PATTERN RECOGNITION FOR RESTRICTED AND NONSTATIONARY DATA

A Dissertation

by

SHUILIAN XIE

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Ulisses M. Braga-Neto
Committee Members,	Edward R. Dougherty Jr
	Krishna R. Narayanan
	Alan Dabney
Head of Department,	Aniruddha Datta

May 2021

Major Subject: Electrical Engineering

ABSTRACT

The standard assumption in classification is that the training data are independent and identically distributed. Indeed, this assumption is so pervasive that it is often applied without mention. In this dissertation, we propose novel methods that address violations of this standard assumption corresponding to 1) restricted sampling and 2) a nonstationary environment.

The first part of this dissertation concerns the bias of classification precision estimation under restricted sampling. Precision and recall have become very popular classification accuracy metrics in the statistical learning literature, under the standard i.i.d. sampling assumption. However, in many cases of interest, as in observational case-control studies for biomarker discovery in cancer studies, the training data are sampled separately from the case and control populations, violating the standard sampling assumption, under which the data is sampled randomly from the mixture of the populations. We present an analysis of the bias in the estimation of the precision of classifiers designed on separately sampled data. The analysis consists of both theoretical and numerical results, which show that classifier precision estimates can display strong bias under separating sampling, with the bias magnitude depending on the difference between the true case prevalence in the population and the sample prevalence in the data. We show that this bias is systematic in the sense that it cannot be reduced by increasing sample size. If information about the true case prevalence is available from public health records, then we propose the use of a modified precision estimator based on the known prevalence that displays smaller bias, which can in fact be reduced to zero as sample size increases under regularity conditions on the classification algorithm. The accuracy of the theoretical analysis and the performance of the precision estimators under separate sampling are confirmed by numerical experiments using synthetic and real data from published observational case-control studies. The results with real data confirmed that under separately-sampled data, the usual precision estimator produces larger, i.e. more optimistic, estimates than the estimator using the true prevalence value.

The second part of this dissertation proposes a state space model approach to classification of nonstationary data. In many applications, the data are collected at different time points. If the time between consecutive acquisition points is large enough, the distribution of data is likely to shift due to natural physical processes, and the standard i.i.d. sampling assumption is violated. This has been known in the statistical learning literature as "population drift" problem. Most attempts to address nonstationarity are ad-hoc and carry no guarantee of optimality. In this dissertation, we propose a state-space methodology, whereby the data are assumed to evolve linearly or nonlinearly under Gaussian observation noise, and applied adaptive filtering methods to estimate the distributional parameters, leading to nonstationary linear and quadratic discriminant analysis (NSLDA and NSQDA) classification rules. Parameter estimation in the linear state-space model is accomplished by a combination of Kalman smoothing and maximum-likelihood estimation by expectation maximization, while particle filtering methods are proposed for the nonlinear state-space model. We have also addressed the case where the time labels of some data are unknown, a situation that often arises in practice, by proposing a hybrid Gaussian mixture modeling (GMM)-Kalman Smoother approach. The accuracy of the proposed nonstationary discriminant analysis rule, as well as its robustness against noise, missing data, and unbalanced training data are demonstrated in numerical experiments, where we compare it to "naive" LDA, QDA, and nonlinear SVM classification rules.

DEDICATION

To my family.

ACKNOWLEDGMENTS

First, I would like to thank my parents, Yanrui and Dinge, and my sisters, Lulu and Jinping, for their financial support and unconditional love. I thank my husband, Cale Scholl, for his non-stop love and support. Without them, I would never have achieved one of the highest goals in my life. I also dedicate this successful educational project to my friends, Shan Wang, Mengyuan Zhang, Liang Liang, Han Xu and Yuan Zhang. Without their help and encouragement, I could not have been motivated and positive.

I would like to take this opportunity to express my deepest gratitude to my Ph.D advisor, Dr. Ulisses Braga-Neto, who made my dissertation possible. Ever since the first days when he taught me Pattern Recognition, Ulisses has been an outstanding mentor and role-model. I thank his brilliant insight, inspiring discussion, patient guidance and generous financial support throughout my Ph.D study. Working with him was indeed a pleasant and rewarding experience. I also want to thank my committee members, Dr. Dougherty, Dr. Narayanan, and Dr. Dabney, for their stimulating discussion during my preliminary examination. I would specially thank Dr. Narayanan, who was also my undergraduate research advisor, for his supervision during my early research experience. I also would like to thank all the members in our research lab, for friendship and memories. Special thanks to Dr. Mahdi Imani, who is a true friend and great collaborator with me.

Last but not the least, I would like to thank all the faculty at Texas A&M University, especially for their interesting and stimulating courses that I took. My learning and life experience at Texas A&M University will definitely benefit the rest of my career.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Professor Braga-Neto and Professors Dougherty and Narayanan of the Department of Electrical & Computer Engineering and Professor Dabney of Department of Statistics. All work for the dissertation was completed by the student, under the advisement of Professor Braga-Neto of the Department of Electrical & Computer Engineering.

Funding Sources

I was supported by the department of Electrical and Computer Engineering, being hired as a Teaching Assistant for four semesters; for the remaining three years of study, I was supported by the National Science Foundation, through NSF award CCF-1718924.

NOMENCLATURE

FP	False Positive
FN	False Negative
TP	True Positive
TN	True Negative
Err	Error
Acc	Accuracy
Prev	Prevalence
Spec	Specificity
Sens	Sensitivity
Prec	Precision
Rec	Recall
AML	Acute Myeloid Leukemia
ALL	Acute Lymphoblastic Leukemia

TABLE OF CONTENTS

LIST OF FIGURES

FIGURE Page

LIST OF TABLES

1. INTRODUCTION

In statistics, a standard assumption is that sampling is unrestricted and stationary, i.e., the sample is independent and identically distributed [1,2]. See Fig. 1.1.

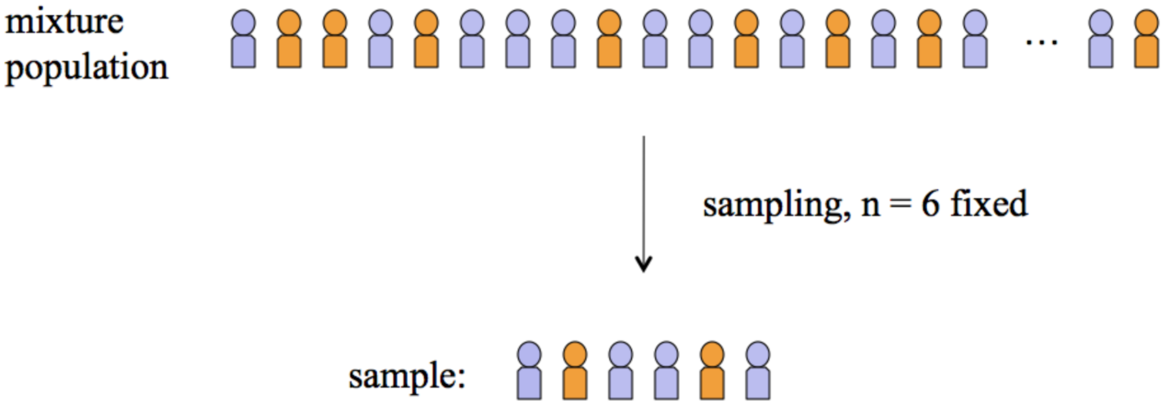


Figure 1.1: Independent and identically distributed sampling example.

Let

$$c_i = \text{“true” proportion of class } i \text{ in the mixture population}$$

$$N_i = \text{number of individuals from class } i \text{ in the sample.}$$

Under random sampling, an i.i.d. sample is drawn from the *mixture* of the populations Π_0 and Π_1. This means that if a sample of size n is drawn for binary classification, then the numbers of sample point $N_i \sim \text{Binomial}(n, c_i)$, and $\hat{c}_i = N_i/n$ is a consistent estimator of c_i, for $i = 0, 1$ (also consistent, by the Bernoulli's Law of Large Numbers [3,4]). Like this example, a consistent estimator of c_0 is $1/3$ if we label 0 for the orange.

However, restriction is any constraint that creates dependencies in the data. For example, separate sampling [5] is common in observational case-control studies in biomedicine [6,7]. That is to say suppose the sampling is not random, in the sense that the ratio $r = \frac{n_0}{n}$ and $1 - r = \frac{n1}{n}$ are chosen prior to the sampling procedure. See Fig. 1.2. Here $n_0 = 4$ and $n_1 = 2$ are not realizations of binomial random variables, but are fixed parameters prior to sampling. Hence, $\hat{c}_0 = n_0/n$ and $\hat{c}_1 = n_1/n$ are not estimators of c_0 and c_1 in any useful sense. However, the inability to consistently estimate c_i matters!

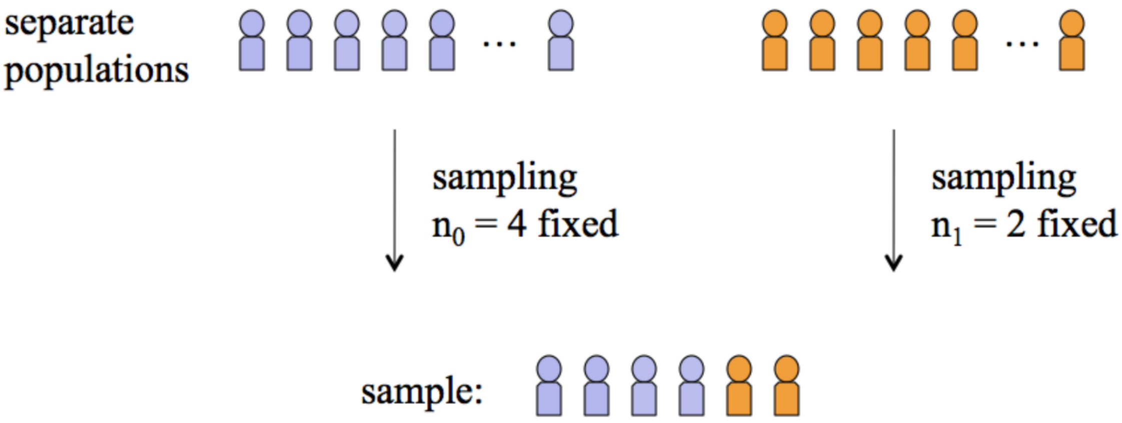


Figure 1.2: Separate sampling example.

Another example of i.i.d. assumption being violated is stationarity. In many time series techniques, being stationary series means data statistical properties like mean, variance, covariance is not changing over time. We could compare these two figures in Fig. 1.3: in the stationary case, mean needs to be constant; while in the nonstationary case, mean depends on time. However, the stationary assumption is likely to be violated if the time between consecutive acquisition points is large enough; then population drift problem should be solved.

A simple example of nonstationary data is - heterogeneous data [8] collected at different time points - when there is population drift, shown in Fig. 1.4. The heterogeneous sample data at time k are given by

$$S_{n[k]} = S_{m[0]} \cup S_{m[1]} \cdots \cup S_{m[k]} \tag{1.1}$$

where $S_{m[k]}$ is a sample from the mixture of populations $\Pi_0[j]$ and $\Pi_1[j]$ for $j = 0, 1, ..., k$. Sample size at time k: $n[k] = m[0] + \cdots + m[k]$.

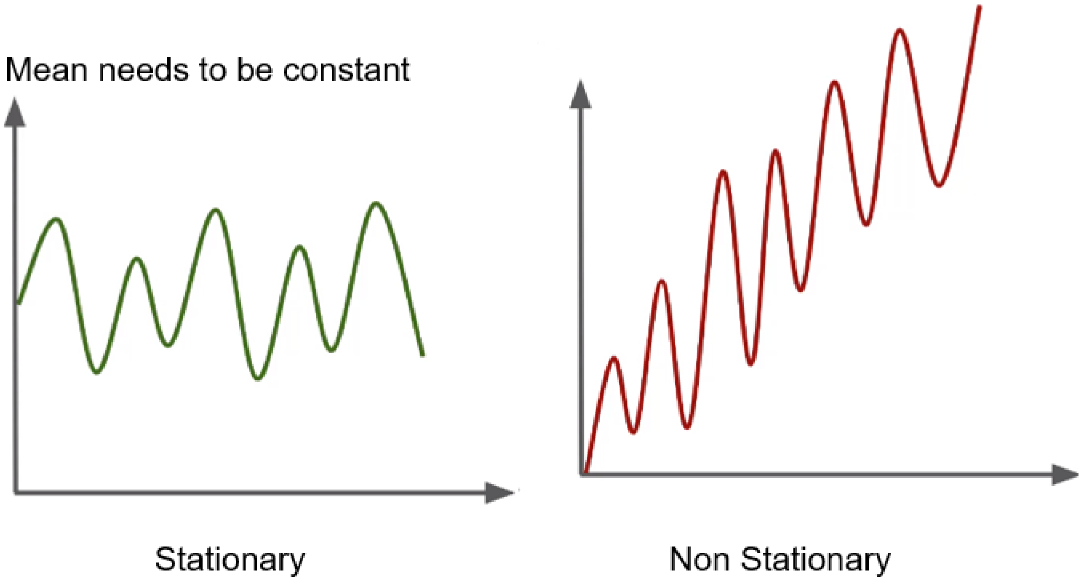


Figure 1.3: A comparison example between stationarity and nonstationarity.

