}\text{tr}(H^T\Sigma_K H\mathcal{I}(f_i)(y - X_1)) + T||H||_\infty^4. \tag{3.55}
\]

Using the definition of $\gamma_1$ and Equation (3.55) we have

\[
\gamma_1 = 2f_Y - \tilde{f}_{Y,X_1}
\]

\[
= 2f_Y - f_i(y - X_1) - \frac{1}{2}\text{tr}(H^T\Sigma_K H\mathcal{I}(f_i)(y - X_1)) - T||H||_\infty^4.
\]

Noting that $\mathcal{I}(f_i)$ is uniformly bounded (Assumption 3.11), and $\int \mathcal{I}(f_i)(y - x)f_X(x)dx = \mathcal{I}(\int f_i(y - x)f_X(x)dx) = \mathcal{I}(f_Y)$, we have

\[
\mathbb{E}[\gamma_1] = f_Y - \frac{1}{2}\text{tr}(H^T\Sigma_K H\mathbb{E}[\mathcal{I}(f_i)(y - X_1)]) + O(||H||_\infty^4)
\]

\[
= f_Y - \frac{1}{2}\text{tr}(H^T\Sigma_K H\mathcal{I}(f_Y)) + O(||H||_\infty^4).
\]

\[\square\]
Lemma 3.9. We have
\[ E[\gamma_1^2] = C + O(||H||_\infty^2) \]
for some \( C \) that does not depend on \( n \) or \( H \).

Proof. Using the definition of \( \gamma_1 \) and the result from Lemma 3.17 on p.85 that \( \tilde{f}_{Y,X_1} = f_e(y - X_1) + T||H||_\infty^2 \) for some random variable \( T \) uniformly bounded in \( H \), we have
\[
\begin{align*}
\gamma_1^2 &= (2f_Y - \tilde{f}_{Y,X_1})^2 \\
&= 4f_Y^2 - 4f_Y\tilde{f}_{Y,X_1} + \tilde{f}_{Y,X_1}^2 \\
&= 4f_Y^2 - 4f_Yf_e(y - X_1) - 4f_YT||H||_\infty^2 \\
&\quad + f_e^2(y - X_1) + 2f_e(y - X_1)T||H||_\infty^2 + T^2||H||_\infty^4.
\end{align*}
\]
Noting that \( f_e \) is bounded by Assumption 3.11, by taking expectations on both sides we have
\[
E[\gamma_1^2] = 4f_Y^2 - 4f_YE[f_e(y - X_1)] + E[f_e^2(y - X_1)] + O(||H||_\infty^2)
\]
\[
= E[f_e^2(y - X_1)] + O(||H||_\infty^2).
\]
\[
= C + O(||H||_\infty^2),
\]
where \( C \) does not depend on \( H \) or \( n \). \( \square \)

Lemma 3.10.
\[ E[\tau_1\gamma_1] = -E[(\mu_{Z|X}(X_1) - \mu_{Z|Y})f_e^2(y - X_1)] + O(||H||_\infty^2) \]

Proof. Recall
\[
\tau_1 = (Z_1 - \mu_{Z|Y})\tilde{f}_{Y,X_1}
\]
\[
\gamma_1 = 2f_Y - \tilde{f}_{Y,X_1}.
\]
By Lemma 3.15 \( E[\tau_1] = O(||H||_\infty^2) \). We have
\[
E[\tau_1\gamma_1] = E[\tau_1(2f_Y - \tilde{f}_{Y,X_1})]
\]
\[
= 2f_YE[\tau_1] - E[(Z_1 - \mu_{Z|Y})\tilde{f}_{Y,X_1}^2]
\]
\[
= -E[(Z_1 - \mu_{Z|Y})\tilde{f}_{Y,X_1}^2] + O(||H||_\infty^2).
\]
By Lemma 3.16 on p.85 \( \tilde{f}_{Y,X_1}^2(y) = f_e^2(y - x) + T||H||_\infty^2 \) for some bounded random variable \( T \). Using this equality and the fact that \( \delta_1 \) is mean 0 and independent of all other random variables, we have
\[
E[\tau_1\gamma_1] = -E[(Z_1 - \mu_{Z|Y})f_e^2(y - X_1)] - E[T(Z_1 - \mu_{Z|Y})||H||_\infty^2 + O(||H||_\infty^2)
\]
\[
= -E[(\mu_{Z|X}(X_1) + \delta_1 - \mu_{Z|Y})f_e^2(y - X_1)] + O(||H||_\infty^2)
\]
\[
= -E[(\mu_{Z|X}(X_1) - \mu_{Z|Y})f_e^2(y - X_1)] + O(||H||_\infty^2).
\]
Lemma 3.11. 
\[ \mathbb{E}[\tau_1^2 \gamma_1] = C + O(||H||^2_{\infty}) \]
where \( C \) is a constant that does not depend on \( H \) or \( n \).

*Proof.* Recall
\[ \tau_1 = (Z_1 - \mu_{Z|Y}) \tilde{f}_{Y,X_1}, \]
\[ \gamma_1 = 2f_Y - \tilde{f}_{Y,X_1} \]

By Lemma 3.16 on p.85, we have
\[ \tilde{f}_{Y,X_1}(y) = f_\epsilon(y - X_1) + T||H||^2_{\infty}, \]
where \( T \) is a bounded random variable. Thus we have
\[
\mathbb{E}[\tau_1^2 \gamma_1] = \mathbb{E}[(Z_1 - \mu_{Z|Y})(f_\epsilon(y - X_1) + T||H||^2_{\infty})]^2 (2f_Y - f_\epsilon(y - X_1) - T||H||^2_{\infty})] \\
= \mathbb{E}[(Z_1 - \mu_{Z|Y})f_\epsilon(y - X_1)]^2 (2f_Y - f_\epsilon(y - X_1)) + O(||H||^2_{\infty}) \\
= C + O(||H||^2_{\infty}).
\]

\[ \square \]

Lemma 3.12. \( \text{Var}(\tau_1 \gamma_1) = O(1) \).

*Proof.* \( \tilde{f}_{Y,X_1} = \int K_H(y - X_1 - \epsilon)f_\epsilon(\epsilon)d\epsilon \) is a convolution of the bounded density \( f_\epsilon \) with \( X_1 \) and \( K_H \), therefore it is bounded. Thus
\[
\gamma_1^2 = (2f_Y - \tilde{f}_{Y,X_1})^2 \leq D
\]
for some \( D \). By Lemma 3.14 on p.83 \( \mathbb{E}[\tau_1^2] = O(1) \). Thus
\[
\text{Var}(\tau_1 \gamma_1) \leq \mathbb{E}[\tau_1^2 \gamma_1^2] \leq D\mathbb{E}[\tau_1^2] = O(1).
\]

\[ \square \]

Lemma 3.13.
\[
\text{Var}(\tau_1) = \mathbb{E}[f_\epsilon^2(y - X_1) \nu_y(X_1)] \\
+ \text{tr}
\left(H^T\Sigma_K \mathbb{E}[\mathcal{I}(f_\epsilon)(y - X_1)f_\epsilon(y - X_1)\nu_y(X_1)]\right) \\
+ O(||H||^4_{\infty}).
\]
CHAPTER 3. NADARAYA–WATSON FOR PREDICTION WITH FEATURE MEASUREMENT ERROR

Proof. By Lemma 3.15 on p.84 $E[\tau_1] = O(||H||_\infty^2)$. Using this bound and Lemma 3.14 on p.83 for the expansion of $E[\tau_1^2]$ we have

$$\text{Var}(\tau_1) = E[\tau_1^2] - E[\tau_1]^2$$

$$= E[\tau_1^2] + O(||H||_\infty^4)$$

$$= E[f_\epsilon^2(y - X_1)\nu_g(X_1)]$$

$$+ \text{tr}\left(H^T \Sigma_K H E[\mathcal{I}(f_\epsilon)(y - X_1)f_\epsilon(y - X_1)\nu_g(X_1)]\right)$$

$$+ O(||H||_\infty^4).$$


$$E[\tau_1^2] = E[f_\epsilon^2(y - X_1)\nu_g(X_1)]$$

$$+ \text{tr}\left(H^T \Sigma_K H E[\mathcal{I}(f_\epsilon)(y - X_1)f_\epsilon(y - X_1)\nu_g(X_1)]\right)$$

$$+ O(||H||_\infty^4).$$

Proof. Define

$$\nu_y(x) = (\mu_{z|x}(x) - \mu_{z|y}(y))^2 + \sigma_\delta^2.$$

Recalling the definition of $\tau_1$, $E[\delta_1^2] = \sigma_\delta^2$, $E[\delta_1] = 0$, and $\delta_1 \perp X_1$, we have

$$E[\tau_1^2] = E[\{(Z_1 - \mu_{z|y})\tilde{f}_{Y,X_1}\}^2]$$

$$= E[\{(\mu_{z|x}(X_1) + \delta_1 - \mu_{z|y})^2 \tilde{f}_{Y,X_1}^2\}]$$

$$= E[\{(\mu_{z|x}(X_1) - \mu_{z|y})^2 + \sigma_\delta^2) \tilde{f}_{Y,X_1}^2\}]$$

$$= E[\nu_y(x)\tilde{f}_{Y,X_1}^2].$$

(3.56)

