The Pennsylvania State University
The Graduate School

ASSESSING CLASSIFICATION UNCERTAINTY ON ASTRONOMICAL OBJECTS WITH MEASUREMENT ERROR

A Thesis in Statistics
by
Sarah Shy

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Master of Science

May 2022

The thesis of Sarah Shy was reviewed and approved by the following:

Hyungsuk Tak
Assistant Professor of Statistics, Astronomy and Astrophysics
Thesis Advisor

Eric D. Feigelson
Distinguished Senior Scholar and Professor of Astronomy and Astrophysics
Professor of Statistics

G. Jogesh Babu
Distinguished Professor of Statistics, Astronomy and Astrophysics

Ephraim M. Hanks
Associate Professor of Statistics
Chair of Graduate Studies

Abstract

Most general-purpose classification methods, such as support-vector machine (SVM) and random forest (RF), fail to account for an unusual characteristic of astronomical data: known measurement error uncertainties. In astronomical data, this information is often given in the data but discarded because popular machine learning classifiers cannot incorporate it. We propose a simulation-based approach that incorporates heteroscedastic measurement error into any existing classification method to better quantify uncertainty in classification. The proposed method first simulates perturbed realizations of the data from a Bayesian posterior predictive distribution of a Gaussian measurement error model. Then, a chosen classifier is fit to each simulation. The variation across the simulations naturally reflects the uncertainty propagated from the measurement errors in both labeled and unlabeled data sets. We demonstrate the use of this approach via two numerical studies. The first is a thorough simulation study applying the proposed procedure to SVM and RF, which are well-known hard and soft classifiers, respectively. The second study is a realistic classification problem of identifying high-z ($2.9 \leq z \leq 5.1$) quasar candidates from photometric data. The data were obtained from merged catalogs of the Sloan Digital Sky Survey, the *Spitzer* IRAC Equatorial Survey, and the *Spitzer*-HETDEX Exploratory Large-Area Survey. The proposed approach reveals that out of 11,847 high-z quasar candidates identified by a random forest without incorporating measurement error, 3,146 are potential misclassifications. Additionally, out of $\sim$1.85 million objects not identified as high-z quasars without measurement error, 936 can be considered candidates when measurement error is taken into account.

Table of Contents

List of Figures **vi**

List of Tables **viii**

Acknowledgments **ix**

Chapter 1
Introduction **1**
1.1 Past Studies 1
1.1.1 Measurement Error Studies 1
1.1.2 Data Simulation Approaches 2
1.2 Our Contributions 3
1.3 Computation and software 5

Chapter 2
Data for classification in astronomy **6**

Chapter 3
Methodology **8**
3.1 Quantifying Overall Uncertainty of a Classifier 10
3.2 Prediction on the Unlabeled Set 11

Chapter 4
Simulation Study **14**
4.1 Support-vector machine 15
4.2 Random Forest 18

Chapter 5
Classifying High-Redshift Quasars **21**
5.1 The Labeled and Unlabeled Data Sets 21
5.2 Classification Accuracy on the Labeled Set 22
5.3 Prediction on the Unlabeled Set via Gaussian Perturbation 24

Chapter 6
Discussion **26**
6.1 Why Bayesian posterior predictive distribution? 26
6.2 The number of perturbed data sets 27
6.3 Correlated measurement error 28
6.4 Gaussian perturbation for clustering 29
6.5 Limitations 29
6.6 Concluding Remarks 30

Bibliography **31**

List of Figures

1.1 The proposed framework uses the original observed data to simulate multiple perturbed data sets. Each perturbed data set is independently drawn from the posterior predictive distribution of a Gaussian measurement error model. We re-fit the classifier to each perturbed data set and calculate any metrics of interest, such as prediction accuracy. This will produce a posterior predictive distribution for each metric, from which we can assess uncertainty by the spread of the distribution. Consequently, the measurement error uncertainty in the original data is propagated through each step in the framework. 3

1.2 Left: For a single observation, a hard classifier outputs a single predicted class label, i.e., the object is predicted to be in that class with probability 1. Center: A soft classifier reports the relative probability of falling into each class. Right: A further softened soft classifier assesses the variability in the class probabilities. The proposed method converts a hard classifier into a soft one (left to center) and further softens soft classifiers (center to right). 4

2.1 In a typical classification problem, a classifier is trained and validated on the labeled set, and then used to make predictions on the unlabeled set. In astronomy, the information about measurement error uncertainty is additionally given in the data set, as illustrated in the right panel. 6

3.1 The perturbed data sets simulated by a posterior predictive distribution of a measurement error model. The information about the measurement error is used only for simulation, and does not appear in the perturbed data sets. Thus, any traditional classification method can be trained on each perturbed data set independently, and the variation of the multiple fits can be ascribed to the measurement error uncertainty. 10

4.1 Left: the true unobserved data. Center: the observed data without considering measurement error, suitable for standard classification methods. Right: the observed data with error bars whose lengths represent 2 standard deviations of measurement error uncertainties (i.e., measurement $\pm 1\sigma$), which the proposed framework utilizes to simulate perturbed data sets. 15

4.2 Two observations in the unlabeled set are superimposed on the 200 observations in the labeled set. 16

4.3 (a) Linear SVM boundary fit to the observed data without accounting for measurement error, colored by the predicted labels. The probability of being in class 1 is 1 for blue dots and 0 for yellow dots. The red line is the fitted decision boundary of the linear SVM. (b) 500 SVM decision boundaries obtained by fitting the SVM on each of the 500 perturbed data sets, are colored in gray. The red line is reproduced from the left panel for a comparison. The estimated probability of belonging to class 1, i.e, $\hat{p}_{i,1}$ for each object i is denoted by a blue-green-yellow color gradient; darker dots are more likely to belong to class 1. Both the training and test sets are plotted, but each decision boundary was derived using the training set only. The length of each error bar is 2 standard deviations of measurement error (i.e., measurement $\pm 1\sigma$). 16

4.4 (a) The probability estimate of being in Class 1 computed from a one-time implementation of random forest without measurement error ($\hat{p}_{i1}$ for each object i, defined in Equation (3.5)) is visualized via color gradation. (b) The probability estimate calculated by the proposed approach that accounts for measurement error ($\hat{p}^{+}_{i1}$ for each object i, defined in Equation (3.6)) is visualized using the same color gradation. Darker dots are more likely to belong to class 1. 19

5.1 Posterior predictive distributions of completeness (left) and efficiency (right), respectively, obtained by Gaussian perturbation with 500 simulations. The mean and standard deviations are superimposed in each panel. The vertical dashed line indicates the value of completeness or that of efficiency obtained without considering measurement error. 23

List of Tables

4.1 Comparison of overall classification accuracy and decision boundary parameters of a linear SVM computed from the labeled set with and without incorporating measurement error uncertainties. The vertical dashed lines in the posterior predictive distributions indicate the values obtained from the SVM without measurement error. 18

5.1 The confusion matrices obtained (a) without and (b) with measurement error. For case (b) the outcome is the average of 500 confusion matrices, and the values in the parentheses are standard deviations. The acronym 'AE' denotes the class 'Anything Else'. 24

5.2 Overlap of label predictions on the prediction set between a single random forest and an ensemble random forest using $R = 500$ simulations. The approach without measurement error is denoted by 'Standard' and the one with measurement error by 'Proposed'. For example, 936 objects are predicted to be high-z quasars with measurement error (Proposed), but to be AE without measurement error (Standard). Also, 3,146 objects are classified as high-z quasars without measurement error (Standard) but as AE with measurement error (Proposed). 24

Acknowledgments

First and foremost, I sincerely thank my advisor, Dr. Tak, whose expertise and mentorship provided invaluable guidance throughout this work. Your professionalism and patience are inspiring, and I am grateful for the opportunity to learn under your guidance. I gratefully recognize Dr. Feigelson, Dr. Timlin, and Dr. Babu for their support and contributions to every step of the project, from formulation to execution. It is a pleasure having you all as my colleagues. I also thank Jackeline Moreno and Weixiang Yu for the helpful discussions about this work early on. Finally, my heartfelt thanks go to my family and friends for their continued support and love.

Chapter 1 Introduction

Classification methods have played an important role in astronomy for centuries and continue to be essential tools in modern astronomy (Feigelson 2012, chap. 9). According to the Astrophysics Data System, the keyword 'classification' appeared in over 6000 refereed astronomy papers annually for the last 5 consecutive years (2017–2021). One critical methodological issue is that most standard classification methods do not have built-in functionality to account for measurement error, a notable property of astronomical data. As a result, to implement these classification methods, we have no choice but to assume that the data are error-free measurements. This assumption inevitably produces over-confident outcomes. It is therefore essential to account for measurement error in classification analyses to reliably classify astronomical objects.

1.1 Past Studies

1.1.1 Measurement Error Studies

Incorporating measurement error into classification analysis is an active area of research across various scientific fields. However, classical statistical treatments of measurement error are, by and large, limited to regression problems, and homoscedasticity is often assumed (Fuller 1987, Carroll et al. 2006, Buonaccorsi 2010). The most common approach in machine learning is to use a weighting scheme, applying heavier weights to more certain measurements (Hoefsloot et al. 2006, Lapin et al. 2014, Hashemi & Karimi 2018, Luo 2019). The inverse of the error variance is often used for the weight, as in the case of classical weighted linear regression. However, these solutions are method-specific, requiring a unique modification to each classification method. Moreover, they consider measurement error in the training set only. Existing method-agnostic approaches to

account for heteroscedastic measurement error are limited to scaling and transforming the raw data according to the measurement error variance (Waaijenborg et al. 2018). Common transformations include the square root and log transformation (van den Berg et al. 2006). Such transformations assume a specific, strictly monotonic, relationship between measurement intensity and measurement error variance. A more general approach that applies to any error structure is needed.

In astronomical literature specifically, the gap in statistical methodology for incorporating heteroscedastic measurement error has been noted for many years (Feigelson & Babu 1998). Various astronomical contributions have been made dating back to Eddington (1913), where the influence of small observational error on flux measurements was first investigated. Petrosian (1992), Kogan (1997), and Hogg & Turner (1998) similarly discuss measurement error on univariate measurements. In regression, Akritas & Bershady (1996), Kelly (2007), Andreon & Hurn (2013), and Sereno (2016) present notable contributions in both classical regression and Bayesian approaches. In time-domain astronomy, the heteroscedastic measurement error has been properly modeled in univariate and multivariate damped random walk processes (Kelly et al. 2009, Hu & Tak 2020) and the more general univariate CARMA(p, q) process (Kelly et al. 2014). However, the disconnect between the astronomical need for incorporating measurement error and the existing statistical methods for doing so is still ongoing for many other data analyses (Feigelson et al. 2021).