Now suppose that the populations are multivariate Gaussian with means $\boldsymbol{\mu}_0\left[k\right]$ and $\boldsymbol{\mu}_1\left[k\right]$ and common covariance matrix $\Sigma\left[k\right]$. Assume further that the means evolve according to a first-order linear equation. By stacking the vectors appropriately, we can write the state-space equation:

$$\boldsymbol{\mu}\left[k+1\right] = A\boldsymbol{\mu}\left[k\right] + \mathbf{w}\left[k\right] \tag{1.2}$$

$$\mathbf{x}_i\left[k\right] = \boldsymbol{\mu}\left[k\right] + \mathbf{v}_i\left[k\right], i = 1, ..., m\left[k\right] \tag{1.3}$$

for k = 0, 1, ... where the transition noise $\mathbf{w}\left[k\right]$ and observation noise $\mathbf{v}_i\left[k\right]$ are zero-mean Gaussian random vectors with covariance matrices R and Σ, respectively.

In the figure, the naive decision boundary is simply made by a Quadratic Discriminant Analysis classifier; while one can draw a Linear Discriminant Analysis classifier boundary at each time point [9]. Even though both decision boundaries make perfect separate for data with different labels, the one doesn't consider nonstationary property fail to explain the data well, and is very likely to give bad accuracy of test dataset.

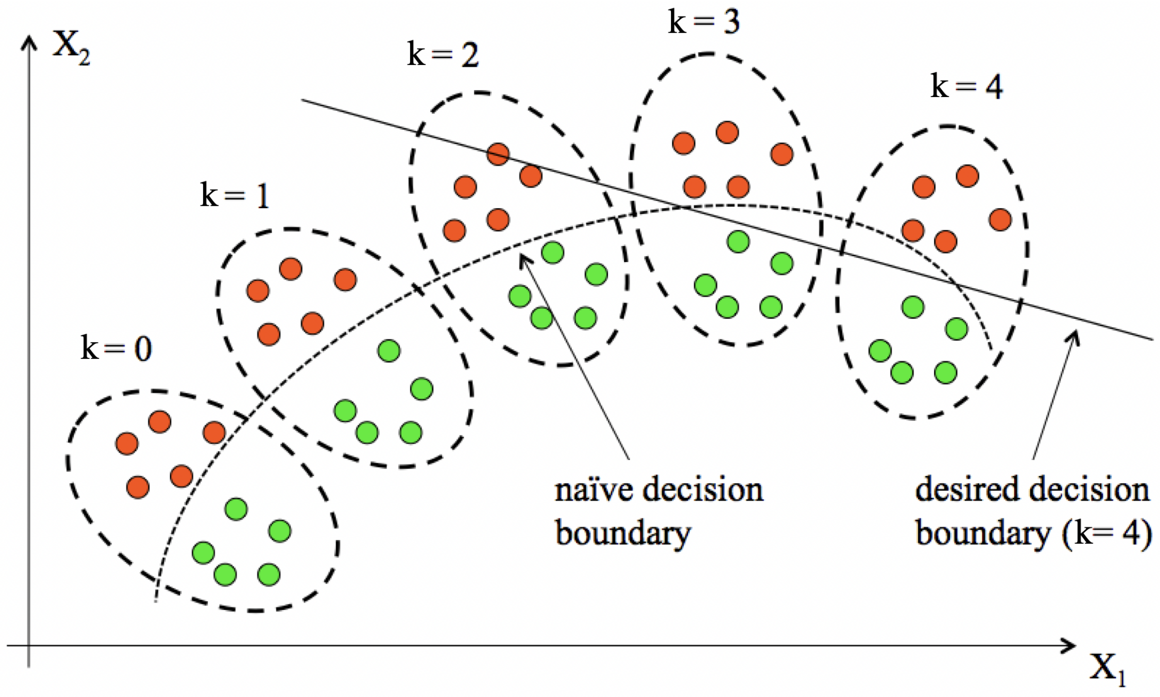


Figure 1.4: An example of nonstationary classification.

1.1 On the Bias of Precision Estimation under Separate Sampling

1.1.1 Background

Biomarker discovery is typically attempted by means of observational case-control studies where classification techniques are applied to high-throughput measurement technologies, such as DNA microarrays [10, 11], next-generation RNA sequencing (RNA-seq) [12], or "shotgun" mass spectrometry [13]. The validity and reproducibility of the results depend critically on the availability of accurate and unbiased assessment of classification accuracy [14, 15].

The vast majority of published methods in the statistical learning literature make the assumption, explicitly or implicitly, that the data for training and accuracy assessment are sampled randomly, or unrestrictedly, from the mixture of the populations. However, observational case-control studies in biomedicine typically proceed by collecting data that are sampled with restrictions. The most common restriction, and the one that is studied in this dissertation, is that the data are sampled separately from the case and control populations. This is always true in studies involving rare diseases, since sampling randomly from the population at large would not yield enough cases. That creates an important issue in the application of traditional statistical learning techniques to biomedical data, because there is no meaningful estimator of case prevalences under separate sampling. Therefore, any methodology that directly or indirectly uses estimates of case prevalence will be severely biased.

Precision and *Recall* have become very popular classification accuracy metrics in the statistical learning literature [16–18]. In practice, these quantities must be estimated from sample data. The recall does not depend on the prevalence, while the precision does. Therefore, in this dissertation, we investigate the bias of precision estimates when the typical separate sampling design used in case-control studies is not properly taken into account.

A similar study was conducted previously into the accuracy of cross-validation under separate sampling [19]. It was shown in that study that the usual "unbiasedness" property of k-fold cross-validation does not hold under separate sampling. In fact, the bias can in fact be substantial and systematic, i.e., not reducible under increasing sample size. In [19], modified k-fold cross-validation estimators were proposed for the class-specific error rates. In the case where the true case prevalence is known, those estimators can be combined into an estimator of the overall error rate, which satisfies the usual "unbiasedness" property of cross-validation.

By contrast, the present research study employs analytical and numerical methods to investigate precision estimation under separate sampling. We show that the usual precision estimator is asymptotically unbiased as sample size increases, under the condition that the classification rule has a finite VC dimension. However, under separate sampling, we show that the usual precision estimator will in general display a systematic bias, which cannot be reduced by increasing sample size, if the observed prevalence of cases in the data is different from the true prevalence in the population of interest, and the bias is larger the more different they are. In particular, the bias tends to be large when the true prevalence is small but the training data contains an equal number of examples from both classes, which is a common scenario in practice. If the true case prevalence is known (e.g., from public health records), then a modified precision estimator that uses the known prevalence is shown to be asymptotically unbiased in the separate sampling case, under the condition that the classification rule is sufficiently stable as sample size increases. All of these theoretical results, and the approximations used to derive them, are verified by numerical experiments using both synthetic and real data from published studies.

1.1.2 Summary of Contributions

This project employs analytical and numerical methods to show that the ordinary precision estimator can display large bias under separate sampling. This is a consequence of the fact that precision is a function of the true prevalence. Case-control studies involving rare diseases are specially affected, since in those studies the true prevalence is small and will almost always differs substantially from the observed prevalence in the data. To address this problem, we propose a modified estimator for precision, which can be applied in case the true prevalence is known. This estimator has small bias that vanishes as sample size increases under certain regularity conditions. In the absence of any knowledge about the true prevalence, precision estimates should be avoided under separate sampling.

1.2 State Space Models to Nonstationary Discriminant Analysis

1.2.1 Background

The standard assumption of statistical learning is that the training data are identically and independently distributed [20–22]. Identical distribution implies that the population is *stationarity*, meaning that it does not change over time. This assumption is likely to be violated in modern "big data" applications of machine learning, including natural language processing, speech recognition, image recognition, and bioinformatics. In these real-world applications, data are complex, with different training points being acquired at different time points. If the time between consecutive acquisition points is large enough, the distribution of the data is likely to shift due to natural physical processes. In practice, there are many circumstances where the stationary assumption is unwarranted because the underlying physical processes are strongly non-stationary, for instance, in the case of tumor classification where genetic structure evolves over time. This kind of behavior has been termed "population drift" in the literature [23], where population drifting means that whole population is assumed to be homogeneous with respect to the distribution of the response variable conditioned on the feature space. Similar problem has been discussed and addressed in [24] and [25].

Changes in population distribution over time is a challenging problem in supervised learning. The book [26] focuses on a specific nonstationary environment known as covariate shift, in which the distributions of inputs (queries) change but the conditional distribution of outputs (answers) is unchanged. [27] discusses the detection and adaptation in nonstationary environment. Once a change has been detected, the classifier needs to adapt the change by learning from the newly available information, and discarding the obsolete one. Previous attempts to address nonstationarity in statistical learning, also known as "population drift," are ad-hoc and carry no guarantee of optimality. Our methodology provides a optimal solution to solve the problem of machine learning application in nonstationary environment, where we utilized the state-space approach. In particular, we combined Kalman Smoother algorithm [28] with Linear Discriminant Analysis (LDA) [29] in different scenarios.

Formally, the standard assumption in binary classification is that there is a feature-label distribution $f(\mathbf{x}, y)$ with classes described by the class-conditional distributions $f(\mathbf{x}|0)$ and $f(\mathbf{x}|1)$, the aim being to construct a classifier $\psi : \mathcal{R}^d \rightarrow \{0,1\}$, with small error rate $\varepsilon[\psi] = P(\psi(\mathbf{X}) \neq Y)$, via some classification rule utilizing sample data, the implicit assumption being that the feature-label distribution is stationary, that is, fixed over time. Linear Discriminant Analysis (LDA) [29] is a simple classification rule, which presents a low degree of overfitting, and is therefore useful for application in small-sample cases. Provided that the underlying feature-label distribution is linearly separable, LDA produces very good classifiers.

1.2.1.1 Nonstationary Discriminant Analysis Overview

In a nonstationary classification problem, there is a feature vector $\mathbf{X}_k \in R^d$ and a label $Y_k \in \{0, ..., c-1\}$ defined at each discrete time $k = 0, 1, \ldots$ (this results from sampling the corresponding continuous-time stochastic processes at discrete points in time). Let $f_k^j(\mathbf{x} \mid Y_k = j)$ be the class conditional distributions and $\pi_k^j = P(Y_k = j)$ be the prior probabilities at time step k. The population drift reflects itself in the changing f_k^j and π_k^j. It is easy to show that an optimal classifier *at each time point* k is given by

$$\psi_k^*(\mathbf{x}) = \operatorname*{argmax}_{j=0,1,\ldots,c-1} D_k^{*,j}(\mathbf{x})\,, \tag{1.4}$$

where the *optimal discriminants* are given by

$$D_k^{*,j}(\mathbf{x}) = \log \pi_k^j + \log f_k^j(\mathbf{x})\,, \tag{1.5}$$

for $j = 0, 1, \ldots, c-1$ and $k = 0, 1, \ldots$ Traditional classification makes the stationarity assumption $f_k^j \equiv f_j$ and $\pi_k^j \equiv \pi_j$, in which case there is a single optimal classifier ψ^*, which is not a function of time.

Typically, the distributional quantities necessary to compute optimal classifiers are not known or only partially known, and (sub-optimal) classifiers need to be designed with the help of sample data. In nonstationary classification, classifiers $\psi_{n,k}$ are designed from data S_n of sample size n to approximate the optimal classification error of ψ_k^* at *a point of time* k where the classifier is supposed to be deployed. This could be a time in the past (in historical studies), but more commonly the present or a future time. The sample data $S_n = \bigcup_k S_k$ consists of a collection of i.i.d. samples $S_k = \{(\mathbf{X}_{k,1}, Y_{k,1}), \ldots, (\mathbf{X}_{k,n_k}, Y_{k,n_k})\}$ from the mixture $\sum_{j=1}^{c-1} \pi_k^j f_k^j$, where $n = \sum_k n_k$ (in practice, data for only a finite number of present and past time points are available). The time labels for each data point could be known explicitly, but they might not be known if this information was not kept while collecting the data, in which case they would need to be estimated. A “naive” classification rule would simply ignore the time label information completely and design a single classifier based on the entire data. Alternatively, different classifiers could be designed based on each subsample S_k, if the time label is known. A nonstationary classification rule that does not ignore the time label and uses the entire data should be to be able to do better than either of these at any point of time (*including* time points not represented in the data), provided that enough data are available and an accurate model for the evolution of the distributions is known.