Let $T$ be a random variable that may change distribution at each appearance but is uniformly bounded for all $H$. By Lemma 3.17 on p.85 we have

$$\tilde{f}_{Y,X_1}(y) = f_\epsilon(y - X_1) + \frac{1}{2}\text{tr}(H^T \Sigma_K H \mathcal{I}(f_\epsilon)(y - X_1)) + T||H||_\infty^4.$$

Noting that $f_\epsilon(y - X_1)$ and $\mathcal{I}(f_\epsilon)(y - X_1)$ are both bounded by Assumption 3.11, by squaring both sides we obtain

$$\tilde{f}_{Y,X_1}^2(y) = f_\epsilon^2(y - X_1) + \text{tr}(H^T \Sigma_K H \mathcal{I}(f_\epsilon)(y - X_1)f_\epsilon(y - X_1)) + T||H||_\infty^4.$$

(3.57)
CHAPTER 3. NADARAYA–WATSON FOR PREDICTION WITH FEATURE MEASUREMENT ERROR

Substituting Equation (3.57) into Equation (3.56) we obtain

\[ \mathbb{E}[\tau_1^2] = \mathbb{E}[f^2_\tau(y - X_1)\nu_y(X_1)] \\
+ \text{tr}\left( H^T \Sigma_{K} H \mathbb{E}[\mathcal{I}(f_\tau)(y - X_1)f_\tau(y - X_1)\nu_y(X_1)] \right) \\
+ O(||H||^4_\infty). \]

Lemma 3.15.

\[ \mathbb{E}[\tau_1] = \frac{1}{2} \text{tr}\left( H^T \Sigma_{K} H \left( \mathcal{I}(\mu_{Z|Y})f_Y + 2\nabla(\mu_{Z|Y})\nabla^T(f_Y) \right) \right) + O(||H||^4_\infty) \]

Hence \( \mathbb{E}[\tau_1] = O(||H||^2_\infty) \).

Proof. Let \( T \) be a random variable uniformly bounded in \( H \) that may change distribution at each appearance. By Lemma 3.17 on p.85

\[ \tilde{f}_{Y,X_1}(y) = f_\tau(y - X_1) + \frac{1}{2} \text{tr}(H^T \Sigma_{K} H \mathcal{I}(f_\tau)(y - X_1)) + T||H||^4_\infty. \]

Noting that \( Z_1 = \mu_{Z|X}(X_1) + \delta_1 \) where \( \mathbb{E}[\delta_1] = 0 \) and \( \delta_1 \) is independent of all other random variables and \( \mathbb{E}[\mu_{Z|X}(X_1)f_\tau(y - X_1)] - \mu_{Z|Y}(y)f_Y(y) = 0 \) (Lemma 3.18 on p.87), we have

\[ \mathbb{E}[\tau_1] = \mathbb{E}\left[ (\mu_{Z|X}(X_1) + \delta_1 - \mu_{Z|Y}) \left( f_\tau(y - X_1) + \frac{1}{2} \text{tr}(H^T \Sigma_{K} H \mathcal{I}(f_\tau)(y - X_1)) + T||H||^4_\infty \right) \right] \\
= \mathbb{E}[\mu_{Z|X}(X_1)f_\tau(y - X_1)] - \mu_{Z|Y}f_Y \\
+ \frac{1}{2} \text{tr}(H^T \Sigma_{K} H \mathbb{E}[\mu_{Z|X}(X_1) - \mu_{Z|Y}]\mathcal{I}(f_\tau)(y - X_1))] + O(||H||^4_\infty) \\
= \frac{1}{2} \text{tr}(H^T \Sigma_{K} H \mathbb{E}[\mu_{Z|X}(X_1)] - \mu_{Z|Y})\mathcal{I}(f_\tau)(y - X_1))] + O(||H||^4_\infty). \]

Note that \( \int \mathcal{I}(f(t - x))g(x)dx = \mathcal{I}(\int f(t - x)g(x)dx) \). Further note that by Lemma 3.18 on p.87 \( \mathbb{E}[\mu_{Z|X}(X_1)f_\tau(y - X_1)] = \mu_{Z|Y}(y)f_Y(y) \). Using these two facts and noting that all Hessian operators (\( \mathcal{I} \)) are taken with respect to \( y \), we have

\[ \mathbb{E}[\mu_{Z|X}(X_1) - \mu_{Z|Y}]}\mathcal{I}(f_\tau)(y - X_1))] = \mathbb{E}[\mu_{Z|X}(X_1)]\mathcal{I}(f_\tau)(y - X_1)] - \mu_{Z|Y}\mathcal{I}(f_\tau)(y - X_1)] \\
= \mathcal{I}(\mathbb{E}[\mu_{Z|X}(X_1)f_\tau(y - X_1)]) - \mu_{Z|Y}\mathcal{I}(\mathbb{E}[f_\tau(y - X_1)]) \\
= \mathcal{I}(\mu_{Z|Y}f_Y) - \mu_{Z|Y}\mathcal{I}(f_Y) \\
= \mathcal{I}(\mu_{Z|Y})f_Y + 2\nabla(\mu_{Z|Y})\nabla(f_Y)^T. \]

Substituting Equation (3.59) into Equation (3.58), we have

\[ \mathbb{E}[\tau_1] = \frac{1}{2} \text{tr}\left( H^T \Sigma_{K} H \left( \mathcal{I}(\mu_{Z|Y})f_Y + 2\nabla(\mu_{Z|Y})\nabla(f_Y)^T \right) \right) + O(||H||^4_\infty). \]
Lemma 3.16. For any $j \in \mathbb{Z}^+$ there exists a random variable $T$ uniformly bounded in $H$ such that

\[
\tilde{f}^j_{Y,X_1}(y) = f^j_{\epsilon}(y - X_1) + T||H||^2_{\infty}.
\]

Proof. Let $T$ be a random variable bounded uniformly in $H$ that may change distribution at each appearance. By Lemma 3.17 on p.85 we have

\[
\tilde{f}_{Y,X_1}(y) = f_{\epsilon}(y - X_1) + \frac{1}{2}\text{tr}(H^T\Sigma_K H\mathcal{I}(f_{\epsilon})(y - X_1)) + T||H||^4_{\infty}.
\]

Since all elements of $\mathcal{I}(f_{\epsilon})$ are uniformly bounded (Assumption 3.11) we have

\[
\tilde{f}_{Y,X_1}(y) = f_{\epsilon}(y - X_1) + T||H||^2_{\infty}.
\]

Since $f_{\epsilon}$ is bounded (Assumption 3.11), we take the $j^{th}$ power of each side and obtain

\[
\tilde{f}^j_{Y,X_1}(y) = f^j_{\epsilon}(y - X_1) + T||H||^2_{\infty}.
\]

Lemma 3.17. For some random variable $T$ uniformly bounded in $H$

\[
\tilde{f}_{Y,X_1}(y) = f_{\epsilon}(y - X_1) + \frac{1}{2}\text{tr}(H^T\Sigma_K H\mathcal{I}(f_{\epsilon})(y - X_1)) + T||H||^4_{\infty}.
\]

Proof. We Taylor expand $\tilde{f}_{Y,X_1}(y)$ around $H = 0$. Using the definition of $\tilde{f}_{Y,X_1}(y)$, the definition of $K_H$, and the substitution $u = H^{-1}(y - X_1 - \epsilon)$ we have

\[
\tilde{f}_{Y,X_1}(y) = \int K_H(y - X_1 - \epsilon)dF_{\epsilon}
\]

\[
= \frac{1}{\det(H)}\int K(H^{-1}(y - X_1 - \epsilon))f_{\epsilon}(\epsilon)d\epsilon
\]

\[
= \int K(u)f_{\epsilon}(y - X_1 - Hu)du.
\]

We perform a third order Taylor expansion of $f_{\epsilon}(y - X_1 - Hu)$ around $y - X_1$. By Assumption 3.11 $f_{\epsilon}$ is four times continuously differentiable, so the expansion is valid point wise. Denote the remainder term by $R(Hu)$. We split the integrals and then show that they exist. We
have
\[
\tilde{f}_{Y,X_1}(y) = f_\epsilon(y - X_1) + \frac{1}{2} \int u^T H^T \mathcal{I}(f_\epsilon)(y - X_1) HuK(u) du + T||H||_\infty^4. \tag{3.65}
\]

Recalling \(\text{tr}(AB) = \text{tr}(BA)\) and \(\int uu^T K(u) du = \Sigma_K\) we have
\[
\int u^T H^T \mathcal{I}(f_\epsilon)(y - X_1) HuK(u) du = \text{tr} \left( \int u^T H^T \mathcal{I}(f_\epsilon)(y - X_1) HuK(u) du \right) = \text{tr} \left( H \left[ \int uu^T K(u) du \right] H^T \mathcal{I}(f_\epsilon)(y - X_1) \right) = \text{tr} \left( H \Sigma_K H^T \mathcal{I}(f_\epsilon)(y - X_1) \right). \tag{3.66}
\]