1.1.2 Data Simulation Approaches

Simulation-based approaches, such as bootstrapping (Efron 1992, Efron & Tibshirani 1994), subsampling (Von Luxburg et al. 2010), and random projections on a lower dimensional space (Achlioptas 2003) are widely used in statistical machine learning. For example, (i) Yu et al. (2013) and Sun (2015) review various ways to simulate data sets through bootstrapping, subsampling, and adding random noise in the context of classification stability and reproducibility. (ii) Cannings (2021) uses random projections to create simulated data sets in lower dimensions and aggregates the results to produce an ensemble classifier with better predictive power. (iii) Malossini et al. (2006) detect potential labeling errors by flipping one label at a time and assessing the sensitivity of the results to a single flipped observation. (iv) Darling & Stracuzzi (2018) characterize the distribution of simulated model outputs for decision trees and k-nearest neighbors by bootstrapping and sampling-without-replacement. However, these works do not incorporate *known* measurement error information. In addition, measurement error in

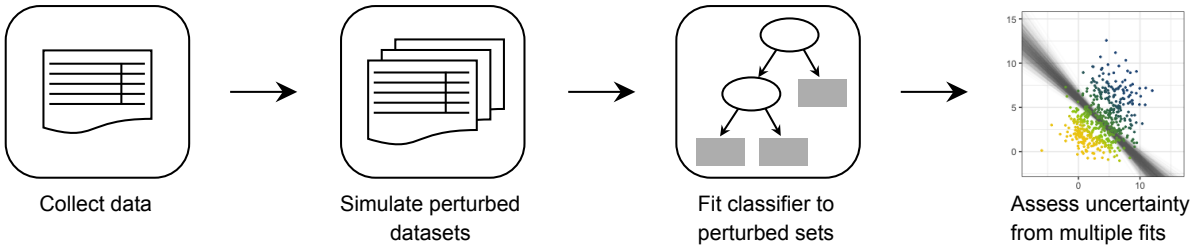


Figure 1.1: The proposed framework uses the original observed data to simulate multiple perturbed data sets. Each perturbed data set is independently drawn from the posterior predictive distribution of a Gaussian measurement error model. We re-fit the classifier to each perturbed data set and calculate any metrics of interest, such as prediction accuracy. This will produce a posterior predictive distribution for each metric, from which we can assess uncertainty by the spread of the distribution. Consequently, the measurement error uncertainty in the original data is propagated through each step in the framework.

the unlabeled data (whose labels are to be predicted) is not considered. Thus, these methods are still inadequate for a large array of classification problems in astronomy.

1.2 Our Contributions

To overcome the limitations of existing methods, we propose a simulation-based approach that incorporates heteroscedastic measurement error into any classification algorithm. In principle, the proposed can quantify the uncertainty of *any* quantity of interest relating to the classification results. First, we simulate multiple pseudo-data sets using the known measurement error uncertainties with minimal assumptions on a Gaussian measurement error model. Second, we fit a classifier on each simulation, using the same method each time. Finally, from the ensemble of classifications, we quantify classification uncertainty by the variation across the fits.

Consequently, our approach propagates uncertainties of heteroscedastic measurement error from both the labeled and unlabeled datasets to the final classification results. Figure 1.1 illustrates this process.

We note that the proposed method is not intended to improve classification accuracy. Classification cannot be more accurate after incorporating additional uncertainty than if we ignore measurement error. Instead, the goal is to present more reliable and informative classification results by incorporating uncertainty from measurement error in a principled way.

We illustrate the proposed approach using two popular classification methods: support-vector machine and random forest (hereafter SVM and RF, respectively). The former

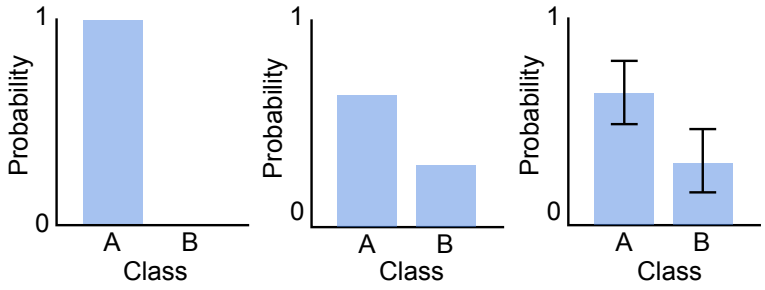


Figure 1.2: Left: For a single observation, a hard classifier outputs a single predicted class label, i.e., the object is predicted to be in that class with probability 1. Center: A soft classifier reports the relative probability of falling into each class. Right: A further softened soft classifier assesses the variability in the class probabilities. The proposed method converts a hard classifier into a soft one (left to center) and further softens soft classifiers (center to right).

is known as a hard classifier because the final classification result is a single predicted class label, i.e., one predicted class with probability 1 (Wahba 2002). The latter is a soft classifier in the sense that for every object, it provides a relative probability (estimate) of belonging to each class. The proposed approach absorbs the known measurement error uncertainties to soften hard classifiers and further soften soft classifiers. Softening hard classifiers means that the information about the probability of belonging to each class, which is naturally given by soft classifiers, becomes available to hard classifiers. Softening soft classifiers means that we are able to assess the variability of the class probabilities provided by soft classifiers. This concept is illustrated in Figure 1.2. We note that the applicability of the proposed approach is not restricted to these two specific classifiers but can be applied to any classifier, including deep learning neural networks.

To better present the advantages of the proposed approach, we conduct a simulation study and a realistic classification analysis of high-redshift ($2.9 \leq z \leq 5.1$) quasars. Both studies show that classification without considering measurement error produces over-confident results when classifying individual objects into two classes. For example, an object is confidently classified into one class even when its measurement error bar spans across the classification boundary. The proposed approach captures this uncertainty, showing that the object is not classified into one class dominantly across all simulations. In the high-z quasar classification, random forest identifies 11,847 high-z quasar candidates out of ~ 1.86 million objects when measurement error is not considered. On the other hand, the proposed approach reveals that 3,146 out of the 11,847 are potential misclassifications when their measurement errors are considered. In addition, of the ~ 1.85 million objects not identified to be high-z quasars without considering measurement error, 936 should be considered potential candidates once we account for

measurement error. These results are based on a simple decision threshold of 0.5, i.e., we have classified an object as a high-z quasar if its estimated probability of being a high-z quasar were greater than 0.5.

The rest of this thesis is organized as follows. In Chapter 2, we briefly review the data format and notation for classification. In Chapter 3, we specify the details of the proposed approach as a generic way to incorporate heteroscedastic measurement error into any standard classification method. Chapter 4 presents a thorough simulation study demonstrating how the proposed procedure can be applied to SVM and RF. In Chapter 5, we present a realistic astronomical application to identify high-z quasar candidates. Finally, we discuss potential limitations of the proposed work and future directions in Chapter 6.

1.3 Computation and software

Computations for this study were made with the R statistical programming language (R Core Team 2020). Random forest and support-vector machine classification were performed with CRAN packages *caret* (Kuhn 2021) and *e1071*. Several other CRAN packages (*dplyr, data.table, magrittr, tictoc, foreach, logger, ggplot2, ggpubr*) were used for data manipulation, parallelization, and graphics. The `R` scripts are available on GitHub for reproducibility[1].

[1] `https://github.com/sarahshy/GaussianPerturbation`

Chapter 2
Data for classification in astronomy

In a typical classification problem, the data are composed of labeled and unlabeled datasets, as shown in the left panel of Figure 2.1. For convenience, we assume that the entire sample of n objects is unified, with the first m observations being labeled and the remaining unlabeled. The i-th row contains measurements of p features for the i-th object, which is denoted by $x_i^{\text{obs}} = \{x_{i1}^{\text{obs}}, x_{i2}^{\text{obs}}, \ldots, x_{ip}^{\text{obs}}\}$. In the labeled set, the class label of the i-th object is denoted by y_i. In practice, the labeled set is often randomly split into training and test sets (e.g., in k-fold cross validation) to train and validate a classifier. Then, the fitted classifier is used to make predictions on the unlabeled dataset.

In astronomy, a measurement of the j-th feature for the i-th object (i.e., x_{ij}^{obs}) is subject

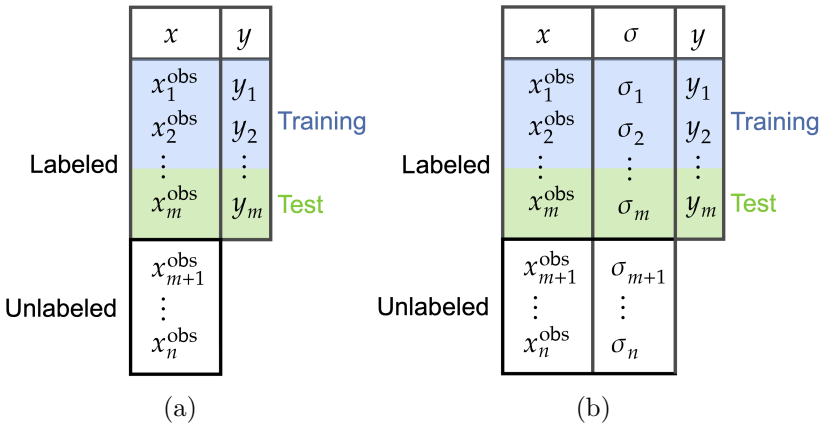


Figure 2.1: In a typical classification problem, a classifier is trained and validated on the labeled set, and then used to make predictions on the unlabeled set. In astronomy, the information about measurement error uncertainty is additionally given in the data set, as illustrated in the right panel.

to measurement uncertainty. This is due to limitations of the telescope and detector sensitivity, exposure time, observing conditions, and data reduction procedures. The common assumption is that the measurement errors are i.i.d. Gaussian noise (Eddington 1913), i.e.,

$$\epsilon_{ij} \sim \mathcal{N}(0,\ \sigma_{ij}^2)$$

The 1σ measurement error uncertainty, σ_{ij}, is directly measured by careful calibration of the instrument and examination of source-free regions of the image or spectrum. The phenomenon where the values of σ_{ij} differ for each object i as well as for each feature j is called *heteroscedasticity*.

Thus, the heteroscedastic measurement error uncertainty is additionally known for both the labeled and unlabeled sets in astronomy, as illustrated in the right panel of Figure 2.1. In the figure, we denote the 1σ uncertainties of p measurements of the i-th object by $\sigma_i = \{\sigma_{i1}, \sigma_{i2}, \ldots, \sigma_{ip}\}$. These measurement error uncertainties are often ignored out of necessity. This work aims at utilizing the entire array of available information, including the heteroscedastic measurement error uncertainties given in the data.