Consider the case where the f_k^j are multivariate Gaussian densities,

$$f_k^j(\mathbf{x} \mid Y_k = j) = \frac{1}{\sqrt{(2\pi)^d \det(\Sigma_k^j)}} \exp\left(-\frac{1}{2}||\mathbf{x} - \boldsymbol{\mu}_k^j||^2_{\Sigma_k^j}\right), \tag{1.6}$$

where $||\mathbf{v}||_M^2 = \mathbf{v}^T M^{-1}\mathbf{v}$. Ignoring constant terms, the optimal discriminants in (1.5) reduce to

$$D_k^{\mathrm{QDA},j}(\mathbf{x}) = \log \pi_k^j - \frac{1}{2}\log|\Sigma_k^j| - \frac{1}{2}||\mathbf{x} - \boldsymbol{\mu}_k^j||^2_{\Sigma_k^j}, \tag{1.7}$$

where "QDA" stands for *quadratic discriminant analysis*, in allusion to the fact that the decision boundaries produced are piecewise quadratic. Under the additional assumption that $\Sigma_k^j \equiv \Sigma_k$ at each time k, the optimal discriminants reduce to

$$D_k^{\mathrm{LDA},j}(\mathbf{x}) = \log \pi_k^j - \frac{1}{2}||\mathbf{x} - \boldsymbol{\mu}_k^j||^2_{\Sigma_k}, \tag{1.8}$$

where "LDA" stands for *linear discriminant analysis*, as in this case the decision boundaries are piecewise linear.

Nonstationary Gaussian discrimination is based on plugging in estimators $\hat{\pi}_k^j$, $\hat{\boldsymbol{\mu}}_k^j$, $\hat{\Sigma}_k$, and $\hat{\Sigma}_k^j$ based on the entire data S_n for the unknown parameters to obtain sample discriminants $D_{n,k}^{\mathrm{QDA},j}(\mathbf{x})$ and $D_{n,k}^{\mathrm{LDA},j}(\mathbf{x})$, and the corresponding sample-based QDA and LDA classifiers:

$$\psi_{n,k}^{\mathrm{QDA}}(\mathbf{x}) = \operatorname*{argmax}_{j=0,1,\ldots,c-1} D_{n,k}^{\mathrm{QDA},j}(\mathbf{x}) \tag{1.9}$$

and

$$\psi_{n,k}^{\mathrm{LDA}}(\mathbf{x}) = \operatorname*{argmax}_{j=0,1,\ldots,c-1} D_{n,k}^{\mathrm{LDA},j}(\mathbf{x}). \tag{1.10}$$

The quality of these classifiers depend on how accurate the parameter estimators are. Indeed, it can be shown [30, Thm. 4.1] that if, for a given k, these estimators converge in probability to the corresponding parameters as sample size increases, then the classification error of $\psi_{n,k}$ converges in probability to the error of the optimal classifier ψ_k^*. In the rest of this paper, we discuss how to define accurate estimators $\hat{\pi}_k^j$, $\hat{\boldsymbol{\mu}}_k^j$, $\hat{\Sigma}_k$, and $\hat{\Sigma}_k^j$ based on linear and nonlinear state-space models of distributional drift.

1.2.1.2 Linear Drift Model Overview

Let us consider a classification problem, in which $\mathbf{j} = \{0, 1, ..., c - 1\}$ denote the set of all classes. Assume $\Pi_k^j = f_k(\mathbf{x}|j)$ is the class conditional distribution of class j at time step k. In non-stationary condition, the class conditional distributions are function of time. Considering the linear model for evolution of the class conditional distributions, we have:

$$\boldsymbol{\mu}_k^j = A^j \boldsymbol{\mu}_{k-1}^j + \mathbf{w}_k^j \,, \tag{1.11}$$

for $k = 1, 2, ...$; where $\mathbf{w}_k^j$is independent zero-mean Gaussian noise processes with known covariances matrices of Q^j, and A^j is state transition matrices for class j.

The data are assumed to be generated through the following measurement process:

$$\mathbf{x}_k^j = C^j \boldsymbol{\mu}_k^j + \mathbf{v}_k^j \,, \tag{1.12}$$

where $\mathbf{v}_k^j \sim \mathcal{N}\left(\mathbf{0}, R^j\right)$, and C^j denote the dynamics of measurement process for class j.

In this project, the ultimate goal is to developed the Linear Discriminant Analysis (LDA) for nonstationary condition given only c sets of measurements. A conventional approach for estimating the mean vectors and covariance matrices of the class conditional distributions is to use the classical Kalman smoother (KS) [31]. Two major cases are considered in this linear model case:

- *Systems with Fully-Known Dynamics:* In this case, all matrices in equations (1.11) and (1.12) are assumed to be known, and the Kalman Smoother algorithm can be directly applied to obtain estimates of the class means $\{\boldsymbol{\mu}_k^j; k = 1, ..., T\}$. A modified version of Kalman Smoother with multiple measurements at one time is discussed in Chapter 3.

- *Systems with Partially-Known Dynamics:* In this case, parameters in equations (1.11) and (1.12) are only partially known. In the conventional case, the well-known maximum-likelihood or Bayesian techniques can be used for estimating these unknown parameters [32–34]. Several modified versions of Kalman Smoother for state space model with unknown parameters are discussed in Chapter 3.

The outputs from the modified Kalman Smoother frameworks are used in our proposed Nonstationary Discriminant Analysis. Performance is assessed in a set of numerical experiments using simulated data, where the average error rates obtained by NSLDA are compared to the error produced by a naive application of LDA to the pooled nonstationary data.

1.2.1.3 Nonlinear Drift Model Overview

In a multiclass nonstationary problem with c classes and T time points, we assume that the centroid of each class is a latent variable that evolves in time according to the following nonlinear model:

$$\mathbf{z}_k^j = \mathbf{f}_k^j\left(\mathbf{z}_{k-1}^j, \mathbf{w}_k^j\right), \tag{1.13}$$

for $j = 0, 1, \ldots, c-1$ and $k = 1, \ldots, T$, where $\mathbf{f}_k^j$ is an arbitrary nonlinear function governing the evolution of class j and $\mathbf{w}_k^j$ defines an i.i.d. transition noise process, which is independent of the $\mathbf{z}_k^j$ process. The initial states $\mathbf{z}_0^j$ are generated from given starting “prior” distributions.

For notational simplicity, we partition the training data into $c \times T$ subsamples

$$S_k^j = \{\mathbf{x}_{k,1}^j, \ldots, \mathbf{x}_{k,n_k^j}^j\}, \tag{1.14}$$

for $j = 0, \ldots, c-1$ and $k = 1, \ldots, T$, where n_k^j are the sample sizes for each class j at time k, adding up to the total sample size n. The data are assumed to satisfy the following general observation model:

$$\mathbf{x}_{k,i}^j = \mathbf{h}_k^j(\mathbf{z}_k^j, \mathbf{v}_{k,i}^j), \tag{1.15}$$

for $i = 0, 1, \ldots, n_k^j$, $j = 0, 1, \ldots, c-1$ and $k = 1, \ldots, T$, where $\mathbf{h}_k^j$ is an arbitrary nonlinear function mapping the latent variables to the observable data and $\mathbf{v}_{k,i}^{j}$ defines an i.i.d. observation noise process, which is independent of the $\mathbf{z}_k^j$ process.

In the non-linear case problem, the ultimate goal is to develop a framework for nonstationary classification when the class-conditional distribution is represented by (1.13) and (1.15). Due to the nonlinearity of the state process and non-Gaussianity of the state process noise, in Chapter 5, we propose using sequential Monte-Carlo (SMC) for estimating the class-conditional distributions. The proposed framework yields several benefits:

- High classification accuracy, as all available data are employed for particle-based estimation of the class conditional distributions at various time points.
- The ability of handling missing data, by using the prediction step of the particle smoother.
- Robustness against unbalanced data, which can be compensated for by picking different number of particles for estimation of the class conditional densities.

1.2.2 Summary of Contributions

In [35], we proposed a novel classification algorithm for nonstationary data, called nonstationary LDA (NSLDA). This new classification rule is model-based, using a state-space equation for evolution of the distribution parameters “population drift”) in different scenarios. Furthermore, we address the case where parameters in linear state space models are unknown by proposing several different Kalman Smoothing frameworks in maximum-likelihood methods. Last, we proposed a general nonlinear, non-Gaussian model for nonstationary data, which allowed us to derive nonstationary discriminant analysis classification rules capable of producing classifiers tuned to the state of the distribution at each time point, while borrowing information from all time points. The high accuracy of the proposed NSLDA classification rule and its ability in handling missing or unbalanced data is demonstrated in a series of numerical experiments.

1.3 Organization

This dissertation contains two different projects I’ve been doing for my PhD research: on the bias of precision estimation under separate sampling and state space models to nonstationary discriminant analysis. This chapter covers background and summary of main contributions for each of the project.

In Chapter 2, we examine the bias of precision and recall estimators under separate sampling. After this, we proposed an unbiased precision estimator, which can be applied in case the true prevalence is known. We performed a set of experiments employing synthetic models and two real-data case studies.

In Chapter 3, we illustrated how linear state-space model approach to nonstationary data in different scenarios. We first review the linear drifts model, then reviewed Kalman Soother that applied multiple observations at one time. In the case when parameters in state space models are unknown, two maximum likelihood method Kalman Smoothing framework are developed - EM-based and GMM-based Kalman Smoother. Different versions of Non-stationary discriminant analysis are proposed. Last we showed synthetic simulations in different scenarios.

In Chapter 4, we extended the cases when the state-space model is non-linear and non-Gaussian. we propose using sequential Monte-Carlo (SMC) techniques [36,37] for efficient estimation of the class-conditional distributions via a finite set of particles, together with corresponding SMC-based Non-stationary discriminant analysis. Performance is assessed in a set of numerical experiments using simulated data.

Finally, Chapter 5 summarizes this dissertation.

2. ON THE BIAS OF PRECISION ESTIMATION UNDER SEPARATE SAMPLING[*]

2.1 Performance Metrics

We provide definitions and properties of the various error rates of interest in this study, including precision and recall. We consider the population error rates, which depend only on the probability distribution (also known as the "feature-label" distribution) governing the problem, as well as the estimated error rates, which attempt to approximate the corresponding population error rates by using sample data, and are thus the main object of our interest.

2.1.1 Population Performance Metrics

In an observational case-control study, there are two populations: Π_0 (control) and Π_1 (case). The *feature* vector $\mathbf{X} \in R^d$ summarizes numerical characteristics of a patient (e.g, blood concentrations of given proteins). The classification problem is how to assign accurately a new observation $\mathbf{X}$ to one of those two populations. Let the *label* $Y \in \{0, 1\}$ be defined as: $Y = 0$ if $\mathbf{X}$ is from the control population Π_0, and $Y = 1$ if $\mathbf{X}$ is from the case population Π_1. The statistical properties of the classification problem are entirely determined by the joint *feature-label* probability distribution $f(\mathbf{x}, y)$ between $\mathbf{X}$ and Y. The feature-label distribution can be decomposed as $f(\mathbf{x}, y) = f(\mathbf{x} \,|\, y) f(y)$, where $f(\mathbf{x} \,|\, y)$, $y = 0, 1$, give the distribution of $\mathbf{X}$ in each population of interest, whereas $f(y)$ is the marginal distribution of the binary random variable Y, specified by

$$P(Y = 0) = 1 - \mathrm{prev} \text{ and } P(Y = 1) = \mathrm{prev}\,, \tag{2.1}$$

where prev denotes the *prevalence*, i.e., is the probability that a randomly selected individual is a case subject. The name comes from the fact that it typically measures the prevalence of a disease in a population of interest. The prevalence plays a fundamental role in the sequel.

A *classifier* $\psi : R^d \to \{0, 1\}$ assigns $\mathbf{X}$ to the control or case population, according to whether

[*] Reprint with permission from "On the Bias of Precision Estimation Under Separate Sampling" by Shuilian Xie and Ulisses M Braga-Neto, 2019. Cancer Informatics, 18:1176935119860822. doi: 10.1177/1176935119860822.

$\psi(\mathbf{X}) = 0$ or $\psi(\mathbf{X}) = 1$, respectively. A classifier's sensitivity and specificity are defined as:

$$\text{sens} = P(\psi(\mathbf{X}) = 1 \mid Y = 1)\,, \tag{2.2}$$

$$\text{spec} = P(\psi(\mathbf{X}) = 0 \mid Y = 0)\,. \tag{2.3}$$

These are accuracy metrics that reflect how closely the classifier agrees with the true population of origin of the individual. The closer both are to 1, the more accurate the classifier is. A noteworthy property of the sensitivity and specificity is that they *do not depend on the prevalence*. In fact, one can write:

$$\text{sens} = \int_{\mathbf{x}:\psi(\mathbf{x})=1} f(\mathbf{x}|1)\, d\mathbf{x}\,, \tag{2.4}$$

so that sensitivity is a function only of $f(\mathbf{x} \mid y)$, not of $f(y)$. A similar expression holds for the specificity, showing that it too is not a function of the prevalence.