Substituting Equation (3.66) into Equation (3.65) we obtain
\[
\tilde{f}_{Y,X_1}(y) = f_\epsilon(y - X_1) + \frac{1}{2} \text{tr}(H^T \Sigma_K H \mathcal{I}(f_\epsilon)(y - X_1)) + T||H||_\infty^4.
\]

Since \(K(u) = K(-u)\forall u\) and \(K\) has four moments (Assumptions 3.4 and 3.5), Expressions (3.61) and (3.63) are 0. The integral in Expression (3.62) exists because the elements of \(\mathcal{I}(f_\epsilon)\) are uniformly bounded and \(K\) has four moments (Assumptions 3.11 and 3.5). The integral in Expression 3.64 exists because it is the difference of integrable functions. By the mean value form of the Taylor remainder we have
\[
\int R(Hu)K(u)du = \int \sum_{\{j: \sum_{k=1}^p j_k = 3\}} \frac{\partial^3 f_\epsilon}{\partial \epsilon_1^i \ldots \partial \epsilon_p^j} (y - X_1) \prod_{k=1}^p \frac{(Hu)_k^{j_k}}{j_k!} K(u) du
\]

where \(\xi = (y - X_1) + tHu\) for some \(t \in [0, 1]\). By Assumption 3.11, \(\frac{\partial^3 f_\epsilon}{\partial \epsilon_1^i \ldots \partial \epsilon_p^j} (\xi)\) is uniformly bounded in \(\xi\). Since \(K\) has fourth moments, \(\int \prod_{k=1}^p u_k^{j_k} K(u) du\) is bounded for all \(j\). Therefore \(\int R(Hu)K(u)du = T||H||_\infty^4\) for some random variable \(T\) bounded uniformly in \(H\). Thus,
Lemma 3.18.

$$\mu_{Z|Y}(y)f_Y(y) = \mathbb{E}[\mu_{Z|X}(X_1)f_c(y - X_1)]$$

Proof. We use the definition $\mu_{Z|Y}(y) = \mathbb{E}[Z|Y = y]$. We have

$$\mu_{Z|Y}(y)f_Y(y) = \mathbb{E}[Z|Y = y]f_Y(y)$$

$$= \left( \int (\mu_{Z|X}(x) + \delta)f_\delta(\delta)p(x|Y = y)\,d\delta \,dx \right) f_Y(y)$$

$$= \left( \int \mu_{Z|X}(x)f_c(y - x)f_X(x)\,dx \right) f_Y(y)$$

$$= \mathbb{E}[\mu_{Z|X}(X_1)f_c(y - X_1)].$$

\[\square\]
Chapter 4

Empirical Study of Noisification

4.1 Introduction

Classification of periodic variables is crucial for scientific knowledge discovery and efficient use of telescopic resources for source follow up [Eyer and Mowlavi, 2008, Walkowicz et al., 2009]. As the size of synoptic surveys has grown, a greater and greater share of the classification process must become automated [Bloom and Richards, 2011]. With *Hipparcos*, it was possible for astronomers to individually analyze and classify each of the 2712 periodic variables observed in the survey. Starting in 2013, Gaia is expected to discover $\sim$ 5 million classical periodic variables over the course of its 4–5-year mission [Eyer and Cuypers, 2000]. LSST, for that matter, may collect on the order of a billion [Borne et al., 2007]. Individual analysis and classification by hand of all periodic variables is no longer feasible.

The need for efficient and accurate source classification has motivated much recent work on applying statistical and machine learning methods to variable star data sets (e.g., Eyer and Blake 2005, Debosscher et al. 2007, Richards et al. 2011, Dubath et al. 2011). In these papers, classifiers were constructed using light curves from a variety of surveys, such as the Optical Gravitational Lensing Experiment (OGLE, Soszynski et al. 2011), *Hipparcos* [Perryman et al., 1997a], The All-Sky Automated Survey (ASAS, Pojmanski et al. 2005), the COnvection, ROtation & planetary Transits survey (CoRoT, Auvergne et al. 2009), and the Geneva Extrasolar Planet Search. Often the intention of these studies is to develop classifiers with high accuracy in classifying sources from surveys other than those used to construct the classifier. For example, Blomme et al. [2011] trained a classifier on a mixture of *Hipparcos*, OGLE, and CoRoT sources and used it to classify sources from the Trans-atlantic Exoplanet Survey (TrES, O’Donovan et al. 2009) Lyr1 field. Dubath et al. [2011] and Eyer et al. [2008] view their work on classification of *Hipparcos* sources as a precursor to classification of yet-to-be collected Gaia light curves. Debosscher and collaborators trained
a classifier on a mixture of OGLE and Hipparcos sources in attempts to classify CoRoT sources [Debosscher et al., 2007, Sarro and Debosscher J., 2008, Debosscher et al., 2009].

It is well known that systematic differences in cadence, observing region, flux noise, detection limits, and number of observed epochs per light curve exist among surveys. Even within surveys there is heterogeneity in these characteristics. Most statistical classifiers assume that the light curves of a known class used to construct the classifier, termed training data, and the light curves of unknown class which we wish to classify, termed unlabeled data, share the same characteristics. This is unlikely to be the case when training and unlabeled light curves come from different surveys, or when the best-quality light curves of sources from each class are used to classify poorly sampled light curves of unknown class from the same survey.

(a) (b)

Figure 4.1: (a) The grey lines represent the CART classifier constructed using Hipparcos data. The points are Hipparcos sources. This classifier separates Hipparcos sources well (0.6% error as measured by cross-validation). (b) Here the OGLE sources are plotted over the same decision boundaries. There is now significant class overlap in the amplitude–fold2P plane (30% error rate). This is due to shifts in feature distributions across surveys.

To illustrate how seriously survey mismatches can deteriorate classification performance, consider the three-class problem of separating Mira variables, Classical Cepheids, and Fundamental Mode RR Lyrae from the Hipparcos and OGLE surveys. From OGLE, we use V-band data. Note that OGLE is far better sampled in I-band than V-band. We use V-
band to create a setting where one set of data is well sampled while the other set is poorly sampled. See Section 4.5.3 and Table 4.2 for more information on these sources.

For each light curve we compute dozens of metrics, termed *features*, that contain important information related to source class (e.g., frequency and amplitude; see Section 4.2 for details on feature selection and extraction). Using the *Hipparcos* light curves we construct a classifier using CART.\(^1\) The resulting classifier uses only two features for separating classes: the amplitude of a best fit sinusoidal model and the 90\(^{th}\) percentile of the slope between phase adjacent flux measurements after the light curve has been folded on twice the estimated period.

Figure 4.1a displays these two features for each *Hipparcos* source with grey lines denoting the class boundaries chosen by CART. Based on the *Hipparcos* light curves, this looks like an excellent classifier as each of the three regions of feature space selected by CART contains sources of only one class. However, examining a subset of the OGLE sources, Figure 4.1b, shows large class overlap on these two features. Here these two features do not separate OGLE sources well. The error rate measured by cross-validation on the *Hipparcos* sources was only 0.6\(^{\circ}\)\(^2\). However, the misclassification rate on the OGLE sources is 30.0%.

Despite what the 30.0% error rate seems to imply, the problem of separating classes in OGLE is not inherently difficult. A CART classifier trained on the OGLE light curves has a cross-validated error rate of 1.3%. While there are many systematic differences between the *Hipparcos* and OGLE surveys, their radically different cadences and number of flux measurements per light curve appear to be driving the increase in misclassification rate. For example, both features in Figure 4.1 depend on the estimate of each source’s period; yet, over 25% of the RR Lyrae in OGLE have incorrectly estimated periods due to poor sampling in the V-band.

A natural question to ask is: If we had observed the *Hipparcos* sources at an OGLE cadence, what classifier would CART have constructed, and how would this have changed the error rate? In this chapter we use *noisification*, a method which matches the cadence of training data and unlabeled data by inferring a continuous periodic function for each training light curve and then extracting flux measurements at the cadence and photometric error level present in the unlabeled light curves. The purpose of noisification is to automatically shift the distribution of features in the training data closer to the distribution of features in the unlabeled data so that a classifier can determine class boundaries as they exist in the unlabeled data. Versions of noisification were introduced in Starr et al. [2010] and Long et al. [2011]. In this chapter, we demonstrate that noisification improves classification accuracy on several simulated and real-world data sets. For instance, on the OGLE – Hipparcos three

---

\(^1\) CART (Classification And Regression Trees) is a popular classifier that forms a sequence of nested binary partitions of feature space. See Breiman et al. [1984] for more on CART.

\(^2\) See 4.2.4 for a definition of cross-validation.
class problem we reduce misclassification rate by 20.0%. Performance increases are greatest when the training data is well sampled at a particular cadence while unlabeled light curves are either poorly time sampled or observed at a different cadence.