Chapter 3
Methodology

We adopt a Gaussian measurement error model, assuming that the observed quantities are noisy measurements centered at the true quantities with known measurement error uncertainty (Eddington 1913). This means that hypothetical repeated measurements of the same object will follow a Gaussian distribution centered at its true value. For each of n observations ($i = 1, 2, \ldots, n$) and p features ($j = 1, 2, \ldots, p$),

$$x_{ij}^{\text{obs}} \sim \mathcal{N}(x_{ij}^{\text{true}},\ \sigma_{ij}^2), \tag{3.1}$$

where x_{ij}^{obs} is the measurement of the j-th feature of the i-th object, x_{ij}^{true} is the corresponding unknown true quantity, and σ_{ij} is the known measurement error uncertainty. Here we assume an independent measurement for each feature, but we also discuss correlated measurements in Section 6.3.

To reflect our lack of knowledge about the true values, x_{ij}^{true}'s, and to minimize the modeling uncertainty, we let the data speak for themselves by adopting a jointly improper flat prior on x_{ij}^{true}. That is, for all i and j,

$$h(x_{ij}^{\text{true}}) \propto 1. \tag{3.2}$$

The resulting posterior distribution of the true value x_{ij}^{true} given the observation x_{ij}^{obs} is a proper Gaussian distribution,

$$x_{ij}^{\text{true}} \mid x_{ij}^{\text{obs}} \sim \mathcal{N}(x_{ij}^{\text{obs}},\ \sigma_{ij}^2), \tag{3.3}$$

satisfying posterior propriety (Hobert & Casella 1996, Tak et al. 2018).

To simulate a perturbed replicate of the observed data, we sample the posterior

predictive distribution of the Gaussian measurement error model,

$$q(x_{ij}^{\text{rep}} \mid x_{ij}^{\text{obs}}) = \int f(x_{ij}^{\text{rep}} \mid x_{ij}^{\text{true}})\ \pi(x_{ij}^{\text{true}} \mid x_{ij}^{\text{obs}})\ dx_{ij}^{\text{true}},$$

where x_{ij}^{rep} is a predicted value of x_{ij}^{obs}. The distributions f and π in the integrand are defined in Equations (3.1) and (3.3), respectively. The distribution of $(x_{ij}^{\text{rep}} \mid x_{ij}^{\text{true}})$ is the same as that of $(x_{ij}^{\text{obs}} \mid x_{ij}^{\text{true}})$ because both x_{ij}^{rep} and x_{ij}^{obs} are noisy measurements given the unknown true value. This posterior predictive distribution is a distribution of predicted data given the observed data, which we obtain by accounting for all possible realizations of the unknown true value x_{ij}^{true} (Gelman et al. 2013, Section 1.3). The resulting posterior predictive distribution of $q(x_{ij}^{\text{rep}} \mid x_{ij}^{\text{obs}})$ is still Gaussian:

$$x_{ij}^{\text{rep}} \mid x_{ij}^{\text{obs}} \sim \mathcal{N}(x_{ij}^{\text{obs}},\ 2\sigma_{ij}^2). \tag{3.4}$$

Here, the prediction variance $2\sigma_{ij}^2$ is inflated from the variance of the measurement error σ_{ij}^2 by a factor of 2. This is to account for the additional uncertainty of the unknown true value x_{ij}^{true} when predicting a future observation x_{ij}^{rep} given the proposed model. In other words, the inflation factor is the result of our choice of prior in (3.2). For example, if we were to use a different Gaussian prior on each, the resulting posterior predictive distribution would be still Gaussian, but with a different variance.

This Gaussian posterior predictive distribution specifies that we can directly simulate replicates of x_{ij}^{obs} by generating x_{ij}^{rep} from the Gaussian distribution in Equation (3.4). We call this replication process *Gaussian perturbation* because x_{ij}^{rep} can be considered a perturbation of x_{ij}^{obs} with random noise drawn from $\mathcal{N}(0,\ 2\sigma_{ij}^2)$. Performing this process repeatedly will produce multiple simulations of the observed data under the proposed Gaussian measurement error model. We denote the r-th replicate of the i-th observation by $x_{i(r)}^{\text{rep}} = \{x_{i1(r)}^{\text{rep}}, x_{i2(r)}^{\text{rep}}, \ldots, x_{ip(r)}^{\text{rep}}\}$ and the r-th simulated data set by $X_{(r)}^{\text{rep}} = \{x_{i(r)}^{\text{rep}} : i = 1, 2, \ldots, n\}$. This simulation process is summarized in Algorithm 1.

Algorithm 1: Gaussian perturbation to simulate R perturbed data sets from the posterior predictive distribution.

input : Observed data x_{ij}^{obs}'s, known measurement error uncertainties σ_{ij}'s
result : R perturbed data sets $X_{(1)}^{\text{rep}}, \ldots, X_{(R)}^{\text{rep}}$
for $r = 1 \ldots R$ **do**
  Sample $x_{ij(r)}^{\text{rep}} \sim \mathcal{N}(x_{ij}^{\text{obs}},\ 2\sigma_{ij}^2)$ for all i and j
end

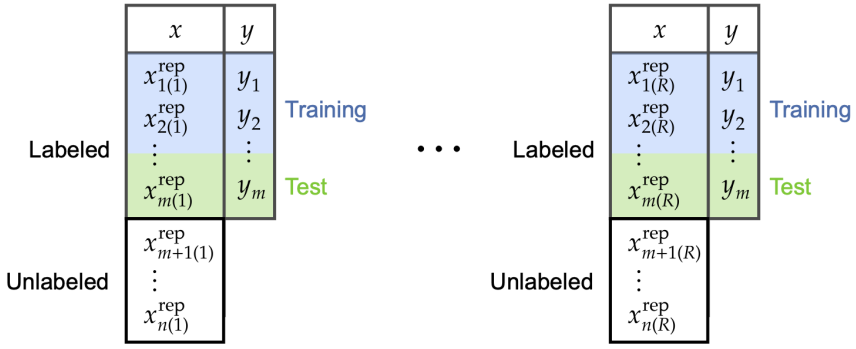


Figure 3.1: The perturbed data sets simulated by a posterior predictive distribution of a measurement error model. The information about the measurement error is used only for simulation, and does not appear in the perturbed data sets. Thus, any traditional classification method can be trained on each perturbed data set independently, and the variation of the multiple fits can be ascribed to the measurement error uncertainty.

We point out that under the proposed framework, the information about the measurement error uncertainty is used only to generate perturbed datasets. Since the error information is incorporated into the ensemble of perturbed datasets, each individual perturbed dataset itself does not contain a column for measurement error, as shown in Figure 3.1. Therefore, any traditional classification method can be fit to each perturbed dataset as if the information about the measurement error were not given. The key point is that the variation across the multiple fits will naturally reflect the known measurement error uncertainty because the variation is ascribed solely to the perturbations under the Gaussian measurement error model.

3.1 Quantifying Overall Uncertainty of a Classifier

To utilize the uncertainty propagated from measurement error, we fit a classifier to each of the R simulated data sets and obtain R classification results. We can then summarize

Algorithm 2: The posterior predictive distribution of any classification metric θ.

input : perturbed data sets $X^{\text{rep}}_{(1)}, \ldots, X^{\text{rep}}_{(R)}$, classification algorithm C, metric θ

result : A sample, $\{\theta_{(1)}, \ldots, \theta_{(R)}\}$, of size R from the posterior predictive distribution of θ

for $r = 1 \ldots R$ **do**

  Fit classifier C to $X^{\text{rep}}_{(r)}$

  Calculate metric of interest $\theta_{(r)}$

end

the variation across the results of any quantity of interest related to the classification, such as classification accuracy and decision rule parameters.

To illustrate this, let θ denote some measure of classification accuracy, and $\theta_{(r)}$ the measured classification accuracy obtained from the labeled test set of the r-th perturbed data set. Then, the distribution of $\theta_{(1)}, \theta_{(2)}, \ldots, \theta_{(R)}$ obtained from R independent fits represents the posterior predictive distribution of the classification accuracy θ. The uncertainty of the classification accuracy θ can be quantified by the spread of this posterior predictive distribution, e.g., using the posterior standard deviation or a credible interval. Algorithm 2 provides a general template for obtaining the posterior predictive distribution for any metric θ that is a function of the classification results. Lastly, we note that this fitting procedure is easily parallelized to reduce computational burden as each fit on a simulation can be performed independently.

3.2 Prediction on the Unlabeled Set

The proposed framework also provides a method to predict each individual's unknown label while accounting for measurement error uncertainty. For hard classifiers, we obtain a set of predicted class labels for each object from the multiple fits. Using these predicted labels, we can estimate class probabilities similar to soft classifiers. We refer to this as *softening hard classifiers*. The distinction between the softened hard classifier and traditional soft classifiers is that the former incorporates measurement error uncertainty into the estimated class probabilities. For soft classifiers, we obtain a set of estimated class probabilities for each object from multiple fits, providing additional information about the uncertainty in classifying an object. We refer to this as *softening soft classifiers* (see Figure 1.2).

We first discuss how to make a prediction via softening a hard classifier to obtain a single predicted label for each object in the unlabeled data set. This procedure is essentially the same as that of a traditional soft classifier. We define K as the number of classes we wish to classify objects into, and

$$\hat{y}_{i(r)} \in \{1, 2, \ldots, K\}$$

as the predicted class label of the i-th object obtained from a fit on the r-th perturbed data set. By conducting R simulations, we obtain R predictions for the i-th object, $\hat{y}_{i(1)}, \ldots, \hat{y}_{i(R)}$. Using these R predictions, we compute the proportion of simulations that

classify object i into class k,

$$\hat{p}_{ik} = \frac{1}{R}\sum_{r=1}^{R} \mathbb{1}\{\hat{y}_{i(r)} = k\}, \tag{3.5}$$

where $\mathbb{1}\{\hat{y}_{i(r)} = k\}$ is an indicator equaling 1 if the r-th simulation places object i into class k and 0 otherwise. This proportion is the estimated probability that object i belongs to class k.

Finally, the predicted class of the i-th object is k if its estimated probability of being in class k, $\hat{p}_{ik}$, is greater than or equal to some decision threshold t between 0 and 1. The threshold is typically chosen by scientists for their purposes (He & Garcia 2009). For example, a simple choice for t in a binary classification problem is to set $t = 0.5$, which is equivalent to predicting the class of the i-th object as class 1 when $\hat{p}_{i1} > \hat{p}_{i0}$. One disadvantage of this simple choice is that it classifies objects into class 1 even when the classification is uncertain, e.g., with $\hat{p}_{i1} = 0.501$ and $\hat{p}_{i0} = 0.499$. Thus, it is desirable to set the threshold according to the purpose of classification in practice. For example, if object i did not have estimated probabilities greater than $t = 0.8$, then we might consider the class of object i as a new class, 'ambiguous' (e.g., unsafe cases in Napierala & Stefanowski 2015). This ambiguous class ensures that uncertain objects are separated from more certain ones, and that the remaining classes contain only objects that we can be confident about. Setting a higher threshold, such as $t = 0.9$, will give even smaller sample in each class but with greater purity. Such a strong decision rule may provide an effective way to remove objects with large measurement errors from further scientific consideration. This threshold approach can also be used for a multi-class problem with more than two classes. That is, the predicted class of object i is k if $\hat{p}_{ik}$ is the only probability estimate greater than t and 'ambiguous' otherwise.