Other common performance metrics for a classifier are the *false-positive* (FP), *false-negative* (FN), *true-positive* (FP), and *true-negative* (FN) rates, given by

$$\text{FP} = P(\psi(\mathbf{X}) = 1, Y = 0)\,, \tag{2.5}$$

$$\text{FN} = P(\psi(\mathbf{X}) = 0, Y = 1)\,, \tag{2.6}$$

$$\text{TP} = P(\psi(\mathbf{X}) = 1, Y = 1)\,, \tag{2.7}$$

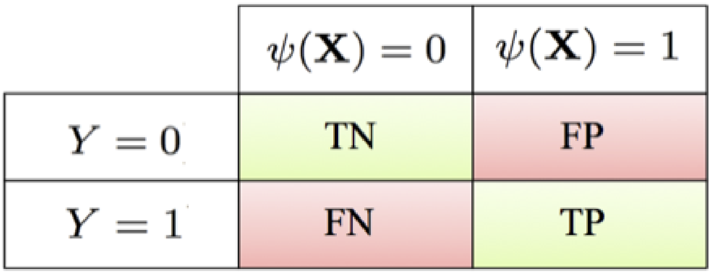
$$\text{TN} = P(\psi(\mathbf{X}) = 0, Y = 0)\,. \tag{2.8}$$

The (overall) classification error and accuracy rates are given by the sum of the appropriate error rates above:

$$\text{Err} = \text{FP} + \text{FN} = P(\psi(\mathbf{X}) \neq Y)\,, \tag{2.9}$$

$$\text{Acc} = \text{TP} + \text{TN} = P(\psi(\mathbf{X}) = Y) = 1 - \text{Err}\,. \tag{2.10}$$

See Fig. 2.1 for an illustration.



	$\psi(\mathbf{X}) = 0$	$\psi(\mathbf{X}) = 1$
$Y = 0$	TN	FP
$Y = 1$	FN	TP

Figure 2.1: Diagram of error (red) and accuracy (green) rates.

Unlike sensitivity and specificity, the previous performance metrics *do* depend on the prevalence. This can be seen easily via the relationships:

$$\begin{aligned}
\mathrm{FP} &= (1 - \mathrm{spec}) \times (1 - \mathrm{prev})\,, && (2.11)\\
\mathrm{FN} &= (1 - \mathrm{sens}) \times \mathrm{prev}\,, && (2.12)\\
\mathrm{TP} &= \mathrm{sens} \times \mathrm{prev}\,, && (2.13)\\
\mathrm{TN} &= \mathrm{spec} \times \mathrm{prev}\,. && (2.14)
\end{aligned}$$

Take, for example, TP and sensitivity; they are intimately (linearly) related, but TP is weighted by the prevalence. If the latter increases or decreases, TP increases and decreases accordingly, but the sensitivity stays the same. Other relationships can easily be obtained, for example:

$$\mathrm{prev} = \mathrm{FN} + \mathrm{TP}, \;\; 1 - \mathrm{prev} = \mathrm{FP} + \mathrm{TN}, \qquad (2.15)$$

$$\mathrm{sens} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}\,, \;\; \mathrm{spec} = \frac{\mathrm{TN}}{\mathrm{TN} + \mathrm{FP}}\,. \qquad (2.16)$$

Finally, we define the precision and recall accuracy metrics. Precision measures the likelihood that one has a true case given that the classifier outputs a case:

$$\mathrm{prec} = P(Y = 1 \mid \psi(\mathbf{X}) = 1)\,. \qquad (2.17)$$

Precision is thus similar to sensitivity; the latter is the likelihood that the classifiers outputs a case when applied on a true case. But the conditioning order is inverted. Applying Bayes' Theorem and using previously-derived relationships reveal that:

$$\mathrm{prec} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} = \frac{\mathrm{sens}\times\mathrm{prev}}{\mathrm{sens}\times\mathrm{prev}+(1-\mathrm{spec})\times(1-\mathrm{prev})}\,. \tag{2.18}$$

On the other hand, recall is simply the sensitivity:

$$\mathrm{rec} = \mathrm{sens} = \frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}}\,. \tag{2.19}$$

It follows that precision depends on the prevalence, but recall does not.

2.1.2 Estimated Performance Metrics

All the performance metrics defined in the previous section require the knowledge of the full feature-label distribution, or at least the class-conditional densities. In practice, however, these quantities are unknown, and thus sample data must be used to estimate the performance metrics. Let $S_n = \{(\mathbf{X}_1, Y_1), \dots, \mathbf{X}_n, Y_n)\}$ be a sample of size n from $f(\mathbf{x}, y)$, known as the *training data*. Consider the *empirical distribution* $p(\mathbf{x}, y)$, which is a discrete distribution putting mass $1/n$ over each pair $(\mathbf{X}_i, Y_i)$, and let $\widehat{P}$ denote the empirical probability measure under $p(\mathbf{x}, y)$. The basic estimation approach in statistical learning is to nominally replace the unknown distribution $f(\mathbf{x}, y)$ by $p(\mathbf{x}, y)$. This leads to the following simple estimator of prevalence:

$$\widehat{\mathrm{prev}} = \widehat{P}(Y=1) = \frac{1}{n}\sum_{i=1}^{n} I_{Y_i=1}\,, \tag{2.20}$$

where $I_A = 1$ if A is true and $I_A = 0$ if A is false. In the case of FP, FN, TP, and TN, and a given classifier ψ, one obtains

$$
\begin{aligned}
\widehat{\mathrm{FP}} &= \widehat{P}(\psi(\mathbf{X}) = 1, Y = 0) = \frac{1}{n}\sum_{i=1}^{n} I_{\{\psi(X_i)=1, Y_i=0\}} \,, & (2.21)\\
\widehat{\mathrm{FN}} &= \widehat{P}(\psi(\mathbf{X}) = 0, Y = 1) = \frac{1}{n}\sum_{i=1}^{n} I_{\{\psi(X_i)=0, Y_i=1\}} \,, & (2.22)\\
\widehat{\mathrm{TP}} &= \widehat{P}(\psi(\mathbf{X}) = 1, Y = 1) = \frac{1}{n}\sum_{i=1}^{n} I_{\{\psi(X_i)=1, Y_i=1\}} \,, & (2.23)\\
\widehat{\mathrm{TN}} &= \widehat{P}(\psi(\mathbf{X}) = 0, Y = 0) = \frac{1}{n}\sum_{i=1}^{n} I_{\{\psi(X_i)=0, Y_i=0\}} \,. & (2.24)
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\widehat{\mathrm{Err}} &= \widehat{\mathrm{FP}} + \widehat{\mathrm{FN}} = \frac{1}{n}\sum_{i=1}^{n} I_{\psi(X_i)\neq Y_i} \,, & (2.25)\\
\widehat{\mathrm{Acc}} &= \widehat{\mathrm{TP}} + \widehat{\mathrm{TN}} = \frac{1}{n}\sum_{i=1}^{n} I_{\psi(X_i)=Y_i} = 1 - \widehat{\mathrm{Err}} \,. & (2.26)
\end{aligned}
$$

These are basic counting estimators; e.g., the FP estimator counts the number of false positive over the training data (and divides that by n, so the result is between 0 and 1). $\widehat{\mathrm{Err}}$ is also known as the *resubstitution* estimator in the pattern recognition literature [15].

We define the remaining performance metrics estimators analogously, using (2.16), (2.18), and (2.19):

$$
\begin{aligned}
\widehat{\mathrm{spec}} &= \frac{\widehat{\mathrm{TN}}}{\widehat{\mathrm{TN}} + \widehat{\mathrm{FP}}} = \frac{\sum_{i=1}^{n} I_{\{\psi(X_i)=0, Y_i=0\}}}{\sum_{i=1}^{n} I_{Y_i=0}} \,,\\
\widehat{\mathrm{prec}} &= \frac{\widehat{\mathrm{TP}}}{\widehat{\mathrm{TP}} + \widehat{\mathrm{FP}}} = \frac{\sum_{i=1}^{n} I_{\{\psi(X_i)=1, Y_i=1\}}}{\sum_{i=1}^{n} I_{\psi(X_i)=1}} \,, \qquad (2.27)\\
\widehat{\mathrm{rec}} = \widehat{\mathrm{sens}} &= \frac{\widehat{\mathrm{TP}}}{\widehat{\mathrm{TP}} + \widehat{\mathrm{FN}}} = \frac{\sum_{i=1}^{n} I_{\{\psi(X_i)=1, Y_i=1\}}}{\sum_{i=1}^{n} I_{Y_i=1}} \,.
\end{aligned}
$$

2.2 Mixture and Separate Sampling

The usual scenario in Statistical Learning is to assume that $S_n = \{(X_1, Y_1), \ldots, (X_n, Y_n)\}$ is an independent and identically distributed (i.i.d.) sample from the true feature-label distribution $f(\mathbf{x}, y)$; i.e., the set of all sample points is independent and each sample point has distribution

$f(\mathbf{x}, y)$, so that

$$f(S_n) \;=\; \Pi_{i=1}^n f(\mathbf{X}_i, Y_i) \;=\; \Pi_{i=1}^n f(\mathbf{X}_i \mid Y_i) \times \Pi_{i=1}^n f(Y_i)\,, \tag{2.28}$$

where all the densities on the right-hand side are as defined previously. That makes S_n a sample from the *mixture* of populations, where each label Y_i is distributed as:

$$P(Y_i = 0) \;=\; 1 - \mathrm{prev} \;\text{ and }\; P(Y_i = 1) = \mathrm{prev}\,, \tag{2.29}$$

for $i = 1, \ldots, n$. Under mixture sampling, $N_0 = \sum_{i=1}^n I_{Y_i=0}$ and $N_1 = \sum_{i=1}^n I_{Y_i=1} = n - N_0$ are binomial random variables, with parameters $(n, 1 - \mathrm{prev})$ and (n, prev), respectively.

By contrast, observational case-control studies in biomedicine typically proceed by collecting data from the populations separately, where the separate sample sizes n_0 and n_1, with $n_0 + n_1 = n$, are pre-determined and nonrandom; i.e., sample occurs with the restriction $N_1 = \sum_{i=1}^n I_{Y_i=1} = n_1$ (or, equivalently, $N_0 = \sum_{i=1}^n I_{Y_i=0} = n_0$). The restriction means that the labels $Y_1, \ldots, Y_n$ are no longer independent, even though the feature vectors $\mathbf{X}_1, \ldots, \mathbf{X}_n$ are independent given the labels. Furthermore, the conditional distributions $f(\mathbf{X}_i \mid Y_i)$ are the same as before. The distribution of the sample is given by

$$f(S_n \mid N_0 = n_0) \;=\; \Pi_{i=1}^n f(\mathbf{X}_i \mid Y_i) \times f(Y_1, \ldots, Y_n \mid N_0 = n_0)\,, \tag{2.30}$$

Under separate sampling, only the order of the labels $Y_1, \ldots, Y_n$ may be random. Thus, $f(Y_1, \ldots, Y_n \mid N_0 = n_0)$ is a discrete uniform distribution over all $\binom{n}{n_0}$ possible orderings. This can also be obtained by direct computation, as follows:

$$\begin{aligned} f(Y_1, \ldots, Y_n \mid N_0 = n_0) \;&=\; \frac{f(Y_1, \ldots, Y_n, N_0 = n_0)}{P(N_0 = n_0)} \\ &= \begin{cases} \frac{\mathrm{prev}^{n_1}(1-\mathrm{prev})^{n_0}}{\binom{n}{n_0}\mathrm{prev}^{n_1}(1-\mathrm{prev})^{n_0}} \;=\; \frac{1}{\binom{n}{n_0}}\,, & \text{if } \sum_{i=1}^n I_{Y_i=0} = n_0, \\ 0, & \text{otherwise.} \end{cases} \end{aligned} \tag{2.31}$$

It is not difficult to verify that under (2.30) and (2.31), the marginal distribution of each label Y_i is given by

$$\begin{aligned} P(Y_i = 1 \mid N_0 = n_0) &= \frac{n_1}{n} \triangleq r\,, \\ P(Y_i = 0 \mid N_0 = n_0) &= \frac{n_0}{n} = 1 - r\,, \end{aligned} \tag{2.32}$$

for $i = 1, \dots, n$, where r is the (fixed) sample size ratio under separate sampling. Comparing (2.29) and (2.32) reveals the main difference between mixture and separate sampling.

2.3 Bias of the Precision Estimator

In this subsection, we present a theoretical large-sample analysis of the bias of the estimators discussed previously, focusing on the precision estimator. Estimation bias is defined as the expectation over the sample data S_n of the difference between the estimated and true quantities.

The situation is clear with the estimator of the prevalence itself, given by (2.20). Under mixture sampling, we have

$$E[\widehat{\mathrm{prev}}] = \frac{1}{n}\sum_{i=1}^{n} E[I_{Y_i=1}] = P(Y_1 = 1) = \mathrm{prev}\,, \tag{2.33}$$

so the estimator is unbiased (in addition, as n increases, $\mathrm{Var}(\widehat{\mathrm{prev}}) \to 0$ and $\widehat{\mathrm{prev}} \to \mathrm{prev}$ in probability, by the law of large numbers). However, under separate sampling,

$$\begin{aligned} E[\widehat{\mathrm{prev}} \mid N_0 = n_0] &= \frac{1}{n}\sum_{i=1}^{n} E[I_{Y_i=1} \mid N_0 = n_0] \\ &= P(Y_1 = 1 \mid N_0 = n_0) = r\,, \end{aligned} \tag{2.34}$$

according to (2.32). This also follows directly from the fact that $\widehat{\mathrm{prev}}$ becomes a constant estimator, $\widehat{\mathrm{prev}} \equiv r$, according to (2.20). Thus,

$$\begin{aligned} \mathrm{Bias}_{\mathrm{sep}}(\widehat{\mathrm{prev}}) &= E[\widehat{\mathrm{prev}} - \mathrm{prev} \mid N_0 = n_0] \\ &= r - \mathrm{prev}\,. \end{aligned} \tag{2.35}$$

Assuming that the sample size ratio $r = n_1/n$ is held constant as n increases (e.g., under the common balanced design case, $n_0 = n_1 = n/2$), then this bias cannot be reduced with increased sample size. Furthermore, the bias is larger the further away prev is from r. In particular, the bias tends to be large when prev is small and $r = 1/2$, which is a common scenario in practice.