This chapter is organized as follows. In Section 4.2 we briefly outline the statistical classification framework and show how it is applied in the context of periodic variables. In Section 4.3 we illustrate the problems that occur when training and unlabeled data come from different surveys. We present noisification, a method for overcoming differences related to number of flux measurements, cadence, and photometric error in Section 4.4. In Section 4.5 we apply noisification to several data sets. Finally in Section 4.6 we discuss possible uses of noisification for upcoming surveys.

4.2 Overview of Classification of Periodic Variables

Here we review a methodology for constructing, implementing, and evaluating statistical classifiers for periodic variables. This approach has been used in many recent works. For a more detailed review of the methodology see Debosscher et al. [2007] or Richards et al. [2011].

4.2.1 Constructing a Classifier

We start with a set of light curves of known class, termed training data and a set of light curves of unknown class, termed unlabeled data. Our goal is to determine the classes for the unlabeled light curves using information present in the training data. Each light curve consists of a set of time, flux, and photometric error measurements. We compute functions of the time, flux, and photometric error, termed features. Features are chosen to contain information relevant for differentiating classes. The same set of features is computed for each light curve. A statistical classification method uses the training data to learn a relationship between features and class and produces a classifier $C$. Given the features, $x$, for a light curve in the unlabeled set, $C(x)$ is a prediction of its class.

4.2.2 Feature Set

We use a total of 62 features to describe each light curve. 50 of these features are described in Tables 4 and 5 of Richards et al. [2011].\(^3\) We use 12 other features, described in Appendix 4.7 of this article. Many of the features that we use are obvious choices e.g., frequency and

\(^3\)We do not use pair_slope_trend, max_slope, or linear_trend.
amplitude. Most of our features, or features very similar to the ones here, have been used in recent work on classification of periodic variables [Kim et al., 2011, Dubath et al., 2011].

4.2.3 Choosing a Classifier

There are many statistical classification methods for constructing the function $C$. Some of the most popular include linear discriminant analysis (LDA), neural networks, support vector machines (SVMs), and Random Forests. In an earlier example we used CART. Each classification method has its own strengths and weaknesses. See Hastie et al. [2009] for an extensive discussion of classification methods. In this work we use the Random Forests classifier developed by Breiman [2001], Amit and Geman [1997], and Dietterich [2000]. Random Forests has been used, with high levels of success, in recent studies of automated variable star classification [Richards et al., 2011, Dubath et al., 2011]. Richards et al. [2011], in a side–by–side comparison of 10 different classifiers using OGLE and *Hipparcos* data, found that Random Forest had the lowest error rate.

4.2.4 Estimating Classifier Accuracy

Usually, researchers want an estimate of how accurate the classifier, $C$, will be when presented with new, unlabeled data. Simply calculating the proportion of times $C$ correctly classifies light curves in the training data is a poor estimate of classifier success, as this typically overestimates classifier performance on unlabeled data. Better assessment of classifier performance on unlabeled data is attained by using training–test set splits or cross–validation. With training–test set splits a fraction of the data, usually between 10% and 30%, is “held out” while the rest of the data is used to train the classifier. Subsequently, the held out observations are classified and the accuracy recorded. This number provides an estimate of how well the classifier will perform on unlabeled observations. In cross–validation, the training–test split is repeated many times, holding out a different set of observations at each iteration. The accuracy of the classifier is recorded at each iteration and then averaged. See Chapter 7 of Hastie et al. [2009] for more information on assessing classifier performance. Cross–validation has been the method of choice for evaluating classifier performance in many of the recent articles on classification of periodic variables.

4.3 Feature Distributions and Survey Systematics

The classification framework described above comes with assumptions and limitations. Of critical importance, statistical classification methods are only designed to produce accurate classifiers when the relationship between features and classes is the same in training and
unlabeled data. This is formalized as follows. Let \( z \) represent the class for a source with features \( x \). Let \( p_{tr}(z|x) \) be the probability of class given features in the training set and \( p_u(z|x) \) be the probability of class given features for unlabeled data. Statistical classifiers are designed to have high accuracy when \( p_{tr}(z|x) = p_u(z|x) \). In the three class example in the introduction, we saw that this was not the case due, in part, to incorrect estimation of periods in the unlabeled (OGLE) light curves. Violating this assumption will also cause cross-validation to make incorrect predictions of classifier accuracy.

In this section we illustrate the complex connection between survey systematics and feature distributions. We show how this connection causes the \( p_{tr}(z|x) = p_u(z|x) \) assumption to break, potentially leading to poor classifier performance on the unlabeled data.

### 4.3.1 Periodic Features

![Distribution of frequency](image1)

![Distribution of frequency](image2)

Figure 4.2: (a) Distribution of frequency (= period\(^{-1}\)) for three source classes observed for entire length of *Hipparcos*. (b) Distribution of frequency for same three sources classes observed for first 365 days of *Hipparcos*. A classifier constructed on the the complete *Hipparcos* light curves is likely to have poor performance on the *Hipparcos* curves truncated to 365 days. This scenario could happen if *Hipparcos* light curves were used to construct a classifier that was then applied to short light curves from the first Gaia data release at 1-2 years into the mission.

Nearly every study of classification of periodic variables has used period (or frequency) as a feature. Often in the training set, the period is correct for a large majority of sources
due to the investigators selecting the highest quality light curves of each source class of interest. However, if periods are estimated incorrectly for the unlabeled data, then a classifier constructed on the training data may not capture the period–class relationship as it exists for the unlabeled data.

For example, it has been suggested that light curves from early Gaia data releases be labeled using classifiers trained on Hipparcos light curves [Eyer et al., 2008, Eyer and et al., 2010]. Figure 4.2a shows a density plot of the estimated frequency for three source classes in Hipparcos\footnote{Sources used in Dubath et al. [2011]} using light curves from the entire 3.5-year survey. The median number of flux measurements per light curve is 91. However, one year into Hipparcos the densities of the estimated frequency for these source classes look significantly different (Figure 4.2b). The median number of flux measurements per light curve is now 29. Thus, even if we assume that Gaia and Hipparcos have similar survey characteristics, a classifier built on the 3.5-year baseline Hipparcos training set will not accurately capture the frequency–class relationship as it exists in 1-year Gaia data. This is due to incorrect estimates of frequency for the 1-year length light curves. Since it is often the case that many features depend on frequency (e.g., Table 4 of Richards et al. [2011] and Section 4.5 of Dubath et al. [2011]), systematic differences in estimates of frequency can alter the distributions of many features.

### 4.3.2 Time-Ordered Flux Measurements

Several recent studies of classification of periodic variables have used features that depend on the time ordering of flux measurements. For example, Dubath et al. [2011] used point–to–point scatter (P2PS), the median of absolute differences between adjacent flux measurements divided by the median absolute difference of flux measurements around the median. Specifically, given some light curve $x$ with time ordered flux measurements $m_0, \ldots, m_k$,

$$ \text{P2PS}(x) = \frac{\mathcal{M}(\{|m_i - m_{i-1}|\}_{i=1}^k)}{\mathcal{M}(\{|m_i - \mathcal{M}(\{m_j\}_{j=0}^k)\}_{i=0}^k)} \quad (4.1) $$

where $\mathcal{M}$ denotes the median. While potentially useful for classification, the behavior of this feature is heavily dependent on the cadence of time sampling. To see this, consider a two class problem where class 1 is sine waves of amplitude 1 with period drawn uniformly at random between 0.25 days and 0.75 days and class 2 is sine waves of amplitude 1 where period is drawn uniformly at random between 2 days and 8 days. Say we observe 20 flux measurements for each source. Figure 4.3 shows the density of P2PS for 200 sources of each class with (a) 30 minutes, (b) 2 days, and (c) 10 days between successive flux measurements. At 30 minutes and 2 days the feature is useful for distinguishing classes, but in opposite directions. At 10 days the feature is no longer useful.
Figure 4.3: Feature distributions can change dramatically with cadence. Plotted are the distributions of the P2PS feature (see Equation (4.1)) for two simulated classes observed at (a) 30 minute, (b) 2 day, and (c) 10 day cadences. A classifier trained on these light curves at one particular cadence may have poor performance when applied to light curves observed at a different cadence due to this change in feature distribution.

The process of how cadence and period produce the P2PS feature density is complex. For class 2 (2 day to 8 day periods) at 30 minute cadence, the flux measurements for each source are often monotonically increasing or decreasing, producing a small numerator relative to denominator in equation (4.1). When the cadence is large relative to the distribution of periods for the source class, the functional shape of the light curve determines the P2PS density. In Figure 4.3c where the cadence is longer than any possible period for either class, the two classes have the same density because they have the same functional shape (sine waves).

Note that this extreme sensitivity to cadence is not based on having 20 flux measurements per light curve. Running these simulations with 100 flux measurements per light curve produces densities of roughly the same shape. Rather, this example suggests how useful P2PS may be for distinguishing between classes in a setting where it may be difficult to determine a correct period (20 flux measurements per light curve), and how sensitive it is to systematic differences in cadence between training and unlabeled data.