For inherently soft classifiers, the output of each fit is a set of estimated class probabilities such as $\hat{p}_{i1}, \hat{p}_{i2}, \ldots, \hat{p}_{iK}$ for the i-th object instead of a single class label. An implementation of random forest, for example, will plant many classification trees (forming a forest), make a label prediction using each tree, and compute $\hat{p}_{ik}$ as formulated in Equation (3.5). Gaussian perturbation provides possible variations of $\hat{p}_{ik}$ via R simulations, i.e., $\hat{p}_{ik(1)}, \hat{p}_{ik(2)}, \ldots, \hat{p}_{ik(R)}$, whose variation reflects the measurement error uncertainty. The average of these R variants provides a more informative estimated class probability by accounting for measurement error, i.e.,

$$\hat{p}^{+}_{ik} = \frac{1}{R}\sum_{r=1}^{R} \hat{p}_{ik(r)}. \tag{3.6}$$

We use the superscript '+' to indicate that the quantity is obtained after accounting for measurement error uncertainties. We have selected this symbol because measurement error uncertainties are typically visualized by cross bars around each 2-dimensional data point, e.g., as shown in the third panel in Figure 4.1. This formulation satisfies $\sum_{k=1}^{K} \hat{p}_{ik}^{+} = 1$, ensuring that the estimated probabilities for each object sum to 1 across all classes. Finally, the class of the i-th object is predicted to be k if $\hat{p}_{ik}^{+}$ is greater than or equal to some threshold t, as is usually done for soft classifiers.

Chapter 4
Simulation Study

We illustrate how Gaussian perturbation can be used to better quantify classification uncertainty and make predictions accounting for measurement error uncertainty. Let us consider the following simulation setting with two features ($p = 2$) and heteroscedastic measurement error uncertainties. We instantiate a set of 'true' data with 200 observations by sampling from two bivariate Gaussian distributions,

$$x_1^{\text{true}}, \ldots, x_{100}^{\text{true}} \sim \mathcal{N}\left(\begin{bmatrix}5\\5\end{bmatrix}, \begin{bmatrix}3 & -0.5\\-0.5 & 2\end{bmatrix}\right)$$
$$x_{101}^{\text{true}}, \ldots, x_{200}^{\text{true}} \sim \mathcal{N}\left(\begin{bmatrix}2\\2\end{bmatrix}, \begin{bmatrix}4 & -1\\-1 & 1\end{bmatrix}\right)$$

where $x_i^{\text{true}} = (x_{i1}^{\text{true}},\ x_{i2}^{\text{true}})$ for all i. The true class labels, denoted as y_i, are known from the setup,

$$y_i = 0, \quad \text{for} \quad i = 1, \ldots, 100$$
$$y_i = 1, \quad \text{for} \quad i = 101, \ldots, 200$$

These data represent *true* values of 200 objects belonging to two classes labeled 0 and 1, and are displayed in the left panel of Figure 4.1.

To simulate noisy observations of the true data, we arbitrarily define the measurement error variance as $\sigma_{ij}^2 = |x_{ij}^{\text{true}}|/2$. These errors are heteroscedastic, where larger values of x_{ij} are subject to greater measurement error. This type of behavior occurs, for example, in astronomical surveys where the measured features are magnitudes. Next, we add measurement errors ε_{ij}'s to the true values to obtain the *observed* data x_{ij}^{obs},

$$x_{ij}^{\text{obs}} = x_{ij}^{\text{true}} + \varepsilon_{ij},$$

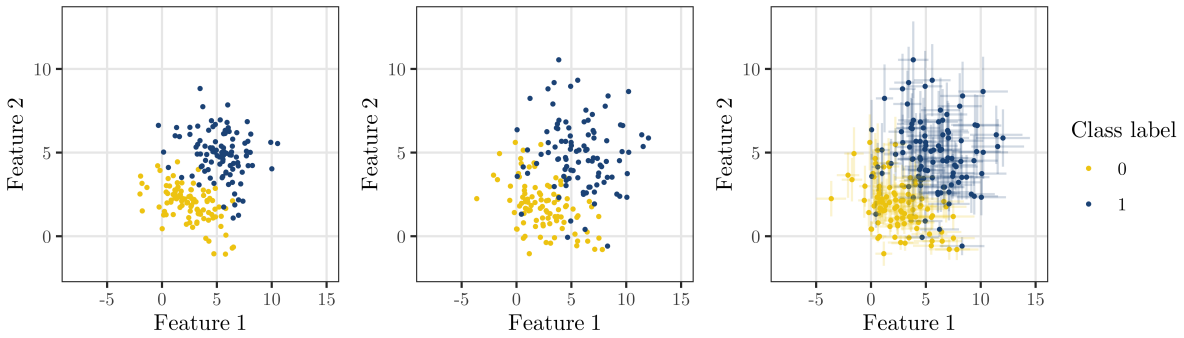


Figure 4.1: Left: the true unobserved data. Center: the observed data without considering measurement error, suitable for standard classification methods. Right: the observed data with error bars whose lengths represent 2 standard deviations of measurement error uncertainties (i.e., measurement $\pm 1\sigma$), which the proposed framework utilizes to simulate perturbed data sets.

where ε_{ij} is sampled from $N(0,\ \sigma_{ij}^2)$. The simulated observations, x_i^{obs}, are shown in the center panel of Figure 4.1, representing the typical data used in standard classification methods (i.e., without accounting for measurement error). In the right panel of Figure 4.1, the 1σ measurement error uncertainties are superimposed on the observed values, representing the information utilized by the proposed framework. These 200 observations serve as the labeled set.

In addition to the labeled set, we carefully construct an unlabeled set of two observations to illustrate how measurement error affects their class prediction under the proposed framework. These observations are highlighted in Figure 4.2; unlabeled object 1 (red circle) lies far from the overlapping area of the two classes with a large error bar and unlabeled object 2 (blue triangle) lies near the overlapping area with a large error bar.

In this simulation study, we use a simple decision threshold $t = 0.5$ because we do not have a specific scientific motivation to choose a higher threshold. That is, the predicted class of the i-th object is 1 if the probability of this object being in class 1 is higher than that of being in class 0 ($\hat{p}_{i1} > \hat{p}_{i0}$ or $\hat{p}_{i1}^{+} > \hat{p}_{i0}^{+}$).

4.1 Support-vector machine

For comparison, we implement a linear SVM with and without considering measurement error. The linear SVM is originally designed to utilize only the measurements $X^{\text{obs}} = \{x_1^{\text{obs}}, \ldots, x_{202}^{\text{obs}}\}$ without incorporating measurement error $\sigma = \{\sigma_1, \ldots, \sigma_{202}\}$. To fit this model, we use 10-fold cross validation on the labeled set (i.e., on the first 200 observations)

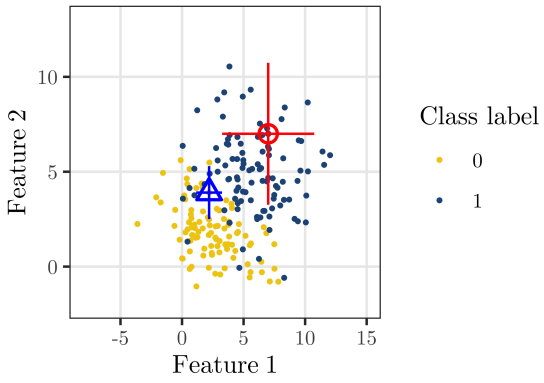


Figure 4.2: Two observations in the unlabeled set are superimposed on the 200 observations in the labeled set.

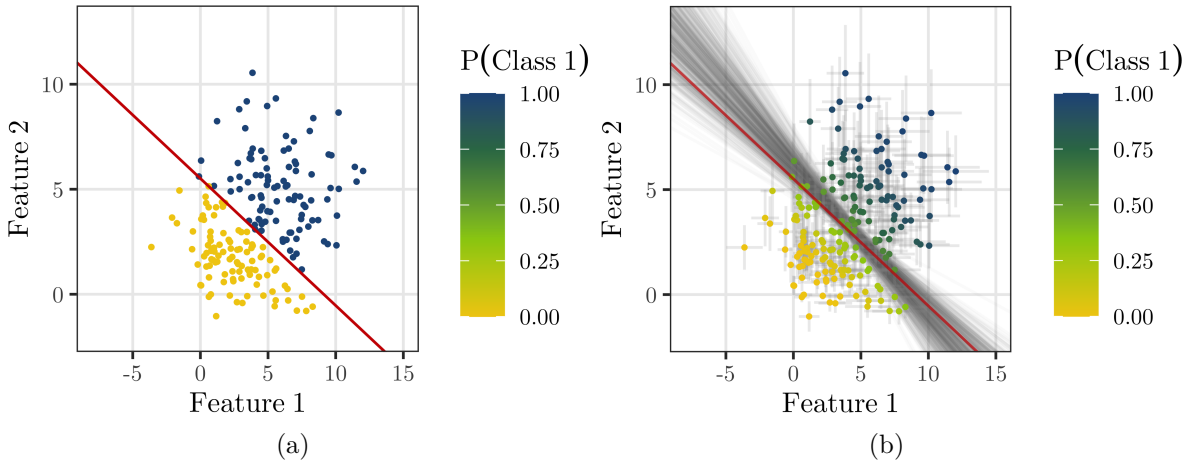


Figure 4.3: (a) Linear SVM boundary fit to the observed data without accounting for measurement error, colored by the predicted labels. The probability of being in class 1 is 1 for blue dots and 0 for yellow dots. The red line is the fitted decision boundary of the linear SVM. (b) 500 SVM decision boundaries obtained by fitting the SVM on each of the 500 perturbed data sets, are colored in gray. The red line is reproduced from the left panel for a comparison. The estimated probability of belonging to class 1, i.e, $\hat{p}_{i,1}$ for each object i is denoted by a blue-green-yellow color gradient; darker dots are more likely to belong to class 1. Both the training and test sets are plotted, but each decision boundary was derived using the training set only. The length of each error bar is 2 standard deviations of measurement error (i.e., measurement $\pm 1\sigma$).

to tune the hyperparameters of the model and estimate classification accuracy. The fitted model is then used to predict the class labels of the two observations in the unlabeled set.