The situation for $\widehat{\mathrm{FP}}$, $\widehat{\mathrm{FN}}$, $\widehat{\mathrm{FP}}$, and $\widehat{\mathrm{TN}}$ is more complicated. First, we are interested in a classifier ψ_n derived by a classification rule from the sample data $S_n = \{(\mathbf{X}_1, Y_1), \ldots, \mathbf{X}_n, Y_n)\}$. Therefore, all expectations and probabilities in the previous sections are conditional on S_n. Under mixture sampling, the powerful *Vapnik-Chervonenkis Theorem* can be applied to show that all of these estimators are asymptotically unbiased, provided that classification rule has a finite *VC Dimension* [20]. This includes many useful classification algorithms such as LDA, linear SVMs, perceptrons, polynomial-kernel classifiers, certain decision trees and neural networks, but it excludes nearest-neighbor classifiers, for example. Classification rules with finite VC dimension do not cut the feature space in complex ways and are thus generally robust against overfitting.

Assuming mixture sampling and a classification algorithm with finite VC dimension $V_{\mathcal{C}}$, it can be shown that (details omitted; see [15, p. 47] for a similar argument)

$$\mathrm{Bias}_{\mathrm{mix}}(\widehat{\mathrm{FP}}) \leq 8\sqrt{\frac{V_{\mathcal{C}} \log(n+1) + 4}{2n}}\,, \tag{2.36}$$

so that the bias vanishes as $n \to \infty$. Similar inequalities apply to $\widehat{\mathrm{FN}}$, $\widehat{\mathrm{FP}}$, and $\widehat{\mathrm{TN}}$. These are distribution-free results, hence vanishingly small bias is guaranteed if $n \gg V_{\mathcal{C}}$, regardless of the feature-label distribution. For linear classification rules, $V_{\mathcal{C}} = d + 1$, where d is the dimensionality of the feature vector. In this case, the $\widehat{\mathrm{FP}}$, $\widehat{\mathrm{FN}}$, $\widehat{\mathrm{FP}}$, and $\widehat{\mathrm{TN}}$ estimators are essentially unbiased if $n \gg d$.

Next we consider the bias of the precision and recall estimators under mixture sampling (the analysis for the sensitivity and specificity estimators is similar; in fact, the former is just the recall estimator). We will make use of the following approximation for the expectation of a ratio of two random variables W and Z (see the Appendix A for the derivation of this approximation and the conditions under which it is valid):

$$E\left[\frac{W}{Z}\right] \approx \frac{E[W]}{E[Z]}\,. \tag{2.37}$$

The approximation is quite accurate if W and Z are around $E[W]$ and $E[Z]$, respectively (it is asymptotically exact as $W \to E[W]$ and $Z \to E[Z]$). For the precision estimator,

$$\begin{aligned} E[\widehat{\text{prec}}] &= E\left[\frac{\widehat{\text{TP}}}{\widehat{\text{TP}}+\widehat{\text{FP}}}\right] \approx \frac{E[\widehat{\text{TP}}]}{E[\widehat{\text{TP}}+\widehat{\text{FP}}]} \\ &\approx \frac{E[\text{TP}]}{E[\text{TP}+\text{FP}]} \approx E\left[\frac{\text{TP}}{\text{TP}+\text{FP}}\right] = E[\text{prec}], \end{aligned} \tag{2.38}$$

for a sufficiently large sample, where we used the previously-established asymptotic unbiasedness of $\widehat{\text{TP}}$, $\widehat{\text{TP}}$, and $\widehat{\text{FN}}$. An entirely similar derivation shows that $E[\widehat{\text{rec}}] = E[\text{rec}]$. Hence, for "well-behaved" classification algorithms (those with finite VC dimension), both the precision and recall estimators are asymptotically unbiased under mixture sampling.

We are not aware of the existence of a VC theory for separate sampling at this time. In order to obtain approximate results for the separate sampling case, we will assume instead that, at large enough sample sizes, the classifier ψ is nearly constant, and invariant to the sample. This assumption is not unrelated to the finite VC dimension assumption made in the case of mixture sampling. Many of the same classification algorithms that have finite VC dimension, such as LDA and linear SVMs, will also become nearly constant as sample size increases. In this case, we have

$$\begin{aligned} E[\widehat{\text{TP}} \mid N_0 = n_0] &= \frac{1}{n}\sum_{i=1}^{n} E[I_{\{\psi(X_i)=1, Y_i=1\}} \mid N_0 = n_0] \\ &= P(\psi(\mathbf{X}_1) = 1, Y_1 = 1 \mid N_0 = n_0) \\ &= P(\psi(\mathbf{X}_1) = 1 \mid Y_1 = 1) P(Y_1 = 1 \mid N_0 = n_0) \\ &= \text{sens} \times r\,, \end{aligned} \tag{2.39}$$

where we used the fact that the event $\{\psi(\mathbf{X}_1) = 1\}$ is independent of N_0 given Y_1 and (2.32). Notice that the equality $P(\psi(\mathbf{X}_1) = 1 \mid Y_1 = 1) = \text{sens}$ depends on the fact that ψ is assumed to be constant, so that $(\mathbf{X}_1, Y_1)$ behaves as an independent test point (also because of a constant ψ, there is no expectation around sens). Hence, $\widehat{\mathrm{TP}}$ is biased under separate sampling, with

$$\mathrm{Bias}_{\mathrm{sep}}(\widehat{\mathrm{TP}}) = \mathrm{sens}\times r - \mathrm{TP} = \mathrm{sens}\times(r - \mathrm{prev})\,. \tag{2.40}$$

As in the case with the bias of $\widehat{\mathrm{prev}}$ under separate sampling, the bias of $\widehat{\mathrm{TP}}$ cannot be reduced with increasing sample size. The bias is in fact larger the more sensitive the classifier is. One can derive similar results for $\widehat{\mathrm{FP}}$, $\widehat{\mathrm{FN}}$, and $\widehat{\mathrm{TN}}$.

The recall estimator is approximately unbiased under separate sampling:

$$\begin{aligned} E[\widehat{\mathrm{rec}} \mid N_0 = n_0] &= E\left[\frac{\widehat{\mathrm{TN}}}{\widehat{\mathrm{TN}} + \widehat{\mathrm{FP}}}\middle| N_0 = n_0\right] \\ &= E\left[\frac{\widehat{\mathrm{TP}}}{\widehat{\mathrm{prev}}}\middle| N_0 = n_0\right] = \frac{E[\widehat{\mathrm{TP}} \mid N_0 = n_0]}{r} \\ &= \frac{\mathrm{sens}\times r}{r} = \mathrm{sens} = \mathrm{rec}\,. \end{aligned} \tag{2.41}$$

This is a consequence of recall's not being a function of the prevalence. However, for the precision estimator,

$$\begin{aligned} E[\widehat{\mathrm{prec}} \mid N_0 = n_0] &= E\left[\frac{\widehat{\mathrm{TP}}}{\widehat{\mathrm{TP}} + \widehat{\mathrm{FP}}}\middle| N_0 = n_0\right] \\ &\approx \frac{E[\widehat{\mathrm{TP}} \mid N_0 = n_0]}{E[\widehat{\mathrm{TP}} + \widehat{\mathrm{FP}} \mid N_0 = n_0]} \\ &= \frac{\mathrm{sens}\times r}{\mathrm{sens}\times r + (1-\mathrm{spec})\times(1-r)} \\ &\neq \frac{\mathrm{sens}\times \mathrm{prev}}{\mathrm{sens}\times\mathrm{prev} + (1{-}\mathrm{spec})\times(1{-}\mathrm{prev})} = \mathrm{prec}\,. \end{aligned} \tag{2.42}$$

The precision estimator is thus biased under separate sampling unless the true prevalence matches exactly the sample ratio $r = n_1/n$; the bias is larger the further away prev is from r.

In case the true prevalence is known, e.g., from public health records and government databases, then we show below that the following estimator of the precision,

$$\widehat{\text{prec}}^{\text{prev}} = \frac{\widehat{\text{sens}} \times \text{prev}}{\widehat{\text{sens}} \times \text{prev} + (1 - \widehat{\text{spec}}) \times (1 - \text{prev})}\,, \tag{2.43}$$

which is based on (2.18), is an asymptotically unbiased estimator of the precision under either mixture or separate sampling. Asymptotic unbiasedness in the mixture sampling case can be shown by repeating the steps in the analysis of the ordinary precision estimator. Under separate sampling, we have

$$\begin{aligned} & E[\widehat{\text{prec}}^{\text{prev}} \mid N_0 = n_0] \\ & \approx \frac{E[\widehat{\text{sens}} \mid N_0] \times \text{prev}}{E[\widehat{\text{sens}} \mid N_0] \times \text{prev} + (1 - E[\widehat{\text{spec}} \mid N_0]) \times (1 - \text{prev})} \\ & = \frac{\text{sens} \times \text{prev}}{\text{sens} \times \text{prev} + (1 - \text{spec}) \times (1 - \text{prev})} = \text{prec}\,, \end{aligned} \tag{2.44}$$

since $E[\widehat{\text{sens}} \mid N_0 = n_0] = \text{sens}$ and $E[\widehat{\text{spec}} \mid N_0 = n_0] = \text{spec}$, as can be easily shown. Hence, $\widehat{\text{prec}}^{\text{prev}}$ is an asymptotically unbiased estimator of the precision under either mixture or separate sampling. The ordinary precision estimator $\widehat{\text{prec}}$ should not be used under separate sampling, or large and irreducible bias may occur. On the other hand, in the impossibility of obtaining information on the true prevalence value, then no meaningful estimator of the precision is possible.

2.4 Simulation Results and Discussion

In this section, we employ synthetic and real-world data to investigate the accuracy of the analysis in the previous section and the performance of the precision estimator under separate sampling. We present results for the bias of the usual and proposed precision estimators under separate sampling, using different classification rules. We also showed corresponding results for mixture sampling and the recall estimator.

2.4.1 Experiments with Synthetic Data

We performed a set of experiments employing synthetic data from a homoskedastic Gaussian model, consisting of 3-dimensional class-conditional distributions $N(\boldsymbol{\mu}_i, \Sigma)$, for $i = 0, 1$, with $\boldsymbol{\mu}_0 = (0, 0, 0)$, $\boldsymbol{\mu}_1 = (0, 0, \theta)$, where $\theta > 0$ is a parameter governing the separation between the classes, and $\Sigma = \mathrm{diag}(\sigma_1^2, \sigma_2^2, \sigma_3^2)$ (i.e., a matrix with $\sigma_1^2, \sigma_2^2, \sigma_3^2$ on the diagonal and zeros off diagonal). We consider two sample sizes, $n = 30$ and $n = 200$, so that we can compare the results for small and large sample sizes. All experiments with separate sampling are performed with sample size ratio $r = \frac{n_1}{n} \in [0.1, 0.9]$. The synthetic data parameters are summarized in Table 2.1.

For each value of r and prev, we repeat the following process 1,000 times, and average the results to estimate expected values:

1. Generate sample data S_n of size n according to r (separate sampling) or prev (mixture sampling);

2. Train a classifier using one of three classification rules [38]: Linear Discriminant Analysis (LDA), 3-Nearest Neighbors (3NN), and a nonlinear Radial-Basis Function Support Vector Machine (RBF-SVM).

3. Obtain recall and precision estimates. Compute both the usual precision estimate $\widehat{\mathrm{prec}}$ and the modified precision estimate $\widehat{\mathrm{prec}}^{\mathrm{prev}}$.

4. Obtain accurate estimates of the true precision values by using a test set of size 10,000.

Parameter	Value
Dimensionality/ feature size	$D = 3$
Mean difference	$\theta = 2$
Covariance matrix	$\sigma_1^2 = 0.5, \sigma_2^2 = 0.5, \sigma_3^2 = 1$
Sample size	$n = 30, 200$
Sample size ratio r	$r = 0.1, 0.3, 0.5, 0.7, 0.9$
True prevalence	$\mathrm{prev} = 0.1, 0.3, 0.5, 0.7, 0.9$

Table 2.1: Synthetic data parameters.

Fig. 2.2 displays the results of the experiment. Notice that there is only one curve for the traditional precision estimator $\widehat{\mathrm{prec}}$ because it does not employ the actual value of prev. The values of $\widehat{\mathrm{prec}}$ and $\widehat{\mathrm{prec}}^{\mathrm{prev}}$ coincide when $\mathrm{prev} = r$, as expected. However, as the values of prev and r become different, their values become quite different, and $\widehat{\mathrm{prec}}^{\mathrm{prev}}$ displays much less bias, i.e., it tracks the true precision much more closely, than $\widehat{\mathrm{prec}}$. At the small sample size $n = 30$, both estimators display bias, which is however much larger overall for $\widehat{\mathrm{prec}}$ than for $\widehat{\mathrm{prec}}^{\mathrm{prev}}$. At the large sample size $n = 200$, the bias of $\widehat{\mathrm{prec}}^{\mathrm{prev}}$ nearly disappears for LDA and is reduced for the other classification rules. We note that, among these classification rules, LDA is the only one with a finite VC dimension; the fact that the bias in this case shrinks to zero as sample size increases confirms the results of the theoretical analysis in the previous section (convergence is quite fast, and quite evident at $n = 200$, due to the fact that the synthetic data is homoskedastic Gaussian). Notice also that the bias of $\widehat{\mathrm{prec}}$ cannot be reduced by increasing sample size, which is also in agreement with the theoretical analysis.