4.3.3 Time-Independent Features

Finally, some of the most useful features for periodic variable classification are simple functions of flux measurements such as estimated amplitude, standard deviation, and skew. Figure 4.4 shows how estimated amplitude of Miras differs in distribution between the Hip-
CHAPTER 4. EMPIRICAL STUDY OF NOISIFICATION

Figure 4.4: Distribution of amplitude for Miras in OGLE and *Hipparcos*. *Hipparcos* Miras do not have very large amplitudes seen in some OGLE sources. The modes of the distributions are different as well.

*Hipparcos* and OGLE surveys. In *Hipparcos* there are no Miras with amplitude greater than 3 mag while roughly 12% of Miras in OGLE have amplitude greater than 3 mag. The mode of the densities is different as well.

There are several possible causes for the difference in shape of these densities. The median difference between last observation time and first observation time for OGLE sources is 1902 days and 1142 days for *Hipparcos*. Since Miras vary in amplitude through each period, it is possible that OGLE is simply observing more periods and picking up on lower troughs and higher peaks than *Hipparcos*. Additionally, many OGLE sources have large mean photometric error (not shown), which may be driving up estimates of amplitude. Also, OGLE and *Hipparcos* sources were observed with different filters, possibly leading to biases in estimated amplitude. Finally OGLE Miras reside in the Large Magellanic Cloud galaxy while *Hipparcos* sources are in the Milky Way, meaning the population of Miras that we are sampling from may have somewhat different properties.

It is also worth noting that the *Hipparcos* catalog light curves are themselves a composite of *Selected* sources chosen for their scientific interest before the mission and a set of *Survey* sources which represent a nearly complete sample to well defined magnitude limits (which depend spectral type and galactic latitude). Figure 4.5 shows boxplots of amplitudes in

---

5The *Hipparcos* Miras were used in Debosscher et al. [2007]. The OGLE sources are V-band data from OGLE III Catalog of Variable Stars: http://ogledb.astrouw.edu.pl/~ogle/CVS/
Figure 4.5: Distribution of amplitude by class for Survey and Selected sources in Hipparcos. The Selected sources have systematically larger amplitudes than the Survey sources. See Table 1 in Dubath et al. [2011] for definitions of these class abbreviations.

_Hipparcos_ for classes with over 50 sources, blocked into Survey and Selected. The Selected sources appear to have larger amplitudes on average than the Survey sources. A statistical classifier trained on this data will discover class boundaries for this mixture of Selected and Survey sources. However if the unlabeled data resemble the Survey sources, these boundaries may not separate classes well.

### 4.4 Noisification

We have shown how differences in survey systematics can alter feature distributions and deteriorate classifier performance. These survey systematics exist between and within surveys. In this section we describe noisification, our solution to addressing training–unlabeled set differences. We use noisification to overcome differences in training–unlabeled feature distributions caused by differences in the number of flux measurements, cadence, and level of photometric error of light curves. Before introducing noisification we discuss a few recent
works in the periodic variable classification literature that account for differences in training and unlabeled data and the extent to which they address distribution shifts discussed in Section 4.3.

### 4.4.1 Related Work

Two recent works, Richards et al. [2012] and Debosscher et al. [2009], have adapted classifiers to address training–unlabeled data set differences by adding unlabeled data to the training set. Richards et al. [2012] applied an active learning methodology to successfully improve classifier performance on ASAS unlabeled data using OGLE and *Hipparcos* training data. Debosscher et al. [2009] used a method similar to self-training [Nigam and Ghani, 2000] where after applying a classifier trained on *Hipparcos* and OGLE sources to CoRoT data, the most confidently labeled CoRoT sources were added to the training data. From this new training set, they constructed a classifier and used it to classify the remaining CoRoT sources.

Both active learning and self-training are designed to work when the feature densities in training and unlabeled data are different, but the feature–class relationship is the same. More formally, if \( p_{tr}(x) \) and \( p_u(x) \) are the feature densities in training and unlabeled data, then Active Learning and self-training are designed to address the setting where \( p_{tr}(x) \neq p_u(x) \), not \( p_{tr}(z|x) \neq p_u(z|x) \). However with our problem, differences in the number of flux measurements, cadence, and photometric error induce different relationships between class and features. For instance, consider the P2PS cadence example in §4.3.2, Figure 4.3. If the left plot, (a), is the training data P2PS class densities and the center plot, (b), is the unlabeled P2PS class densities, then moving data from (b) to (a) (as is done with Active Learning and self-training) would produce class densities that are a mixture of (a) and (b). Training a classifier on a mixture of (a) and (b) densities is unlikely to produce a classifier that has high accuracy on data with the classes densities in (b).

A method that comes closer to addressing class–feature distribution differences was used in Debosscher et al. [2009] to overcome aliasing in period estimation. There the authors found that the 13.97\(^{-1}\) day orbital frequency of the CoRoT mission caused spurious spectral peaks and induced incorrect period estimation for sources. Their solution was to disregard spectral peaks at the orbital frequency.

Effectively, Debosscher et al. [2009] asked the question “What would the value of this light curve’s period feature have been if it had been observed at a cadence matching the training data.” In their case, the answer is fairly straightforward. However it is much less clear how to correct other features in a similar manner. If the unlabeled sources are observed for 10 days, then it is likely that estimates of amplitude are biased. But by how much? If the source is a Mira, then likely by a lot, but if the source is an RR Lyrae possibly not at
all. So in order to correct amplitude estimates we need to know, or have some idea, of the class of the unlabeled source. But this returns to the goal of classification in the first place.

In Long et al. [2011] this approach was termed denoisification. For each unlabeled source the authors estimated a distribution across features representing uncertainty on what the feature values would have been if the source had been observed at a cadence, noise–level, and number of flux measurements in the training data. This distribution was combined with a classifier constructed on training data in order to classify unlabeled sources. While denoisification was superior to not adjusting for training–unlabeled distribution differences, the method did not achieve as large performance increases as noisification.

Noisification overcomes training–unlabeled set differences by altering the training set so that the number of flux measurements, cadence, and photometric error match that of the unlabeled data. A classifier can then use this “noisified” training data to determine class boundaries as they exist for the unlabeled data. Noisification was introduced in Starr et al. [2010]. Long et al. [2011] described a specific version of noisification appropriate for when training and unlabeled data have different numbers of flux measurements but are otherwise identical. Here we describe a far more general version of noisification which can be used across surveys when unlabeled sources have a systematically different number of flux measurements, cadence, and photometric error than the training data. Code written in Python and R is available for implementing noisification of light curves.\(^6\)

### 4.4.2 Implementation of Noisification

Given a set of training light curves, we first estimate a period for each.\(^7\) Next, we smooth the period folded light curves, turning each set of flux measurements into a continuous periodic function. Select a light curve \(x\) from the training set, and then at random choose a light curve, \(l\) from the unlabeled set. Let \(g\) be the smooth periodic function associated with \(x\). Let \(l_{i,1}, l_{i,2}, \text{ and } l_{i,3}\) represent the time, flux and photometric error for epoch \(i\) of light curve \(l\). Say there are \(m\) flux measurements for light curve \(l\). We now extract flux measurements from the periodic function \(g\) matching the cadence and photometric error present in \(l\). Specifically, if we let \(x_{i,1}, x_{i,2}, \text{ and } x_{i,3}\) be the time, flux, and photometric error of light curve \(x\) noisified to light curve \(l\), then we have,

\[
\begin{align*}
    x_{i,1} &= l_{i,1} \\
    x_{i,2} &= g(l_{i,1} + \alpha) + \epsilon_i \\
    x_{i,3} &= l_{i,3}
\end{align*}
\]

\(^6\)Code available here: [https://github.com/jpl2116/lightcurves](https://github.com/jpl2116/lightcurves)

\(^7\)Noisification assumes we have training sources that are of high enough quality that we can estimate periods accurately.
for $i \in \{1, \ldots, m\}$ where

$$
\epsilon_i \sim N(0, l_{i,3}^2)
$$

$$
\alpha \sim \text{Uniform}[0, p]
$$

$\alpha$ is a phase offset drawn uniformly at random between 0 and the period of $g, p$. This represents that fact that we are equally likely to start observing a source at any point in its phase. $\epsilon_i$ is the photometric error added to each flux measurement.

The cadence and level of photometric error in this new, noisified version of light curve $x$ now match that of the unlabeled data. Repeat this process for every training light curve. Then derive features for the noisified training data, train a classifier on these observations, and classify the unlabeled light curves using this classifier. We call this process noisification because if our training data consists only of well-sampled light curves and our unlabeled data consists mainly of poorly sampled light curves, then the technique effectively adds noise to features in the training data to more closely match the characteristics of the unlabeled features. See Figure 4.6 for a concise description of the algorithm.