We then apply Gaussian perturbation to the labeled and unlabeled sets to generate 500 simulated data sets, $X^{\text{rep}}_{(1)}, \ldots, X^{\text{rep}}_{(500)}$. For each data set, we fit the linear SVM via 10-fold cross-validation, and predict on the unlabeled set.

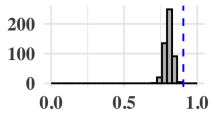
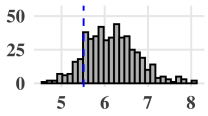
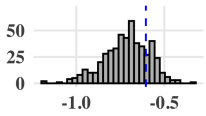
The left panel of Figure 4.3 shows the labeled data and the resulting red decision boundary without accounting for measurement error. The right panel visualizes the result of incorporating measurement error using the proposed approach. The 500 gray decision boundaries obtained by fitting the linear SVM on each of the 500 perturbed data sets appear to form a decision 'band'. This band represents the uncertainty of the decision boundary due to measurement error, i.e., the variation around the red line.

We first compare the classification accuracy between the standard approach and the proposed approach considering measurement error. Without accounting for measurement error, the cross-validated classification accuracy is 0.91. Accounting for measurement error, the average cross-validated classification accuracy is 0.80 with a standard deviation of 0.03; see Table 4.1 for details. The variation in accuracy stems from the variation in the SVM decision boundary due to measurement error. The former (0.91) is more than three standard deviations away from the accuracy obtained when accounting for measurement error (i.e., larger than 0.89). This indicates that overconfidence about our measurements can lead to exaggerated results and potential bias, as is common in linear regression fits with/without measurement error (Akritas & Bershady 1996). The lower accuracy of the proposed method is unsurprising as the presence of measurement error has further blurred the separation between the classes.

The estimated probability of each object belonging to class 1, $\hat{p}_{i1}$, is calculated as defined in Equation (3.5) and visualized using a blue-green-yellow color gradient in the right panel of Figure 4.3; darker dots are more likely to belong to class 1. Objects that are closer to the decision band are less certain, with estimated class probabilities near 0.5. In the left panel of Figure 4.3, colors are either blue or yellow without gradation because SVM predicts each object's class as a single label with probability 1.

Finally, we compare the prediction results of the two observations in the unlabeled set. As shown in Figure 4.2, unlabeled object 1 (red circle) lies far from the intersection of the two classes and is near the majority of class 1 objects with large measurement error. The SVM without measurement error classifies this into class 1 with probability 1. On the other hand, SVM with measurement error makes a less certain prediction with an estimated probability of being class 1 equal to 0.78. Even for such a seemingly clean-cut object that is far from the decision band, 22% of the perturbed values of this object would lie below the decision boundary if its large measurement error were considered. Object 2 (blue triangle) lies on the intersection with large measurement error. Without considering measurement error, this object's class is predicted to be class 0 with probability 1. With measurement error, however, the uncertainty of the object is reflected, with an estimated

Table 4.1: Comparison of overall classification accuracy and decision boundary parameters of a linear SVM computed from the labeled set with and without incorporating measurement error uncertainties. The vertical dashed lines in the posterior predictive distributions indicate the values obtained from the SVM without measurement error.

Method	**Accuracy**	**SVM decision boundary**	
		Intercept	Slope
SVM w/o measurement error	0.91	5.51	-0.60
SVM w/ measurement error	0.80 ± 0.03	6.15 ± 0.60	-0.71 ± 0.13
Posterior predictive dist'n			

probability of belonging to class 0 equal to 0.46.

Consequently, even though single predicted labels are consistent under both approaches, we note that the proposed approach provides more elaborate uncertainty quantification.

4.2 Random Forest

We fit a random forest on the same data of 202 objects, repeating the same 10-fold cross validation procedure. The cross-validated classification accuracy of random forest without accounting for measurement error is 0.87. Considering measurement error, the average classification accuracy decreases to 0.77 with standard deviation 0.03. This result is similar to that obtained by SVM in Section 4.1. It indicates that random forest is also sensitive to the measurement error and that ignoring measurement error can exaggerate classification performance.

Next, we visually compare the estimated probabilities obtained with and without measurement error. The left panel of Figure 4.4 displays the estimated probabilities of belonging to class 1 obtained without measurement error ($\hat{p}_{i1}$ for each object i, defined in Equation (3.5)) using the same color gradient as in Figure 4.3. Random forest clearly shows some color gradation near the overlapping area, unlike SVM's strict blue-yellow distinction around the decision boundary. However, the gradient is sporadic rather than continuously changing near the overlapping area, with dark green dots scattered even in the predominantly yellow area of class 0 (i.e., near the lower-left corner). It turns out that these dark green dots in the yellow-dominant region are known to belong to class 1.

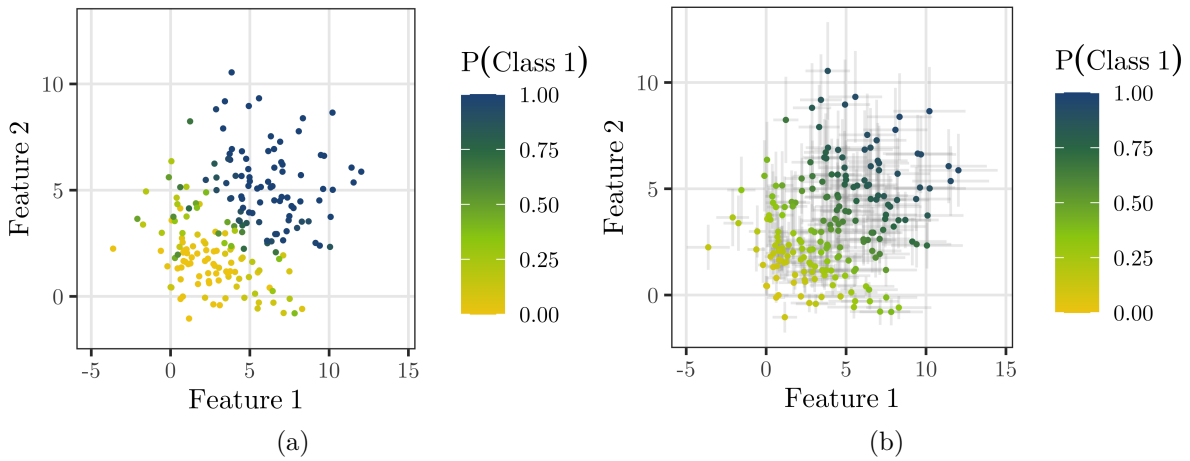


Figure 4.4: (a) The probability estimate of being in Class 1 computed from a one-time implementation of random forest without measurement error ($\hat{p}_{i1}$ for each object i, defined in Equation (3.5)) is visualized via color gradation. (b) The probability estimate calculated by the proposed approach that accounts for measurement error ($\hat{p}^{+}_{i1}$ for each object i, defined in Equation (3.6)) is visualized using the same color gradation. Darker dots are more likely to belong to class 1.

A comparison between the left and middle panels of Figure 4.1 indicates that these data has become isolated from class 1 while simulating noisy measurements. From this result, one may conclude that random forest without measurement error has worked very well, identifying these dark green dots even deep inside the yellowish area. However, we point out that this standard random forest approach is not concerned with measurement error, i.e., why these data became isolated from the majority of class 1 in the beginning. Instead, it assumes that the data are perfectly measured. It has just worked well under this assumption, producing relatively high probabilities of these data belonging to class 1, even though their *error-free* measurements put them near the majority of class 0.

On the other hand, the proposed approach reflects the possibility that these class 1 data may be located near the majority of class 0 simply due to measurement error. The estimated probabilities obtained by Gaussian perturbation with measurement error, i.e., $\hat{p}^{+}_{i1}$ in Equation (3.6), are visualized in the right panel of Figure 4.4 using the same color gradient. Compared to the left panel (obtained without measurement error), the right panel shows a thick greenish band. This band occupies most of the previously yellow-dominant region, leaving a smaller yellow area in the lower-left corner. This indicates that classifying objects into class 0 has become less certain after we account for measurement error. This phenomenon can be ascribed to the isolated class 1 points residing within the majority of the class 0 data. Since each of these isolated class 1

data is used for training in 9 folds during 10-fold cross-validation, the trees of a random forest routinely identify nearby data to be class 1. When aggregated, this increases the probability of being in class 1 for the nearby class 0 data. The act of perturbing the data using Gaussian perturbation spreads this uncertainty deeper into the lower-left corner. For example, half of perturbed values of the isolated class 1 data are generated even further away from class 1 (i.e., closer to the lower-left corner), and each of them increases the probability of nearby data being in class 1. Thus, due to the average effect of the proposed approach, the previously yellow-dominant area has become greenish, reflecting more uncertainty in classifying objects into class 0.

Lastly, we predict the classes of the two observations in the unlabeled set. Unlabeled object 1 (red circle in Figure 4.2) is near the majority of class 1 objects with large measurement error. Random forest without measurement error classifies this into class 1 with an estimated probability of being in class 1 ($\hat{p}_{i0}$) equal to 1. With measurement error incorporated, the probability estimate ($\hat{p}^{+}_{i1}$) decreases to 0.79 with standard deviation 0.29, appropriately reflecting its large measurement error. It turns out that both approaches predict this unlabeled object to be in class 1 with a simple threshold of $t = 0.5$ (i.e., $\hat{p}_{i1} > \hat{p}_{i0}$ and $\hat{p}^{+}_{i1} > \hat{p}^{+}_{i0}$ for $i = 201, 202$). But we note that the proposed approach provides the more comprehensive information about its prediction uncertainty.

Object 2 (blue triangle in Figure 4.2) is located in the overlapping area with large measurement error. Without considering measurement error, this object's class is predicted to be class 1 with estimated probability $\hat{p}_{i1} = 0.51$. This probability implies that our class prediction for this observation is as good as a coin toss. By considering measurement error, however, the proposed approach reveals ambiguity in a more elaborate way, producing the probability estimate $\hat{p}^{+}_{i1} = 0.45$ with standard deviation 0.32. Consequently, the label prediction for this object becomes completely different, i.e., predicted to be class 1 without measurement error and class 0 with measurement error. It implies that measurement error can play a significant role in predicting unknown labels in practice. We also note that while the estimated probabilities are similar, the proposed approach tells more about the uncertainty involved in the prediction than a simple coin toss.

Chapter 5
Classifying High-Redshift Quasars

Following the work of Timlin et al. (2018), we consider the problem of identifying high-z quasar candidates ($2.9 \leq z \leq 5.1$) in a catalog data set merged from multiple sources. Here we adopt only the random forest classifier to demonstrate the application of the Gaussian perturbation framework, although other classifiers can be similarly applied to the dataset. Also, we adopt the same decision threshold as Timlin et al., setting the threshold $t = 0.5$ as is done in our simulation study.