To examine more closely the effect of the difference between prev and r on precision estimation, Fig. 2.3 plots bias estimates for $\widehat{\mathrm{prec}}$ and $\widehat{\mathrm{prec}}^{\mathrm{prev}}$ as a function of the absolute difference between prev and r, using the same data employed in Figure 2.2. It can be seen that the bias is always positive, indicating optimistic precision estimates. In nearly all cases, $\widehat{\mathrm{prec}}^{\mathrm{prev}}$ has a smaller bias than $\widehat{\mathrm{prec}}$, and when prev is far from r, the difference in bias becomes quite large.

2.4.2 Case Studies with Real Data

Here we further investigate the bias of precision estimation under separate sampling using real data from three published studies:

- **Leukemia Study.** This publication [39] used a tumor microarray dataset containing two types of human acute leukemia: acute myeloid leukemia (AML) and acute lymphoblastic leukemia (ALL). Gene expression measurements were taken from $15,154$ genes from 72 tissue specimens, 47 of which of ALL type (class 0) and 25 of AML type (class 1), so that $r = 0.347$. The estimator $\widehat{\mathrm{prec}}^{\mathrm{prev}}$ was computed using the value $\mathrm{prev} = 0.222$, which is the incidence rate of ALL over AML in the U.S. population [40].

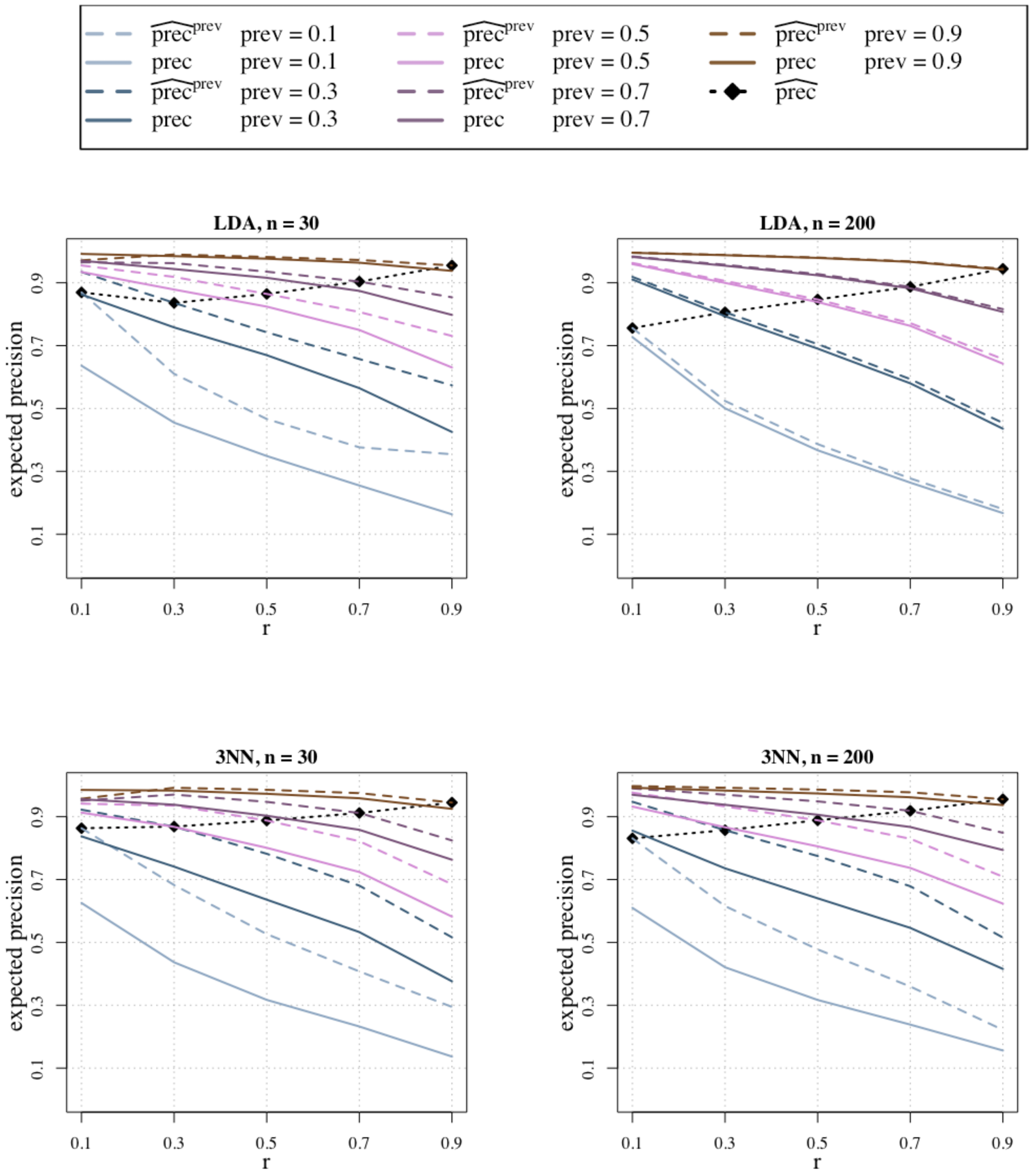


Figure 2.2: Average true precision (solid curves), average usual precision estimate $\widehat{\text{prec}}$ (dash-diamond curves), and average modified precision estimate $\widehat{\text{prec}}^{\text{prev}}$ (dashed curves), for LDA, 3NN and RBF-SVM, with sample sizes $n = 30$ and $n = 200$, and different prevalence values, as a function of the sample size ratio.

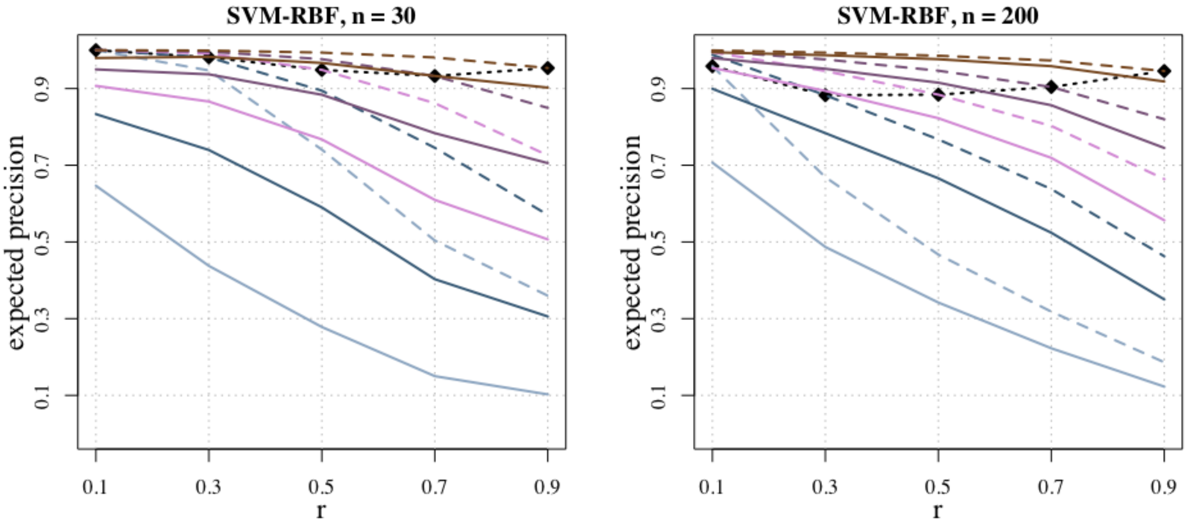


Figure 2.2 Continued

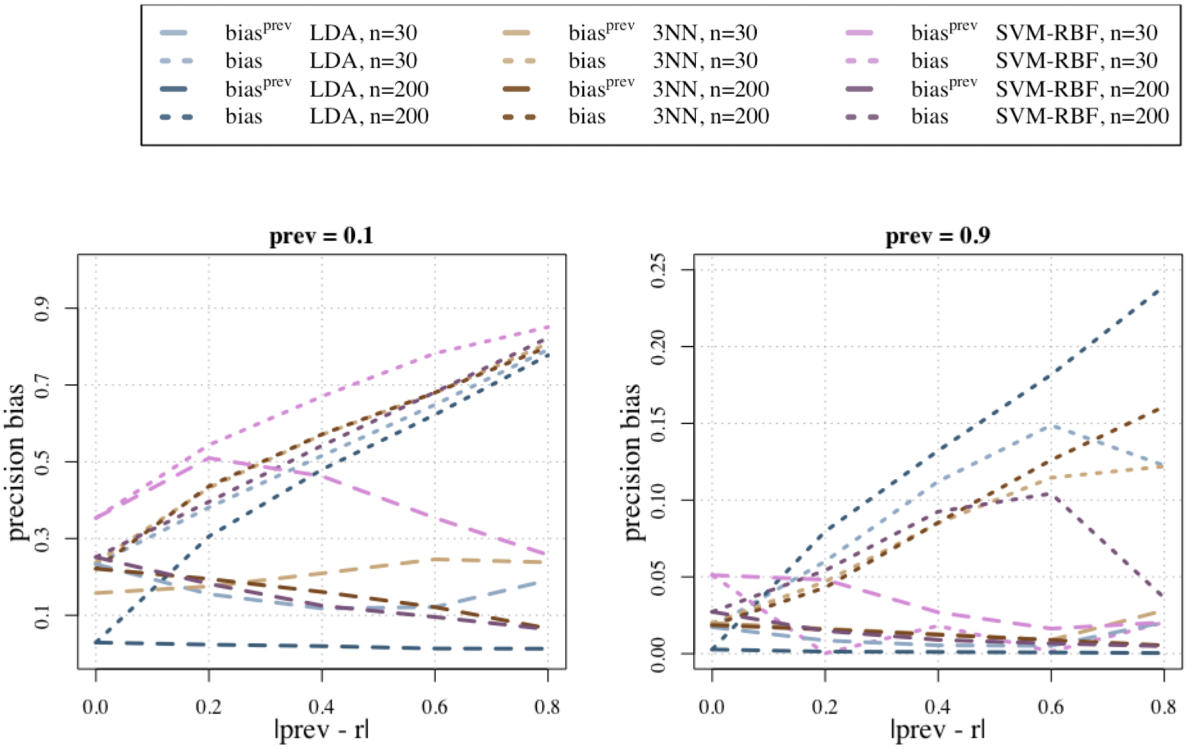


Figure 2.3: Estimated bias of the usual precision estimator $\widehat{\text{prec}}$ (dotted curves), and the modified precision estimator $\widehat{\text{prec}}^{\text{prev}}$ (dashed curves) for LDA, 3NN and RBF-SVM, with sample sizes $n = 30$ and $n = 200$, and different prevalence values, as a function of the absolute difference between true prevalence and sample size ratio.