<table>
<thead>
<tr>
<th>Noisification Algorithm</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. smooth training light curves, turning them into continuous periodic functions</td>
</tr>
<tr>
<td>2. extract flux measurements from these functions so that the number of flux measurements, cadence, and photometric error match the unlabeled data</td>
</tr>
<tr>
<td>3. derive features from these altered (noisified) training data light curves</td>
</tr>
<tr>
<td>4. construct a classifier using these light curve features</td>
</tr>
<tr>
<td>5. apply classifier to unlabeled sources</td>
</tr>
</tbody>
</table>

Figure 4.6: Description of the light curve noisification algorithm.

### 4.4.3 Remarks on Noisification

There are a few important points to note about this procedure. First, if the training and unlabeled data have the same cadence and photometric error, then smoothing the training light curves is not necessary. This would be the case, for example, if we had a set of training light curves of known class with many flux measurements ($\sim 100$) from one survey and we wanted to classify an unlabeled set of poorly sampled light curves ($\sim 30$ flux measurements) of similar cadence and photometric error level from the same survey as the training data.
Then we could simply take the training light curves, truncate them at 30 flux measurements, train a classifier on the truncated curves, and apply this classifier to the unlabeled light curves. This setting has the added benefit that no error will be introduced by smoothing the light curves. In this case the training sources do not need to be periodic.

Secondly, the procedure as described is most appropriate if all of the unlabeled data have similar numbers of flux measurements, cadence and photometric error. If this is not the case, then we can repeat the procedure several times using different subsets of the unlabeled data which share similar properties. For example, if unlabeled light curves have either around 20 or around 70 flux measurements, then we could break the unlabeled data into two sets and classify each set using a separate run of the noisification procedure. The more subsets of the unlabeled data one uses, the closer the noisified training data gets to the unlabeled data. The tradeoff is computational burden. With \( n \) training light curves and \( m \) unlabeled light curves, noisifying to precisely match the properties of each unlabeled light curve requires deriving features for \( nm \) light curves. In Section 4.5 we explore how much one can gain from dividing the unlabeled data into subsets.

With noisification, the unlabeled light curve, \( l \), at which to noisify training light curve \( x \), \( \alpha \) and \( \epsilon \) are all random. Thus, repeating the noisification process several times and obtaining several classifiers offers potential for improvement in classifier performance over running the process once. We study this in Section 4.5. While building several classifiers may be a good idea, it is important not to train a classifier using several noisified versions of the same light curve as the training data would no longer be independent. This can cause classifiers to overfit the data, hurting classifier performance.

Note that noisification is classifier independent. We use Random Forests in this work, but noisification can be used in conjunction with essentially any statistical classification method. Here we use Super Smoother for transforming training light curves into continuous periodic functions [Friedman, 1984]\(^8\). The method used for inferring continuous training curves is separate from the the rest of the noisification process. Splines and Nadaraya-Watson methods are other possibilities. Splines are described in 5.4 of Hastie et al. [2009]. See Hall [2008] for using Nadaraya-Watson with periodic variables.

Finally we stress that this implementation of noisification is limited to addressing differences between training and unlabeled sets caused by number of flux measurements, cadence, and photometric error. We do not correct for differences in feature distributions due to observing regions, detection limits, or filters.

\(^{8}\)Fortran code here: \texttt{http://www-stat.stanford.edu/~jhf/ftp/progs/supsmu.f}. We used automatic span selection (span= 0.0) and a high frequency penalty of \( \alpha = 1.0 \). These choices were based on visual inspection of smoothing fits to light curves.
CHAPTER 4. EMPIRICAL STUDY OF NOISIFICATION

102

Survey Source Classes\(^a\) F / LC\(^b\) # Train # Unlabeled

Simulated RR Lyrae, Cepheid, β Persei, β Lyrae, Mira 200-200 500 500

OGLE\(^c\) RR Lyrae DM, MM Cepheid, β Persei, β Lyrae, WU Majoris 261-474 358 165

Table 4.1: Light curves used in Sections 4.5.1 and 4.5.2.

\(^a\) In the case of the simulated data, the light curves were made to resemble these classes.

\(^b\) F / LC is the first and third quartiles of flux measurements per light curve for training.

\(^c\) We use every light curves of these classes analyzed in Richards et al. [2011].

4.5 Experiments

4.5.1 Noisification within a Survey

To get a sense how noisification performs in a controlled setting, we first test the method using training and unlabeled data from the same survey, but with systematically differing number of flux measurements. This resembles the real–life situation where well sampled light curves of known class are used as training data to classify poorly sampled curves of unknown class from the same survey. The cadence and levels of photometric error are assumed to match in the training and unlabeled data. We are also free from worrying about survey characteristics that noisification does not address. We perform two experiments, one using a simulated light curve data set and one using an OGLE light curve data set. See Table 4.1 for data set information.

After splitting each data set into training and unlabeled sets, we downsample the light curves in the unlabeled data set to 10 through 100 flux measurements in multiples of 10. Now the unlabeled data sets resemble the training in every way except for the number of flux measurements per light curve. To each of the ten unlabeled data sets we apply four classifiers and compute classification accuracy on the unlabeled data sets. Figure 4.7 provides error rates for the four classifiers applied to the 10 unlabeled sets from (a) simulated and (b) OGLE. The four classifiers are:

1. **naive** (black circles): Random Forest constructed on the unaltered training data

2. **unordered** (red triangles): noisify every training light curve by matching the number of flux measurements in the training set and unlabeled set, but we choose a random, non-contiguous set of epochs (cadence information is lost)

\(^9\) Here the OGLE curves are in I-band.
3. **1x noisification** (green plus): noisification without smoothing as described in Section 4.4

4. **5x noisification** (blue x) “1x noisification” repeated five times as discussed in Section 4.4

![Figure 4.7: Noisification results for (a) simulated light curves and (b) OGLE light curves. 5x Noisification (blue x) improves over making no adjustments for training–unlabeled data set differences (black circles) at all numbers of flux measurements.](image)

The results in Figure 4.7 suggest that noisification can significantly increase classification performance when the unlabeled data is poorly sampled. With OGLE, “naive” misclassifies around 32% of light curves with 30 flux measurements while “5x noisification” misclassifies around 21%. Based on the difference between the “unordered” and “1x / 5x noisification” procedures, it appears that having a training cadence that matches the cadence of the unlabeled data can improve classification performance. We explore this in more detail later when training and unlabeled data come from surveys with different cadences. The “5x noisification” advantage over “1x noisification” is fairly modest. Repeatedly noisifying the training data and averaging the resulting classifiers reduces variance and leaves bias unchanged, so we see no way that using “5x noisification” instead of “1x noisification” could hurt classifier performance. For the remainder of the chapter, noisification refers to “5x noisification.”
Figure 4.8: Variable importances for the OGLE “1x noisified” classifier on (a) 10 flux measurement and (b) 100 flux measurement training sets. When the training data has few flux measurements non-periodic features are most important because periods cannot be estimated correctly. See Section 4.2 of Dubath et al. [2011] for an explanation of feature importance.

To investigate how noisified classifiers differ, we plot feature importances for the “1x noisification” classifier for 10 and 100 flux measurements for the OGLE data (see Figure 4.8). Random Forest feature importance measures were introduced by Breiman [2001] and have been used in recent studies of periodic variables to gain an understanding of which features Random Forests considers most highly when assigning a class to a light curve. See Dubath et al. [2011] Section 4.1 for a complete description of feature importance. Figure 4.8 shows that skew is very important for both classifiers. Notice that the 100 flux measurement classifier ranks several period based features as being important – scatter_res_raw, freq_signif, and freq1_harmonics_freq_0 – while the 10 flux measurement classifier does not. The interpretation is clear: when classifying light curves with 10 flux measurements, features that require a correct period will not be very useful. The process of noisifying light curves causes the classifier to recognize this and make use of class information present in other features.

In these two examples, light curves in the unlabeled data set always had one of 10 possible number of flux measurements (10, 20, ... 100). The noisified light curves had exactly the same number of flux measurements as the unlabeled data. In practice, we will need to classify light curves with any number of flux measurements. It may be computationally challenging
CHAPTER 4. EMPIRICAL STUDY OF NOISIFICATION

Figure 4.9: The 10-point, 50-point, and 100-point noisified classifiers applied to all of the (a) simulated and (b) OGLE unlabeled sets. The 50-point and 100-point noisified classifiers perform well on all the unlabeled data sets with more than 30 flux measurements for both simulated and OGLE.

to construct noisified classifiers for every possible number of flux measurements. To test how sensitive error rates are to how light curves are noisified, we took the noisified classifiers for 10, 50, and 100 flux measurements and applied them across all 10 of the unlabeled data sets. Figure 4.9 shows the results for the (a) simulated and (b) OGLE data. We plot the error rates of these three classifiers along with the error rate of the classifier noisified to the number of flux measurements actually in the unlabeled data set (the “5x noisified” classifiers from Figure 4.7). The results show that for these examples the error rates are fairly insensitive to exactly how many flux measurements we use in the noisified classifier. For the OGLE data, the classifier noisified to 10 flux measurements performs well until unlabeled light curves have around 70 flux measurements. Additionally the 50–flux and 100–flux noisified classifiers perform well for unlabeled data sets with between 30 and 100 flux measurements.