5.1 The Labeled and Unlabeled Data Sets

We collect labeled and unlabeled photometric catalog data from several sources, following Timlin et al. (2018). Their scientific goal is to identify high-z quasar candidates from the unlabeled set when the prevalence rate of high-z quasars in the labeled set is only $\sim$3%. The labeled data consist of optical photometric data from the Sloan Digital Sky Survey (SDSS) combined with data from the Spitzer IRAC Equatorial Survey (SpIES; Timlin et al. 2016) and the Spitzer-HETDEX Exploratory Large-Area survey (SHELA; Papovich et al. 2016). The combined data are restricted to Stripe 82 objects (Annis et al. 2014, Jiang et al. 2014) for which deep optical photometry is available in the five optical SDSS filters (*ugriz*; Fukugita et al. 1996). Photometric errors for each object in each band are also provided.

The 5,487 quasars in the labeled set are spectroscopically confirmed as high-z quasars from the quasar catalog of Richards et al. (2015). The remaining 643,952 objects in the labeled set, consisting of stars, galaxies, and unclassified quasars, are collapsed into a single class called 'AE' (Anything Else). The unlabeled set is assembled using matched optical+MIR photometric data restricted to Stripe 82 and consists of 1,862,968 objects. See Timlin et al. for further details regarding how the labeled and unlabeled sets are

assembled. We note that the labeled and unlabeled data sets of this work are not exactly the same as those in Timlin et al. due to different time of data collection.

5.2 Classification Accuracy on the Labeled Set

Using these labeled and unlabeled observations with their measurement error uncertainties, we first generate $R = 500$ simulated data sets via Gaussian perturbation. In each simulation, we apply 10-fold cross validation to the labeled data set to train and test a random forest classifier. Considering the class imbalance (only about 3% of objects are high-z quasars), we use stratified sampling to form 10 cross validation sets, each with 3% high-z quasars and 97% AE. That is, we randomly divide the labeled set of high-z quasars and that of AE into 10 equally-sized pieces, and pair one piece of high-z quasar with one piece of AE.

To assess the classification accuracy in the presence of substantial class imbalance in the data, we use two measures, *completeness* and *efficiency* (Timlin et al. 2018). This is because the typical measure for classification accuracy (the total number of correct classifications divided by the total number of objects) is dominated by the result of the much larger AE group. Completeness — also commonly known as sensitivity, recall, or true positive rate — evaluates the number of correctly identified high-z quasars out of all known high-z quasars in the test set. Efficiency — also referred to as precision or positive predictive value — evaluates the number of correct high-z quasar classifications out of the total number of objects classified as high-z quasar.

In terms of these two measures, the single-run random forest without measurement error shows evidence of over-confidence in its classification accuracy. Without considering measurement error, completeness is 82.3% based on a single implementation of random forest. Incorporating measurement error via Gaussian perturbation, completeness becomes 82.1% on average across 500 simulations with standard deviation 0.1%. The value of completeness without measurement error is on the upper bound of the two standard deviation range obtained from the proposed approach with measurement error. Thus, the outcome without measurement error can be considered nearly over-confident in completeness. Similarly, efficiency is 88.6% without measurement error and 88.4% $\pm$ 0.08% with measurement error. The value of efficiency without measurement error is higher than the two standard deviation range obtained with measurement error, indicating that the former is over-confident in efficiency. We display these results in Figure 5.1. The vertical dashed line in each panel indicates the value of completeness or efficiency obtained from

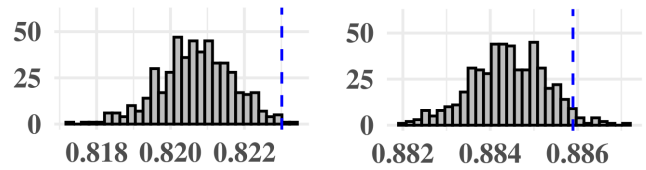


Figure 5.1: Posterior predictive distributions of completeness (left) and efficiency (right), respectively, obtained by Gaussian perturbation with 500 simulations. The mean and standard deviations are superimposed in each panel. The vertical dashed line indicates the value of completeness or that of efficiency obtained without considering measurement error.

the one-time implementation of random forest without considering measurement error.

To discuss false-positive and false-negative rates, we display confusion matrices in Table 5.1 obtained with and without accounting for measurement error on the labeled set. The confusion matrix obtained via Gaussian perturbation is averaged over 500 confusion matrices, one from each simulation. The single run of random forest identifies 2,392 false-positives and 3,705 false-negatives, shown in the upper-right and lower-left cells, respectively. On the other hand, the proposed approach via Gaussian perturbation shows 2,214.9 $\pm$ 18.7 false-positives and 3701.1 $\pm$ 20.7 false-negatives when measurement error is considered.

The first column of the two tables are nearly identical; almost the same number of true high-z quasars are misclassified to AE (false-negatives) regardless of whether we consider measurement error or not. This implies that these misclassified high-z quasars might be located deep inside the feature space of AE objects even after accounting for the measurement error. In the second column, however, more than 100 true AE objects are misclassified as high-z quasars (false-positives) without measurement error, while correctly identified as AE with measurement error. These AE objects may be distant from the majority of AE objects in the feature space, but not as distant if we consider their measurement error uncertainties. From a practical standpoint, we also note that reducing the number of false positives is extremely important for follow-up observations. This is because a reduction of even $\sim 5\%$ of false-positives is not only useful for particular scientific goals, but also saves time, cost, and effort when further investigating the candidates.

Table 5.1: The confusion matrices obtained (a) without and (b) with measurement error. For case (b) the outcome is the average of 500 confusion matrices, and the values in the parentheses are standard deviations. The acronym 'AE' denotes the class 'Anything Else'.

(a) Without measurement error

Predicted\True	High-z Quasars	AE
High-z Quasars	16,935	2,392
AE	3,705	626,407

(b) With measurement error

Predicted\True	High-z Quasars	AE
High-z Quasars	$16,938.9 \pm 20.7$	$2,214.9 \pm 18.7$
AE	$3,701.1 \pm 20.7$	$626,584.1 \pm 18.7$

5.3 Prediction on the Unlabeled Set via Gaussian Perturbation

We also predict labels of the newly observed objects in the unlabeled set with and without measurement error. To account for measurement error, we compute estimated probabilities of being high-z quasars using Gaussian perturbation, $\hat{p}_{i1}^{+}$ in Equation (3.6) for $i = 1, 2, \ldots, n$. Then, we classify the i-th object in the unlabeled set as a high-z quasar if $\hat{p}_{i1}^{+} > 0.5$ (or equivalently $\hat{p}_{i1}^{+} > \hat{p}_{i0}^{+}$).

Table 5.2 summarizes these results. It shows that both approaches are consistent in predicting high-z and AE objects on the diagonal cells. However, 936 objects are predicted to be high-z quasars with measurement error, but to be AE without measurement error. Also, 3,146 objects are classified as high-z quasars without measurement error but as

Table 5.2: Overlap of label predictions on the prediction set between a single random forest and an ensemble random forest using $R = 500$ simulations. The approach without measurement error is denoted by 'Standard' and the one with measurement error by 'Proposed'. For example, 936 objects are predicted to be high-z quasars with measurement error (Proposed), but to be AE without measurement error (Standard). Also, 3,146 objects are classified as high-z quasars without measurement error (Standard) but as AE with measurement error (Proposed).

Proposed \Standard	High-z Quasars	AE
High-z Quasars	8,701	936
AE	3,146	1,850,185

AE with measurement error. Consequently, a single run of random forest without measurement error identifies $11,847$ high-z quasars from $\sim$2 million objects in the unlabeled set (i.e., the sum of the first column). Taking measurement error into account, random forest via Gaussian perturbation identifies 9,637 high-z quasars (i.e., the sum of the first row), which is substantially lower than the number of high-z quasars identified without measurement error.

This difference reveals important aspects of the classification with and without measurement error. The top-left cell of Table 5.2 shows that the two approaches are consistent in predicting 8,701 objects as high-z quasars. However, the top-right cell indicates that 936 object are classified as AE without measurement error, while classified as high-z quasars after considering measurement error. In other words, these 936 objects are potential candidates for high-z quasars previously hidden under the assumption that the measurement has no uncertainty. Also the bottom-left cell of Table 5.2 indicates that 3,146 objects, which are classified as high-z quasars without measurement error, are classified as AE after accounting for measurement error. This means that these 3,146 objects might be potential misclassifications (i.e., not high-z quasars) of the standard approach that does not account for measurement error.

Chapter 6
Discussion

6.1 Why Bayesian posterior predictive distribution?

Under the Gaussian measurement error model in Equation (3.1), the observed data are measurements of the unknown true values with Gaussian noise, i.e.,

$$x_{ij}^{\text{obs}} \sim \mathcal{N}(x_{ij}^{\text{true}},\ \sigma_{ij}^2),$$

where σ_{ij}^2 is known for all i and j.

Based on this model, one naive way to simulate replicates of the current data is to add Gaussian noise to the observation with *the same* measurement error uncertainty. That is,

$$x_{ij}^{\text{rep}} \sim \mathcal{N}(x_{ij}^{\text{obs}},\ \sigma_{ij}^2).$$

Under the Gaussian measurement error model, it makes more sense for replicates to be distributed around x_{ij}^{true}. This is because the model assumes that (hypothetical) repeated measurements under the same condition are distributed as a Gaussian distribution centered at x_{ij}^{true}. The replicates obtained with the naive approach would be consistent with this assumption only if $x_{ij}^{\text{obs}} = x_{ij}^{\text{true}}$. In practice, however, this condition is impossible to meet due to non-zero measurement error ($\sigma_{ij} > 0$), violating the key idea of the Gaussian measurement error model.

A Bayesian approach provides a straightforward way to simulate replicates centered at x_{ij}^{true}, avoiding the inconsistency of the naive approach. By setting a prior distribution on x_{ij}^{true}, we can derive and sample the resulting posterior distribution given the observed data x_{ij}^{obs}, i.e., $\pi(x_{ij}^{\text{true}} \mid x_{ij}^{\text{obs}})$. This posterior distribution captures all possible variations of x_{ij}^{true} given the data. Then, we can easily generate a replicate from a Gaussian distribution

centered at each possible realization of x_{ij}^{true} given the data. This is consistent with the assumption behind the Gaussian measurement error model. In addition, by accounting for all possible values of the unknown true value, x_{ij}^{true}, the uncertainty of x_{ij}^{true} is naturally reflected in the replicates. This is what the following posterior predictive distribution does by encoding the uncertainty of x_{ij}^{true} into the replicates given the observed data:

$$q(x_{ij}^{\text{rep}} \mid x_{ij}^{\text{obs}}) = \int f(x_{ij}^{\text{rep}} \mid x_{ij}^{\text{true}})\ \pi(x_{ij}^{\text{true}} \mid x_{ij}^{\text{obs}})\ dx_{ij}^{\text{true}}.$$

We note that the proposed Gaussian perturbation in Equation (3.4) is a specific posterior predictive distribution obtained with the improper flat prior distribution on x_{ij}^{true}. Different priors on x_{ij}^{true} lead to different posterior predictive distributions with possibly more complicated sampling steps.