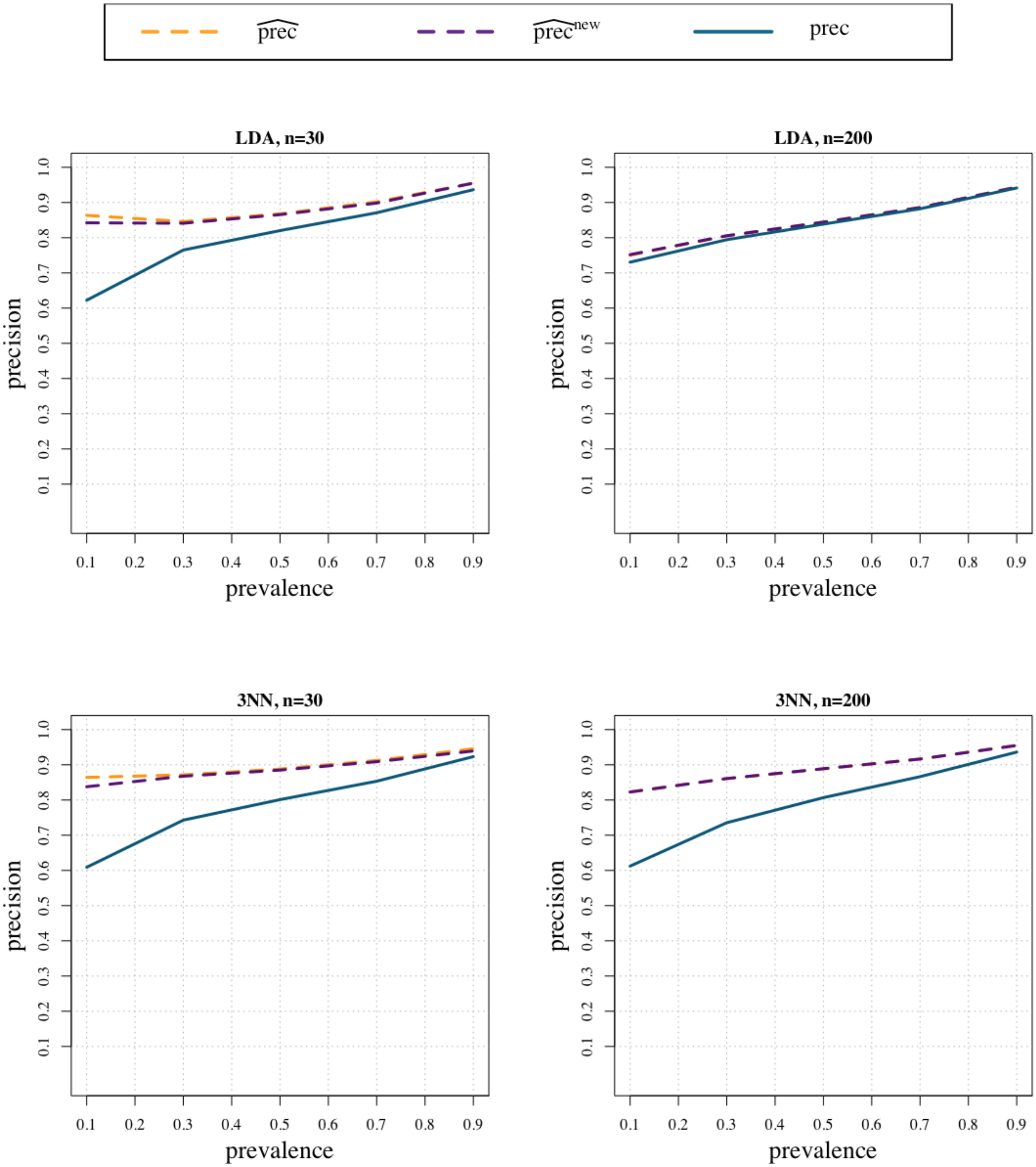


Figure 2.4: Precision under mixture sampling, as a function of prev for different classification rules and different sample size using synthetic data. Average true precision values (solid blue curve) and average precision estimates $\widehat{\mathrm{prec}}$ (dashed orange curve) and $\widehat{\mathrm{prec}}^{\mathrm{new}}$ (dashed purple curve)

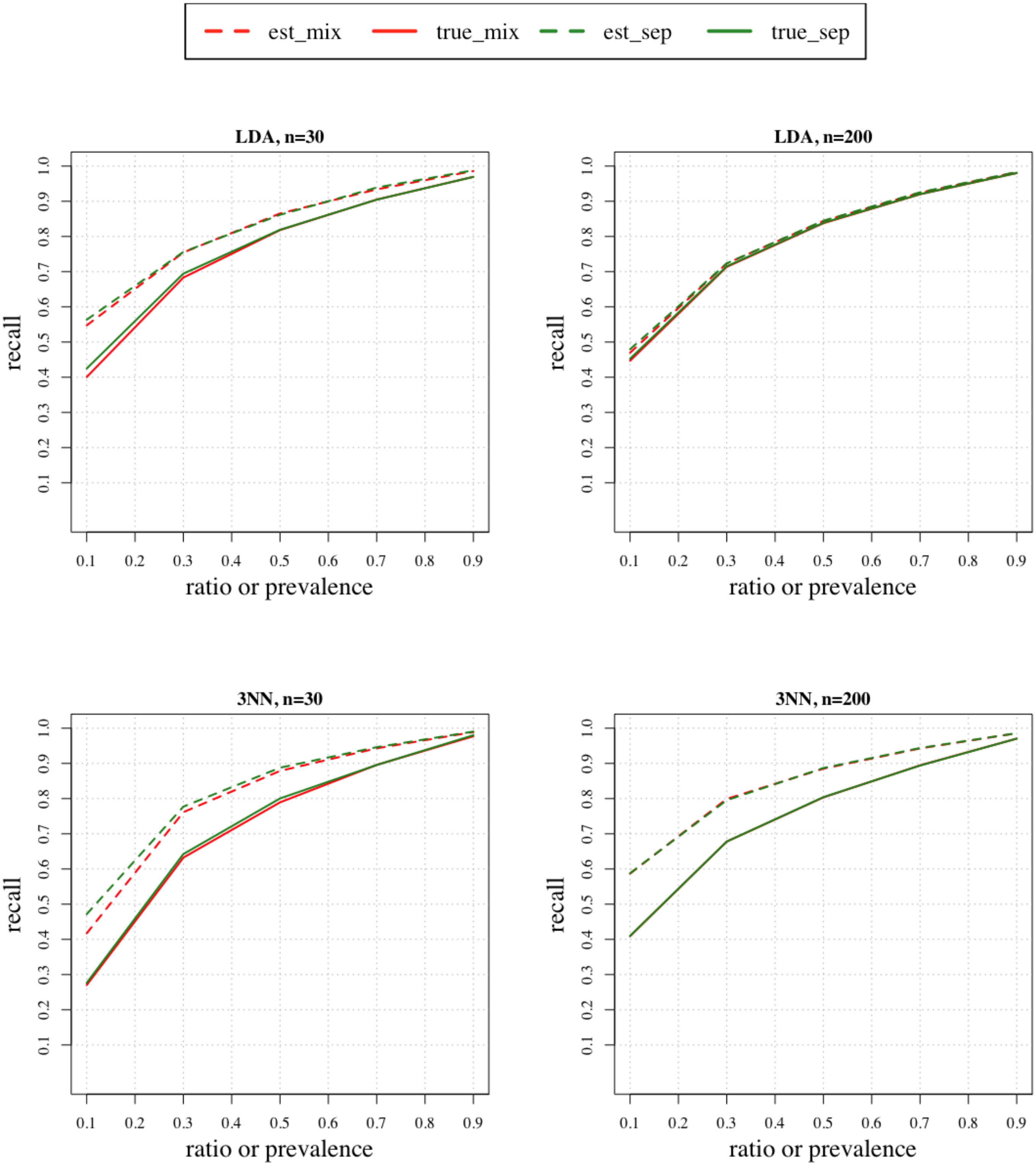


Figure 2.5: Recall under mixture and separate sampling, as a function of r or prev for different classification rules and different sample size using synthetic data. Average true recall values (solid curves) and average recall estimates (dashed curves). Mixture sampling (red) and separate sampling (green).

- **Breast Cancer Study.** The second publication [41] employed the Wisconsin Breast Cancer (Original) Dataset from the University of California-Irvine (UCI) Machine Learning Repository [42, 43], which has been used by several groups to investigate breast cancer classification methods [44, 45]. The dataset consists of 699 instances, 458 and 241 of which are from benign and malignant tumors, respectively, and 10 features corresponding to cytological characteristics of breast fine-needle aspirates. According to [46], fewer than 20% of breast lumps are malignant, therefore we used used $\mathrm{prev} = 0.2$ in the computation of the modified precision estimator $\widehat{\mathrm{prec}}^{\mathrm{prev}}$.

- **Liver Disease Study.** The final publication [47] employed a liver disease dataset, also from the UCI Machine Learning Repository. This data set contains 5 blood test attributes and 345 records, of which 145 belong to individuals with liver disease (class 0) and 200 measurements are taken from healthy individuals (class 1), so that $r = 0.42$. This dataset was donated to UCI in 1990, when the prevalence rate for chronic liver disease in the US was $\mathrm{prev} = 0.1178$ [48], which we use as the prevalence in the computation of the $\widehat{\mathrm{prec}}^{\mathrm{prev}}$ estimator.

All three studies used libraries from the Weka machine learning environment [49] to compute usual precision estimates on separately-sampled data, while ignoring true prevalences, for different classification rules: Naive Bayes (NB) [50], C4.5 decision tree [51], Back-Propagated Neural Networks, 3NN and Linear SVM [38]. We reproduced the analysis in all three papers using Weka, obtaining almost exactly the same $\widehat{\mathrm{prec}}$ estimates reported in those papers, and added for comparison the $\widehat{\mathrm{prec}}^{\mathrm{prev}}$ using the prevalence values described above. The results, displayed in Fig. 2.6, show that, without exception, the usual precision estimates $\widehat{\mathrm{prec}}$ are larger than the more accurate $\widehat{\mathrm{prec}}^{\mathrm{prev}}$ estimates, in agreement with the previously observed fact that $\widehat{\mathrm{prec}}$ displays a larger (optimistic) bias. The bias is particularly large in the case of the liver disease study, reflecting the fact that among the three data sets, this is the one where the value of prev and r differ the most.

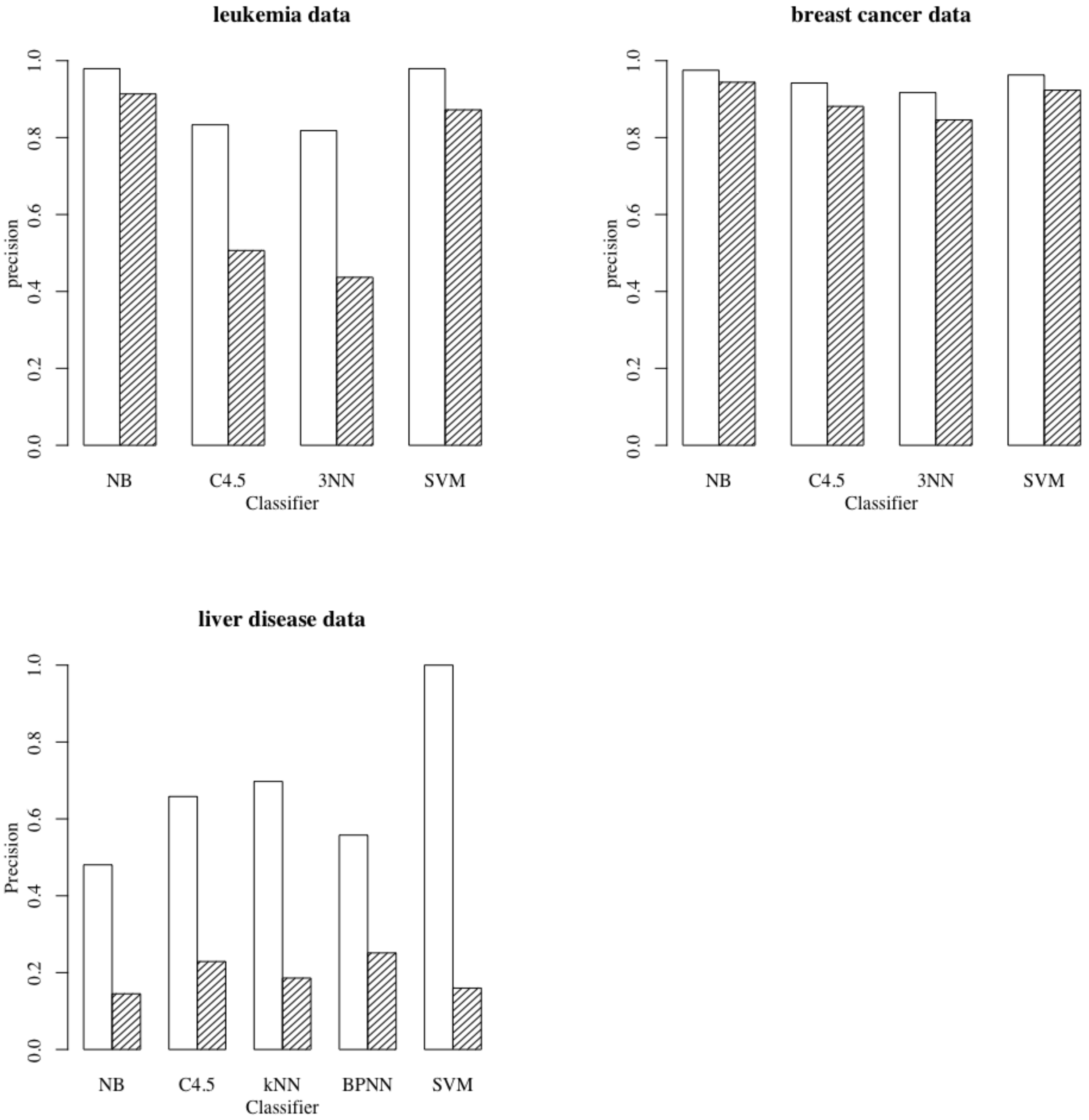


Figure 2.6: Precision estimates for different classification rules using separately-sampled leukemia, breast cancer, and liver disease data. The white bars depict the usual estimated precision estimates, while the shaded bars are for the precision estimates using the true case prevalences.