4.5.2 Noisification with Smoothing

We now address the challenge of training a classifier on a survey with one cadence to classify light curves of a different cadence. In order to ensure that all differences between training and unlabeled data are due to issues addressed by noisification (number of flux measurements,
cadence, photometric error) we use the simulated light curve prototypes from Section 4.5.1 for both training and unlabeled data sets. We sample the light curves at actual *Hipparcos* and OGLE light curve cadences used in previous studies [Richards et al., 2011, Debosscher et al., 2007].

Systematic differences exist between the OGLE and *Hipparcos* survey cadences. OGLE is a ground based survey with flux measurements taken at multiples of one day plus or minus a few hours. The sampling for these curves is quite regular with few large gaps. In contrast, *Hipparcos* light curves tend to be sampled in bursts, with several measurements over the course of less than a day followed by long gaps.

In practice, one data set (say, *Hipparcos*) would be used to train a classifier in order to classify sources in the other (say, OGLE). However since these light curves are simulated, and we have labels for both sets, we create training and unlabeled data sets at *Hipparcos* and OGLE cadences so we can study the challenge of constructing a classifier on *Hipparcos* for use on OGLE sources and vice versa. We begin by generating 1000 simulated light curves using the class templates from Section 4.5.1. For 500 of these curves we randomly select an OGLE cadence and sample flux measurements and photometric errors from this cadence. We then take these 500 curves and downsample them to have 10,...,100 flux measurements in multiples of 10. The original 500 curves cadenced to OGLE is the OGLE training set, and the downsampled curves are the 10 OGLE unlabeled data sets. We repeat this process for the other 500 simulated curves at *Hipparcos* cadences.

In order to test the efficacy and necessity of various aspects of the noisification process, we apply several classifiers to each of the unlabeled data sets. Figure 4.10 shows the accuracy of these methods treating (a) OGLE and (b) *Hipparcos* as the unlabeled data. For the left plot with OGLE unlabeled light curves the classifiers are trained on:

1. **OGLE cadence naive** (black circle): unaltered OGLE light curves

2. **Hipparcos cadence noisified** (red triangle): *Hipparcos* light curves truncated to match length of unlabeled set, but not smoothed (cadence is different between training and unlabeled)

3. **Hipparcos smoothed to OGLE – noisified** (green plus): *Hipparcos* light curves after they have been smoothed, cadenced at OGLE, and truncated to match length of unlabeled curves

4. **OGLE cadence noisified** (dark blue x): noisified OGLE light curves (cadence already matches unlabeled set so smoothing unnecessary)

5. **Hipparcos naive** (light blue diamonds): unaltered *Hipparcos* light curves
Not addressing cadence, flux measurement, and photometric error mismatches by training on full length Hipparcos light curves leads to poor performance (light blue diamond). Noisifying these Hipparcos sources by truncation improves performance (red diamonds). However we gain significantly by correcting for cadence differences by smoothing (green plus). It is encouraging to see that by smoothing the Hipparcos training set and noisifying we can do as well as if we had started with OGLE cadence curves (dark blue x and green plus).

The right plot of Figure 4.10 displays the same information with Hipparcos as the unlabeled cadence. Note that the line markings have been changed to preserve relationship of training set to unlabeled set. The overall picture is similar to the OGLE data, except that convergence of error rates happens much more quickly. At 60 flux measurements there is little difference among any of the classifiers.

Figure 4.10: Simulated light curves where the unlabeled data is observed at a (a) OGLE or (b) Hipparcos cadence. By smoothing the training set and extracting flux measurements to match that of the unlabeled data (green plus), we improve performance over only matching number of flux measurements (red triangle).

The difference in error rates between classifiers trained on data noisified to the cadence of the unlabeled data and those that are not suggests that at low number of flux measurements feature distributions are different for the OGLE and Hipparcos cadences. To investigate this in Figure 4.11 we plot densities of amplitude for simulated light curves with 10 flux measurements at the OGLE and Hipparcos cadences. To keep things simple we show two
class densities – Miras and not Miras. It is clear here that for the OGLE cadence amplitude
is not a particularly useful feature for separating Miras from other sources whereas for the
Hipparcos cadence it is. Due to the regular sampling at one to two day intervals, 10 flux
measurement OGLE curves have only captured part of a Mira period. Hence the amplitude
of the source looks much smaller than it actually is. In contrast the large gaps between
flux measurements in Hipparcos cadences result in us observing a much larger piece of phase
space and thus obtaining a better estimate of amplitude.

![Graphs showing amplitude feature distributions for Mira versus other classes for 10 flux mea-
surements at (a) OGLE and (b) Hipparcos cadence. The feature is very useful for separating
Miras from non-Miras at the Hipparcos cadence because of the irregular time sampling. Here
we see how important it is for training and unlabeled data to have matching cadences, not
just number of flux measurements.]

Figure 4.11: Amplitude feature distributions for Mira versus other classes for 10 flux mea-
surements at (a) OGLE and (b) Hipparcos cadence. The feature is very useful for separating
Miras from non-Miras at the Hipparcos cadence because of the irregular time sampling. Here
we see how important it is for training and unlabeled data to have matching cadences, not
just number of flux measurements.

4.5.3 Using Hipparcos to Classify OGLE

Now that we have studied noisification in some controlled settings, we test the method on
the original problem proposed in Section 4.1. Recall that we are classifying Miras, RR
Lyrae AB, and Classical Cepheids Fundamental Mode using light curves from Hipparcos as
the training data and V-band OGLE as the unlabeled data. In Section 4.1 we saw that
training a classifier on the Hipparcos curves and applying it directly to OGLE resulted in
poor performance due, in part, to differences in number of flux measurements, cadence, and photometric error between the two data sets.

Table 4.2 highlights some important differences between the Hipparcos and V-band OGLE sources. See Udalski et al. [2008b], Soszynski et al. [2008, 2009b,a] for descriptions of OGLE III photometry and these three source classes. We use all OGLE III sources from the LMC belonging to the three classes of interest.

There are systematically fewer flux measurements in OGLE than in Hipparcos. Unlike the previous example with I-band OGLE, the V-band OGLE curves here are fairly sparse. 25% percent of the flux measurements are spaced 16 or more days apart. Perhaps the most striking difference between surveys is in the class proportions. RR Lyrae AB make up 26.6% of light curves in Hipparcos and 84.1% of light curves in OGLE. This is most likely due to Hipparcos magnitude limits which result in undersampling the intrinsically faint RR Lyrae AB relative to Mira and Classical Cepheids AB.

To classify the OGLE sources, we noisify all the Hipparcos light curves to OGLE cadence at 10 through 100 flux measurements in multiples of 10. We then construct classifiers on each of these sets, resulting in 10 noisified classifiers. Each OGLE light curve is classified using the classifier with the closest number of flux measurements. So for an unlabeled OGLE light curve with 27 flux measurements, we classify it using the noisified classifier constructed on the 30-flux measurement training set.

Table 4.3 displays a confusion matrix for the classifier constructed on the unmodified Hipparcos light curves when it is applied to the OGLE light curves. Table 4.4 shows the error rate using the noisification procedure. The overall error rate drops from 27% to 7% as a result of using noisification. This is driven by the drop in error rate for RR Lyrae AB (31% error using unmodified classifier, 7% after noisification) and the prevalence of RR Lyrae AB in OGLE. The error rate for Classical Cepheids F actually increases from 2% to 10% while

\[\text{Table 4.2: Training and unlabeled set characteristics for example in Section 4.1 and Subsection 4.5.3.}\]

\[
\begin{array}{|c|c|c|c|c|c|}
\hline
\text{Survey} & \text{# Sources} & \text{Class Probs.} & \text{F / LC} & \text{Time Diff} & \text{Error} \\
\hline
\text{Hipparcos}^{(a)} \text{ (training)} & 357 & (0.45,0.27,0.28) & 81-119 & 0.01-0.25 & 0.015-0.034 \\
\text{OGLE (unlabeled)} & 20605 & (0.09,0.84,0.07) & 36-74 & 5.1-16.0 & 0.022-0.050 \\
\hline
\end{array}
\]

\[a\text{ Class probs. is the class proportion of (Classical Cepheids F, RR Lyrae AB, Mira).}\]
\[b\text{ F / LC is the first and third quartiles of flux measurements per light curve for training.}\]
\[c\text{ Time Diff is the first and third quartiles of time differences in days between successive flux measurements.}\]
\[d\text{ Error is the first and third quartiles of estimated photometric error in magnitude for all flux measurements.}\]
\[e\text{ Light curves and classifications from Richards et al. [2011].}\]

\[10\text{These OGLE III sources are available here: http://ogledb.astrouw.edu.pl/~ogle/CVS/}.\]
for Miras it is roughly the same.