A potential extension of the current work is to adopt a proper prior distribution on x_{ij}^{true} based on physical knowledge. A two-level Gaussian hierarchical model (Efron & Morris 1975) is an example of modeling a population distribution of each known class via a Gaussian prior distribution. For example, the unknown true values x_{ij}^{true} in class 1 are assumed to be from one Gaussian population distribution, and those in class 2 from another Gaussian population distribution, etc. However, this elaborate modeling approach requires more computation prior to perturbation by sampling the more complicated posterior distribution of x_{ij}^{true} before we sample the replicates x_{ij}^{rep}'s.

6.2 The number of perturbed data sets

One important question regarding the proposed Gaussian perturbation is how we determine the number of perturbed data sets, R, in practice. In principle, the larger the value of R is, the smaller the resulting Monte Carlo error is. This is because the Monte Carlo error converges to zero at rate $O(1/\sqrt{R})$ (Liu 2008, chap. 1.1). Thus, it needs to be chosen large enough to capture the shape of the posterior predictive distribution without missing any important feature of the distribution, e.g., heavy-tailedness or multimodality.

In the bootstrapping literature, it is known that 200 replicates of the data are typically needed and even 25 are usually informative (Efron & Tibshirani 1994, pp. 48 and 52). Since the current work is a resampling method similar to parametric bootstrapping, setting $R = 200$ might be large enough. In our numerical studies, we set $R = 500$ because 500 pseudo-data sets have been enough to capture the shapes of posterior predictive distributions, as shown in Table 4.1 and Figure 5.1. Though not reported here, we have

confirmed that more perturbed data sets (e.g., $R = 1000$) do not change the shape of each posterior predictive distribution meaningfully.

6.3 Correlated measurement error

In this work, we assume that measurement error in each feature is independent of other features because the measurements of p features come with only p measurement error uncertainties. This does *not* mean that the correlation does not exist. In fact, modeling correlations among measurement errors is not unusual in astronomy, e.g., Sereno (2016) for linear regression.

One particular difficulty arises in dealing with correlated measurement errors. The information about correlations among the p measurement errors is necessary for constructing a full covariance structure of multiple measurement errors across features. But this information is typically not given in the data, which means that we need to *estimate* all of the $p(p-1)/2$ correlations from the data. We note that these correlations are essentially the same as those among p measurements (given the true values) because

$$\begin{aligned}\mathrm{Corr}(x_{ij}^{\mathrm{obs}},\ x_{ij'}^{\mathrm{obs}}) &= \mathrm{Corr}(x_{ij}^{\mathrm{true}} + \epsilon_{ij},\ x_{ij'}^{\mathrm{true}} + \epsilon_{ij'}) \\ &= \mathrm{Corr}(\epsilon_{ij},\ \epsilon_{ij'})\end{aligned}$$

for any $j \neq j'$. Therefore, we can estimate these correlations using the sample correlations in practice.

There are possibly many ways to model these correlations within the Gaussian perturbation framework. A simple but naive approach is to construct the full $p \times p$ covariance matrix V by filling out off-diagonal elements with the estimated correlations. Then the resulting posterior predictive distribution of a perturbed observation is simply a multivariate version of (3.4), i.e.,

$$x_i^{\mathrm{rep}} \sim \mathcal{N}_p(x_i^{\mathrm{obs}},\ 2V).$$

The notation x_i^{rep} denotes $\{x_{i1}^{\mathrm{rep}}, \ldots, x_{ip}^{\mathrm{rep}}\}$ and, as defined before, $x_i^{\mathrm{obs}} = \{x_{i1}^{\mathrm{obs}}, \ldots, x_{ip}^{\mathrm{obs}}\}$. One disadvantage of this approach is that it does not properly account for the uncertainty of estimating correlations.

A Bayesian approach can be one alternative that enables modeling the correlations (or covariances) with their priors. Physical knowledge will be useful in constructing scientifically motivated priors, but flat priors between -1 and 1 would also be practical

if such information were not available. Then, the uncertainty of unknown correlations will be additionally reflected in the perturbed data sets. The resulting implementation, however, may become more computationally intensive as an inevitable cost for more elaborate Bayesian modeling.

6.4 Gaussian perturbation for clustering

The proposed Gaussian perturbation can naturally be extended to unsupervised learning problems, such as clustering analysis in astronomy. Extensive work has already been proposed relating to cluster stability evaluation via data perturbation, i.e., assessing how stable a clustering algorithm is to small perturbations in the data (Rand 1971, Breckenridge 1989, Levine & Domany 2001, Bhattacharjee et al. 2001, Zhang et al. 2020). In fact, simulating pseudo-data sets and ensembling the results have proven to be successful in achieving more robust clustering performance (Fridlyand & Dudoit 2001, Fern & Brodley 2003, Monti et al. 2003, Moeller & Radke 2006). Gaussian perturbation can serve as a novel perturbation scheme especially suitable for assessing the variability in clustering results of astronomical clustering problems. This is because, as in the case of supervised learning, most standard clustering methods do not account for heteroscedastic measurement error uncertainties given in the astronomical data either. While accuracy metrics are not available in clustering tasks as in classification, the variability of any clustering metric can be similarly assessed, such as Silhouette Coefficient (Rousseeuw 1987), Rand Index (Rand 1971), or any other measure of similarity or dispersion.

6.5 Limitations

Every statistical method has its own pros and cons, and the proposed method is no exception. The current work adopts a Gaussian measurement error model, but the Gaussian assumption may not always be sufficient to describe the complex nature of astronomical measurements well. For example, features such as color index at low signal-to-noise can have errors that are both non-Gaussian and asymmetrical (Babu & Mahabal 2016). One may require different distributional assumptions depending on how the measurements were collected and how the measurement error uncertainties were calculated. Therefore, it is desirable to extend the current work to encompass various measurement error models, such as a mixture of Gaussians and Student's t measurement errors (Tak et al. 2019).

Also, the computational cost of Gaussian perturbation increases linearly in the number of simulations, which adds onto the computational cost of a chosen classification method. In fact, the proposed approach increases the original computation cost of a classification method by a factor of R, the number of perturbed data sets. However, since any classification analysis can be independently conducted for each perturbed data set, the procedure is embarrassingly parallel over multiple cores. For example, the computational cost of the proposed approach can be restricted to that of a classification method, as long as 500 cores are available. Theoretically, the computational burden of the proposed approach can match that of a standard classification method if one has access to R CPU cores.

6.6 Concluding Remarks

Astronomical data are unusual in the sense that each measurement is accompanied by heteroscedastic measurement error whose uncertainty is known. These uncertainties are often ignored in astronomical classification problems because standard classification methods, such as support-vector machine and random forest, cannot incorporate them. This work proposes Gaussian perturbation as a simulation-based way to incorporate the measurement error uncertainties into any standard classification methods to better quantify classification uncertainty. The key idea is to simulate pseudo-data sets from a posterior predictive distribution of a Gaussian measurement error model, using the known heteroscedastic measurement error uncertainty. Then, any chosen standard classification method can be fit on each of these simulations. The resulting variation of a quantity of interest across the multiple fits naturally reflects the measurement error uncertainty as it has been propagated through every step of the procedure. We have illustrated this procedure via an extensive simulation study using SVM and random forest. Additionally, we have demonstrated its potential for astronomical applications through the problem of classifying high-z quasars from astronomical catalog data.

We note that this is not the only work raising a question about how to incorporate the unusual feature of astronomical data into astronomical data analyses, as described in the introduction. More recently, a small group of astronomers and one of the co-authors of this work had an active discussion about how to incorporate measurement error in astronomical data analyses during a three-day workshop "Petabytes to Science" held in Boston in Nov, 2019[1]. Later, one of the workshop organizers carefully examined

[1] `https://petabytestoscience.github.io/workshop-iii`

how the noises of input data propagate to a result of deep learning regression via an analytically tractable single pendulum experiment (Caldeira & Nord 2020). We hope that this work adds a momentum for the community to continue a discussion about incorporating measurement error in various contexts of astronomical data analyses.

Bibliography

Achlioptas, D. (2003), 'Database-friendly random projections: Johnson-lindenstrauss with binary coins', *Journal of computer and System Sciences* **66**(4), 671–687.

Akritas, M. G. & Bershady, M. A. (1996), 'Linear regression for astronomical data with measurement errors and intrinsic scatter', *The Astrophysical Journal* **470**, 706.

Andreon, S. & Hurn, M. (2013), 'Measurement errors and scaling relations in astrophysics: a review', *Statistical Analysis and Data Mining: The ASA Data Science Journal* **6**(1), 15–33.

Annis, J., Soares-Santos, M., Strauss, M. A., Becker, A. C., Dodelson, S., Fan, X., Gunn, J. E., Hao, J., Ivezić, Ž., Jester, S. et al. (2014), 'The sloan digital sky survey coadd: 275 deg2 of deep sloan digital sky survey imaging on stripe 82', *The Astrophysical Journal* **794**(2), 120.

Babu, G. J. & Mahabal, A. (2016), 'Skysurveys, light curves and statistical challenges', *International Statistical Review* **84**(3), 506–527.

Bhattacharjee, A., Richards, W. G., Staunton, J., Li, C., Monti, S., Vasa, P., Ladd, C., Beheshti, J., Bueno, R., Gillette, M. et al. (2001), 'Classification of human lung carcinomas by mrna expression profiling reveals distinct adenocarcinoma subclasses', *Proceedings of the National Academy of Sciences* **98**(24), 13790–13795.

Breckenridge, J. N. (1989), 'Replicating cluster analysis: Method, consistency, and validity', *Multivariate Behavioral Research* **24**(2), 147–161.

Buonaccorsi, J. P. (2010), *Measurement error: models, methods, and applications*, CRC press.

Caldeira, J. & Nord, B. (2020), 'Deeply uncertain: comparing methods of uncertainty quantification in deep learning algorithms', *Machine Learning: Science and Technology* .

Cannings, T. I. (2021), 'Random projections: Data perturbation for classification problems', *Wiley Interdisciplinary Reviews: Computational Statistics* **13**(1), e1499.

Carroll, R. J., Ruppert, D., Stefanski, L. A. & Crainiceanu, C. M. (2006), *Measurement error in nonlinear models: a modern perspective*, CRC press.

Darling, M. C. & Stracuzzi, D. J. (2018), Toward uncertainty quantification for supervised classification, Technical report, Sandia National Lab.(SNL-NM), Albuquerque, NM (United States).

Eddington, A. (1913), 'On a formula for correcting statistics for the effects of a known error of observation', *Monthly Notices of the Royal Astronomical Society* **73**, 359–360.

Efron, B. (1992), Bootstrap methods: another look at the jackknife, *in* 'Breakthroughs in statistics', Springer, pp. 569–593.

Efron, B. & Morris, C. (1975), 'Data analysis using stein's estimator and its generalizations', *Journal of the American Statistical Association* **70**(350), pp. 311–319.
URL: *http://www.jstor.org/stable/2285814*

Efron, B. & Tibshirani, R. J. (1994), *An introduction to the bootstrap*, CRC press.

Feigelson, E. (2012), 'Classification in astronomy: Past and present', *Advances in Machine Learning and Data Mining for Astronomy* pp. 3–10.

Feigelson, E. & Babu, G. (1998), Statistical methodology for large astronomical surveys, *in* 'Symposium-International Astronomical Union', Vol. 179, Cambridge University Press, pp. 363–370.

Feigelson, E. D., De Souza, R. S., Ishida, E. E. & Babu, G. J. (2021), 'Twenty-first-century statistical and computational challenges in astrophysics', *Annual Review of Statistics and Its Application* **8**, 493–517.

Fern, X. Z. & Brodley, C. E. (2003), Random projection for high dimensional data clustering: A cluster ensemble approach, *in* 'Proceedings of the 20th international conference on machine learning (ICML-03)', pp. 186–193.

Fridlyand, J. & Dudoit, S. (2001), Applications of resampling methods to estimate the number of clusters and to improve the accuracy of a clustering method, Technical report, Technical Report 600, Department of Statistics, UC Berkeley.

Fukugita, M., Shimasaku, K., Ichikawa, T., Gunn, J. et al. (1996), The sloan digital sky survey photometric system, Technical report, SCAN-9601313.

Fuller, W. A. (1987), *Measurement error models*, Vol. 305, John Wiley & Sons.

Gelman, A., Carlin, J. B., Stern, H. S., Dunson, D. B., Vehtari, A. & Rubin, D. B. (2013), *Bayesian data analysis*, CRC press.

Hashemi, M. & Karimi, H. (2018), 'Weighted machine learning', *Statistics, Optimization and Information Computing* **6**(4), 497–525.

He, H. & Garcia, E. A. (2009), 'Learning from imbalanced data', *IEEE Transactions on knowledge and data engineering* **21**(9), 1263–1284.

Hobert, J. P. & Casella, G. (1996), 'The effect of improper priors on gibbs sampling in hierarchical linear mixed models', *Journal of the American Statistical Association* **91**(436), 1461–1473.
URL: *http://www.jstor.org/stable/2291572*

Hoefsloot, H. C., Verouden, M. P., Westerhuis, J. A. & Smilde, A. K. (2006), 'Maximum likelihood scaling (mals)', *Journal of Chemometrics: A Journal of the Chemometrics Society* **20**(3-4), 120–127.

Hogg, D. W. & Turner, E. L. (1998), 'A maximum likelihood method to improve faint-source flux and color estimates', *Publications of the Astronomical Society of the Pacific* **110**(748), 727.

Hu, Z. & Tak, H. (2020), 'Modeling stochastic variability in multiband time-series data', *The Astronomical Journal* **160**(6), 265.

Jiang, L., Fan, X., Bian, F., McGreer, I. D., Strauss, M. A., Annis, J., Buck, Z., Green, R., Hodge, J. A., Myers, A. D. et al. (2014), 'The sloan digital sky survey stripe 82 imaging data: Depth-optimized co-adds over 300 deg2 in five filters', *The Astrophysical Journal Supplement Series* **213**(1), 12.

Kelly, B. C. (2007), 'Some aspects of measurement error in linear regression of astronomical data', *The Astrophysical Journal* **665**(2), 1489.

Kelly, B. C., Bechtold, J. & Siemiginowska, A. (2009), 'Are the Variations in Quasar Optical Flux Driven by Thermal Fluctuations?', *The Astrophysical Journal* **698**(1), 895–910.

Kelly, B. C., Becker, A. C., Sobolewska, M., Siemiginowska, A. & Uttley, P. (2014), 'Flexible and scalable methods for quantifying stochastic variability in the era of massive time-domain astronomical data sets', *The Astrophysical Journal* **788**(1), 33.

Kogan, G. (1997), 'Gamma ray burst source statistics in the presence of stochastic errors', *Astronomy and Astrophysics* **324**.

Kuhn, M. (2021), *caret: Classification and Regression Training*. R package version 6.0-90.
URL: *https://CRAN.R-project.org/package=caret*

Lapin, M., Hein, M. & Schiele, B. (2014), 'Learning using privileged information: Svm+ and weighted svm', *Neural Networks* **53**, 95–108.

Levine, E. & Domany, E. (2001), 'Resampling method for unsupervised estimation of cluster validity', *Neural computation* **13**(11), 2573–2593.

Liu, J. S. (2008), *Monte Carlo Strategies in Scientific Computing*, Springer, New York, NY.

Luo, K. (2019), Classification with Measurement Error in Covariates Or Response, with Application to Prostate Cancer Imaging Study, PhD thesis, The University of Western Ontario.

Malossini, A., Blanzieri, E. & Ng, R. T. (2006), 'Detecting potential labeling errors in microarrays by data perturbation', *Bioinformatics* **22**(17), 2114–2121.

Moeller, U. & Radke, D. (2006), 'Performance of data resampling methods for robust class discovery based on clustering', *Intelligent Data Analysis* **10**(2), 139–162.

Monti, S., Tamayo, P., Mesirov, J. & Golub, T. (2003), 'Consensus clustering: a resampling-based method for class discovery and visualization of gene expression microarray data', *Machine learning* **52**(1), 91–118.

Napierala, K. & Stefanowski, J. (2015), 'Abstaining in rule set bagging for imbalanced data', *Logic Journal of the IGPL* **23**(3), 421–430.

Papovich, C., Shipley, H. V., Mehrtens, N., Lanham, C., Lacy, M., Ciardullo, R., Finkelstein, S. L., Bassett, R., Behroozi, P., Blanc, G. A., de Jong, R. S., DePoy, D. L., Drory, N., Gawiser, E., Gebhardt, K., Gronwall, C., Hill, G. J., Hopp, U., Jogee, S., Kawinwanichakij, L., Marshall, J. L., McLinden, E., Mentuch Cooper, E., Somerville, R. S., Steinmetz, M., Tran, K. V., Tuttle, S., Viero, M., Wechsler, R. & Zeimann, G. (2016), 'The Spitzer-HETDEX Exploratory Large-area Survey', *The Astrophysical Journal Supplement Series* **224**(2), 28.

Petrosian, V. (1992), Luminosity function of flux-limited samples, *in* 'Statistical challenges in modern Astronomy', Springer, pp. 173–194.

R Core Team (2020), *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria.
URL: *https://www.R-project.org/*

Rand, W. M. (1971), 'Objective criteria for the evaluation of clustering methods', *Journal of the American Statistical association* **66**(336), 846–850.

Richards, G. T., Myers, A. D., Peters, C. M., Krawczyk, C. M., Chase, G., Ross, N. P., Fan, X., Jiang, L., Lacy, M., McGreer, I. D. et al. (2015), 'Bayesian high-redshift quasar classification from optical and mid-ir photometry', *The Astrophysical Journal Supplement Series* **219**(2), 39.

Rousseeuw, P. J. (1987), 'Silhouettes: a graphical aid to the interpretation and validation of cluster analysis', *Journal of computational and applied mathematics* **20**, 53–65.

Sereno, M. (2016), 'A bayesian approach to linear regression in astronomy', *Monthly Notices of the Royal Astronomical Society* **455**(2), 2149–2162.

Sun, W. (2015), Stability of machine learning algorithms, PhD thesis, Purdue University.

Tak, H., Ellis, J. A. & Ghosh, S. K. (2019), 'Robust and accurate inference via a mixture of gaussian and student's t errors', *Journal of Computational and Graphical Statistics* **28**(2), 415–426.

Tak, H., Ghosh, S. K. & Ellis, J. A. (2018), 'How proper are bayesian models in the astronomical literature?', *Monthly Notices of the Royal Astronomical Society* **481**(1), 277–285.

Timlin, J. D., Ross, N. P., Richards, G. T., Lacy, M., Ryan, E. L., Stone, R. B., Bauer, F. E., Brandt, W. N., Fan, X., Glikman, E., Haggard, D., Jiang, L., LaMassa, S. M., Lin, Y.-T., Makler, M., McGehee, P., Myers, A. D., Schneider, D. P., Urry, C. M., Wollack, E. J. & Zakamska, N. L. (2016), 'SpIES: The Spitzer IRAC Equatorial Survey', *The Astrophysical Journal Supplement Series* **225**(1), 1.

Timlin, J. D., Ross, N. P., Richards, G. T., Myers, A. D., Pellegrino, A., Bauer, F. E., Lacy, M., Schneider, D. P., Wollack, E. J. & Zakamska, N. L. (2018), 'The clustering of high-redshift ($2.9 \leq z \leq 5.1$) quasars in sdss stripe 82', *The Astrophysical Journal* **859**(1), 20.

van den Berg, R. A., Hoefsloot, H. C., Westerhuis, J. A., Smilde, A. K. & van der Werf, M. J. (2006), 'Centering, scaling, and transformations: improving the biological information content of metabolomics data', *BMC genomics* **7**(1), 1–15.

Von Luxburg, U. et al. (2010), 'Clustering stability: an overview', *Foundations and Trends® in Machine Learning* **2**(3), 235–274.

Waaijenborg, S., Korobko, O., Willems van Dijk, K., Lips, M., Hankemeier, T., Wilderjans, T. F., Smilde, A. K. & Westerhuis, J. A. (2018), 'Fusing metabolomics data sets with heterogeneous measurement errors', *PloS one* **13**(4), e0195939.

Wahba, G. (2002), 'Soft and hard classification by reproducing kernel hilbert space methods', *Proceedings of the National Academy of Sciences* **99**(26), 16524–16530.

Yu, B. et al. (2013), 'Stability', *Bernoulli* **19**(4), 1484–1500.

Zhang, L., Lin, L. & Li, J. (2020), 'Cps analysis: self-contained validation of biomedical data clustering', *Bioinformatics* **36**(11), 3516–3521.