<table>
<thead>
<tr>
<th>Predicted</th>
<th>ClsC</th>
<th>Mira</th>
<th>RRLAB</th>
<th>True</th>
<th>ClsC</th>
<th>Mira</th>
<th>RRLAB</th>
<th>True</th>
<th>ClsC</th>
<th>Mira</th>
<th>RRLAB</th>
<th>True</th>
</tr>
</thead>
<tbody>
<tr>
<td>ClsC</td>
<td>1799</td>
<td>0</td>
<td>34</td>
<td>0.02</td>
<td>1644</td>
<td>1</td>
<td>188</td>
<td>0.1</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mira</td>
<td>58</td>
<td>1360</td>
<td>20</td>
<td>0.05</td>
<td>18</td>
<td>1381</td>
<td>39</td>
<td>0.04</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RRLA</td>
<td>5358</td>
<td>78</td>
<td>11898</td>
<td>0.31</td>
<td>1168</td>
<td>76</td>
<td>16090</td>
<td>0.07</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Err.Rate</td>
<td>0.75</td>
<td>0.05</td>
<td>0</td>
<td>0.27</td>
<td>0.42</td>
<td>0.05</td>
<td>0.01</td>
<td>0.07</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 4.3: Confusion matrix for classifier constructed on the unmodified Hipparcos light curves and applied to OGLE. Rows are true class and columns are predictions. The overall error rate is driven by the performance on the most abundant class, RR Lyrae AB.

Table 4.4: Confusion matrix for classifier constructed on noisified Hipparcos light curves. Rows are true class and columns are predictions. The overall error rate has dropped to .07 from .26. This is due to better predicting RR Lyrae AB sources. The error rate on Classical Cepheids has actually increased.

Part of the reason why noisification increases the error rate for Classical Cepheids appears due to differences in distribution of frequency caused by Hipparcos magnitude limits. Figure 4.12 displays frequency density in *Hipparcos*, 35-45 flux length OGLE, and *Hipparcos* noisified to 40 flux for Cepheids (4.12a), RR Lyrae (4.12b), and Miras (4.12c). Noisification has not changed the density at all for the Cepheid sources (the blue and orange density almost exactly overlap) for the Cepheids. Visual inspection of OGLE periods revealed that they were correct. This suggests that the frequency distribution for Cepheids is fundamentally different in *Hipparcos* and OGLE. This is likely due to magnitude limits in *Hipparcos* and OGLE.

Lower frequency Cepheids are intrinsically brighter, so we can see them from further away. These low frequency Cepheids are over-represented in *Hipparcos*. In contrast OGLE is closer to a random sample of Cepheids in the Large Magellanic Cloud (LMC). If it is there, we see it. Since this survey difference is not caused by number of flux measurements, cadence,
or photometric error, the current implementation of noisification does not correct for it. Notice that in Figure 4.12 right plot, the noisification procedure has shifted the distribution of RR Lyrae frequencies in *Hipparcos* to more closely match that in OGLE. Here much of the density mismatch was due to error in estimation of frequency due to having few flux measurements. Noisification helps us overcome this survey difference.

Figure 4.12: Density of frequency in OGLE light curves with 35–45 flux measurements (black solid), *Hipparcos* before noisification (blue dots) and after noisification to 40 flux measurements (orange dashed) for (a) Classical Cepheids F, (b) RR Lyrae AB, and (c) Miras. Noisification of Cepheid periods does not match training and unlabeled densities because of survey differences not related to number of flux measurements, cadence or photometric error.

Noisification is successful at matching other feature distributions. Figure 4.13 displays the densities of P2PS for each sources class in 4.13a *Hipparcos*, 4.13b OGLE, and 4.13c *Hipparcos* noisified. There is a great deal of difference between *Hipparcos* and OGLE densities. However the noisified *Hipparcos* source densities appear to closely match the densities of OGLE.

4.6 Conclusions

We have highlighted how differences between training and unlabeled light curves induce different feature distributions. We then showed how these shifts in distribution can cause high error rates, even on problems where the unlabeled data is well separated in feature space. Common methods to evaluate classifier performance, such as cross–validation, do not detect these shifts in distribution and may give a false impression of classifier quality as they only reveal how well a classifier performs on data that is similar to the training set.

We developed a methodology, noisification, for overcoming differences between training and unlabeled data sets. As implemented in this study, noisification addresses differences
Figure 4.13: (a) P2PS in Hipparcos un-noisified data. The feature appears useful for separating RR Lyrae from Miras and Classical Cepheids. (b) P2PS in OGLE for light curves with 35–45 flux measurements. Now Classical Cepheids have nearly the same density as RR Lyrae. A classifier trained on the un-noisified Hipparcos light curves would not capture this relationship. (c) P2PS for Hipparcos light curves noisified to OGLE cadence with 40 flux measurements. The densities now closely resemble the OGLE light curves.

due to the number of flux measurements, cadence, and photometric error. On several simulated and real-world examples, noisification greatly improved classifier performance. In the Hipparcos training–OGLE unlabeled example, noisification reduced the misclassification rate by 20%.

We hope these findings motivate practitioners to carefully consider differences between training and unlabeled data sets. In general, we recommend using training sets that match as closely as possible the unlabeled set of interest rather than training sets that are high signal–to–noise. As demonstrated in many examples, high signal–to–noise light curves often work poorly as training sets when the unlabeled light curves are of low quality. This is due to the classifier discovering class boundaries in feature space as they exist in the training set, not as they exist in the unlabeled set.

This study has made us skeptical of attempts to identify a single set of features that is generically sufficient for separating a set of classes of periodic variables. Useful features change depending on how sources are observed. The Random Forest importance plots (Figure 4.8) and the P2PS simulation (Subsection 4.3.2) illustrate this. When implementing noisification, we recommend starting with large feature sets, even including features that are not useful for separating classes in the training data. These features may become useful for separating classes once the light curves have been noisified.

While we have studied noisification in the context of classification, it could also be applied
to other problems. For example, novelty detection and unsupervised learning (clustering) methods are likely to work poorly when training and unlabeled data sets have systematic differences. Noisifying light curves offers a way to overcome these differences.

Noisification may also be extended from what is implemented here to account for differences not related to number of flux measurements, cadence, and level of photometric error. For example, known censoring thresholds in the unlabeled data could be incorporated into the training data by removing, or marking as censored, flux measurements which would not have been observed in the unlabeled data set due to magnitude limits.

In the future, we will apply noisification to light curves from more surveys using larger, highly multi-class training sets. In parallel, we are developing a theoretical understanding of how noisification works and the problems for which it is most suitable. Of particular interest is how noisification performs when there are survey differences not addressed by noisification. This was the case with the Cepheid frequencies in the three-class Hipparcos–OGLE problem.

Upcoming surveys pose a challenge based in their size and their novelty. Not only will Gaia and LSST detect orders of magnitude more periodic variables than previous surveys, the sources they collect will have different properties than any training data we currently have. Noisification offers the potential to bridge some of these differences, enabling us to optimize scientific discovery.

4.7 Description of Features

We used 62 features in this work. Fifty of these features came from Tables 4 and 5 in Richards et al. [2011]. We did not use the features pair_slope_trend, max_slope, or linear_trend from these tables. We used 12 additional features. Five are from Dubath et al. [2011]. The remaining seven are:

1. **fold2P_slope_10percentile** 10th percentile of slopes between adjacent flux measurements after the light curve has been folded on twice the estimated period

2. **fold2P_slope_90percentile** 90th percentile of slopes between adjacent flux measurements after the light curve has been folded on twice the estimated period

3. **freq_frequency_ratio_21** ratio of the second to first frequency determined by lomb-scargle (\(\frac{f_2}{f_1}\) from Table 4 in Richards et al. [2011])

4. **freq_frequency_ratio_31** ratio of the third to first frequency determined by lomb-scargle (\(\frac{f_3}{f_1}\) from Table 4 in Richards et al. [2011])

\[^{11}\text{scatter_res_raw, medperc90_2p_p, p2p_scatter_2p raw, P2PS}\ (\text{named P2p_scatter in Dubath et al. [2011]}), \text{and p2p_scatter pfold over mad}\]
5. freq_amplitude_ratio_21 ratio of amplitude for frequency 2 to amplitude for frequency 1 ($\frac{A_{2,1}}{A_{1,1}}$ from Table 4 in Richards et al. [2011])

6. freq_amplitude_ratio_31 ratio of amplitude for frequency 3 to amplitude for frequency 1 ($\frac{A_{3,1}}{A_{1,1}}$ from Table 4 in Richards et al. [2011])

7. p2p_ssqr_diff_over_var\textsuperscript{12} the sum of squared differences in successive flux measurements divided by the variance of the flux measurements

\textsuperscript{12}From Kim et al. [2011]
Bibliography


L. Eyer and J. Cuypers. Predictions on the number of variable stars for the gaia space mission and for surveys such as the ground-based international liquid mirror telescope. In *IAU Colloq. 176: The Impact of Large-Scale Surveys on Pulsating Star Research*, volume 203, pages 71–72, 2000.


BIBLIOGRAPHY