3. LINEAR STATE-SPACE-MODELS TO NONSTATIONARY DISCRIMINANT ANALYSIS[*]

3.1 Linear Drift Model

Let us consider a classification problem, in which $\mathbf{j} = \{0, 1, ..., c - 1\}$ denote the set of all classes. Assume $\Pi_k^j = f_k(\mathbf{x}|j)$ is the class conditional distribution of class j at time step k. In non-stationary condition, the class conditional distributions are function of time. Using the Gaussian assumption for class conditional distributions, we have: $f_k(\mathbf{x}|j) \sim \mathcal{N}\left(\boldsymbol{\mu}_k^j, P_k^j\right)$. Considering the linear model for evolution of the class conditional distributions, we have:

$$\boldsymbol{\mu}_k^j = A^j \boldsymbol{\mu}_{k-1}^j + \mathbf{w}_k^j, \tag{3.1}$$

for $k = 1, 2, ...$; where $\mathbf{w}_k^j$is independent zero-mean Gaussian noise processes with known covariances matrices of Q^j, and A^j is state transition matrices for class j.

In most real-world problems, only noisy data from the class conditional distributions of the system is available for decision making task. This can be due to various sources of errors that might affect process of acquiring data. In this project, the data are assumed to be generated through the following measurement process:

$$\mathbf{x}_k^j = C^j \boldsymbol{\mu}_k^j + \mathbf{v}_k^j, \tag{3.2}$$

where $\mathbf{v}_k^j \sim \mathcal{N}\left(\mathbf{0}, R^j\right)$, and C^j denote the dynamics of measurement process for class j.

In this project, the ultimate goal is to developed the Linear Discriminant Analysis (LDA) for nonstationary condition given only c sets of measurements, where $S_{N_k^j}^j = \{\mathbf{X}_1^j, ..., \mathbf{X}_{N_k^j}^j\}$ denotes N_k^j measurements at time step k from equation 3.1 and 3.2 . Toward this, one needs to accurately

[*] Parts of this section are reprinted with permission from "Nonstationary linear discriminant analysis" by S. Xie, M. Imani, E. R. Dougherty and U. M. Braga-Neto, 2017 51st Asilomar Conference on Signals, Systems, and Computers, 2017, pp. 161-165, doi: 10.1109/ACSSC.2017.8335158.

estimate the class conditional distributions. A conventional approach for estimating the mean vectors and covariance matrices of the class conditional distributions is to use the classical Kalman smoother (KS) [31]. Two cases are considered in this project:

- *Systems with Fully-Known Dynamics:* In this case, all matrices in equations (3.1) and (3.2) are assumed to be known, and the Kalman Smoother algorithm can be directly applied to obtain estimates of the class means $\{\boldsymbol{\mu}_k^j; k = 1, ..., T\}$ required for the proposed NSLDA classifier. A modified version of Kalman Smoother with multiple measurements at one time is discussed in Section 3.2.

- *Systems with Partially-Known Dynamics:* In this case, parameters in equations (3.1) and (3.2) are only partially unknown. In this case, unknown parameters must be learned from data, simultaneously with the class mean themselves. Two versions of modified Kalman Smoother have been proposed in Section 3.3 for this purpose, where parameters such as initial states $\boldsymbol{\mu}_1^j$, matrices A^j, Q^j, and the corresponding time labels of available measurements might be unknown.

3.2 Kalman Smoother

For class label $j \in \{0, 1, ..., c-1\}$, the state model and measurement processes are

$$\begin{aligned} \boldsymbol{\mu}_k^j &= A^j \boldsymbol{\mu}_{k-1}^j + B^j \mathbf{w}_k^j \,, \\ \mathbf{x}_k^j &= C^j \boldsymbol{\mu}_k^j + D^j \mathbf{v}_k^j \,, \end{aligned} \tag{3.3}$$

where $\mathbf{w}_k^j$ and $\mathbf{v}_k^j$ are i.i.d. Gaussian white noise at time k for class label j, with $\mathbf{w}_k^j \sim \mathcal{N}(0, Q^j)$, and $\mathbf{v}_k^j \sim \mathcal{N}(0, R^j)$. For the i.i.d. sample S_k at time k, let the j -label points be written in vector form as $\mathbf{x}_k^j = (\mathbf{x}_{k,1}^j, \mathbf{x}_{k,2}^j, ..., \mathbf{x}_{k,N_k^j}^j)$, where N_k^j denotes the number of measurements labeled j at time k. The goal is to use these available measurements for each class to optimally estimate the means of classes through the fixed time interval.

Since data are assumed to be available off-line, the state estimation is referred to as smoother. The optimal minimum mean-square error smoother for the linear-Gaussian state space model is

Kalman Smoother [28]. Here, multiple measurements have been observed at each time point, as oppose to the regular Kalman Smoother which is developed for state estimation of systems with single observation at each time point. Here, the modified version of the Kalman Smoother capable of processing multiple measurements at each time step has been used. As most of smoothing techniques, the method has two main processes, *forward* and *backward*, as described in the following paragraphs.

In the forward Process, assuming known initialized states $\boldsymbol{\mu}^j_{1,n^j_1} = \boldsymbol{\mu}^j_1$ and the corresponding error covariance matrix $P^j_{1,n^j_1} = P^i_1$, two well-known *prediction* and *update* steps should be followed.

However, due to the existence of multiple measurements at each time step, one prediction is followed by several update steps to estimate the mean and error covariance matrix of each class at each time point. The process is summarized as follows:

Forward Process:

- Initialization:

$$\hat{\boldsymbol{\mu}}^j_{1,n^j_1} = \boldsymbol{\mu}^j_1, \, P^j_{1,n^j_1} = P^j_0 \,.$$

- Prediction Step: for $k = 2, ..., T$,

$$\begin{aligned}
\hat{\boldsymbol{\mu}}^j_{k,0} &= A^{j^T} \hat{\boldsymbol{\mu}}^j_{k-1,n^j_{k-1}} \,, \\
P^j_{k,0} &= A^j P^j_{k-1,n^j_{k-1}} A^{j^T} + B^j Q^j B^{j^T} \,.
\end{aligned}$$

- Update Step: for $k = 1, ..., T$, and for $i = 1, ..., N^j_k$, do:

$$\begin{aligned}
K^j_{k,i} &= P^j_{k,i-1} C^{j^T} (C^j P^j_{k,i-1} C^{j^T} + D^j R^j D^{j^T})^{-1} \,, \\
\hat{\boldsymbol{\mu}}^j_{k,i} &= \hat{\boldsymbol{\mu}}^j_{k,i-1} + K^j_{k,i} \left(\mathbf{x}^j_{k,i} - C^j \hat{\boldsymbol{\mu}}^j_{k,i-1} \right) \,, \\
P^j_{k,i} &= \left(I_d - K^j_{k,i} C^j \right) P^j_{k,i-1} \,,
\end{aligned}$$

where $K^j_{k,i}$ is the *Kalman Gain*, which is a function of the relative certainty of the measurements and the forward filter estimate. As time increases, the prediction error covariances P_{k,N^j_k} should converge to steady-state values. The backward process starts with the final filtering estimate, which is extrapolated backwards in time as follows [28]:

Backward Process:

- Initialization: for $k = T - 1, ..., 1$, do:

$$\hat{\boldsymbol{\mu}}^j_{T|T} = \hat{\boldsymbol{\mu}}^j_{T,N^j_T}, \; P^j_{T|T} = P^j_{T,N^j_T}.$$

- Backward Step:

$$\begin{aligned} L^j_k &= P^j_{k,N^j_k} A^{j^T} \left(P^j_{k+1,0}\right)^{-1}, \\ \hat{\boldsymbol{\mu}}^j_{k|T} &= \hat{\boldsymbol{\mu}}^j_{k,N^j_k} + L^j_k \left(\hat{\boldsymbol{\mu}}^j_{k+1|T} - \hat{\boldsymbol{\mu}}^j_{k+1,0}\right), \\ P^j_{k|T} &= P^j_{k,N^j_k} + L^j_k \left(P^j_{k+1|T} - P^j_{k+1,0}\right) L^{j^T}_k, \end{aligned}$$

where L^j_k is the *Smoother Gain*, which does not depend on the backward estimate. Its form reveals just a correction of the forward estimate using only the values computed in the forward process. In addition, the smoothed estimate $\hat{\boldsymbol{\mu}}^j_{k|T}$ does not depend on the smoothed covariance $P^j_{k|T}$. To obtain the smoothed estimate, only the forward state estimate and the smoothed gain have to be stored.

3.3 Adaptive Filter for Systems with Unknown Dynamics

When the parameters in equations (3.1) and (3.2) are only partially unknown, the Kalman Smoother , which is developed for state estimation with known linear-Gaussian state space equations, must be modified by performing an adaptive filter for simultaneous identification and state estimation of linear-Gaussian state space model. For example, when parameters initial states $\boldsymbol{\mu}^j_1$, matrix A^j and state noise covariance Q^j in eq (3.1) are unknown, we use the expectation maximization (EM) [32, 52] in combination with Kalman Smoother to estimate unknown parameters, simultaneously with the class mean themselves. However, in the case where corresponding time labels of available measurements are missing, we use Gaussian mixture model (GMM) [53] combined with Kalman Smoother to estimate time label and other unknown parameters.

3.3.1 EM-based Kalman Smoother

Ordinary maximum likelihood estimation attempts to find the value of the unknown parameter θ that maximizes the "incomplete" log-likelihood function. The EM algorithm considers instead the "complete" log-likelihood function, which includes the unknown state sequence, the assumption being that maximizing the complete log-likelihood is easier than maximizing the incomplete one.

For the jth class, the EM algorithm obtains a sequence of parameter estimates θ_n^j. Given the current estimates θ_n^j, the algorithm obtain the next estimate θ_{n+1}^j in the sequence by computing (E-step) the function as:

$$\mathcal{Q}(\theta_n^j, \hat{\theta}_n^j) = E\left[\log p\left(\boldsymbol{\mu}_{1:T}^j, \mathbf{x}_{1:T}^j \mid \theta^j\right) \mid \mathbf{x}_{1:T}^j, \hat{\theta}_n^j\right], \tag{3.4}$$

and then maximizing (M-step) this function:

$$\hat{\theta}_{n+1}^j = \underset{\theta^j}{\operatorname{argmax}}\, \mathcal{Q}(\theta^j, \hat{\theta}_n^j)\,. \tag{3.5}$$

Using (3.1) and (3.2), the joint log probability $\log p\left(\boldsymbol{\mu}_{1:T}^j, \mathbf{x}_{1:T}^j\right)$ in (3.4) is a sum of quadratic terms, and can be written as:

$$\begin{aligned}
\log p\left(\boldsymbol{\mu}_{1:T}^j, \mathbf{x}_{1:T}^j\right) \propto &- \sum_{k=1}^{T}\left(\frac{1}{2}\left[\mathbf{x}_k^j - C^j\boldsymbol{\mu}_k\right]^T R_k^{j^{-1}}\left[\mathbf{x}_k^j - C^j\boldsymbol{\mu}_k\right]\right) - \frac{T}{2}\log|R^j| \\
&- \sum_{k=2}^{T}\left(\frac{1}{2}\left[\boldsymbol{\mu}_k^j - A^j\boldsymbol{\mu}_{k-1}\right]^T Q_k^{j^{-1}}\left[\boldsymbol{\mu}_k^j - A^j\boldsymbol{\mu}_{k-1}\right]\right) - \frac{T-1}{2}\log|Q^j| \\
&- \frac{1}{2}\left[\boldsymbol{\mu}_1^j - \boldsymbol{\pi}_1\right]^T V_1^{j^{-1}}\left[\boldsymbol{\mu}_1^j - \boldsymbol{\pi}_1\right] - \frac{1}{2}\log|V_1^j|
\end{aligned} \tag{3.6}$$

where $\hat{\boldsymbol{\mu}}^j_{k|T}$ and $P^j_{k|T}$ can be obtained from the Kalman smoother tuned to parameter $\hat{\theta}^j_n$, and $P^j_{k,k-1|T}$ is defined as:

$$P_{k-1,k-2|T} = L^j_{k-1}\left(P_{k,k-1|T} - A^j P^j_{k-1|n_{k-1}}\right) L^{j^T}_{k-2} + P^j_{k-1|n_{k-1}} L^{j^T}_{k-2}\,, \tag{3.7}$$

which can be computed backward with terminal condition

$$P_{T-1,T-2|T} = \left(I_d - K^j_{T|n_T} C^j\right) A^j P_{T-1|N_{T-1}}\,. \tag{3.8}$$

The parameters of this system are C^j (output matrix), R^j (output noise covariance), A^j (state dynamics matrix), Q^j (state noise covariance), $\boldsymbol{\pi}^j_1$ (initial state mean), V^j_1 (initial state covariance). It can be shown that the E-Step computation in equation (3.5) can be performed by setting the derivative of $\mathcal{Q}(\theta^j, \hat{\theta}^j_n)$ to zero, which leads in to closed-form solutions:

- Output matrix:

$$\begin{aligned}
\frac{\partial \mathcal{Q}(\theta^j, \hat{\theta}^j_n)}{\partial C^j} &= -\sum_{k=1}^{T} R^{j^{-1}} \mathbf{x}^j_k \hat{\boldsymbol{\mu}}^{j^T}_{k|T} + \sum_{k=1}^{T} R^{j^{-1}} C^j \left(P^j_{k|T} + \hat{\boldsymbol{\mu}}^j_{k|T} \hat{\boldsymbol{\mu}}^{j^T}_{k|T}\right) = 0 \\
\hat{C}^j_{n+1} &= \left(\sum_{k=1}^{T} \mathbf{x}^j_k \hat{\boldsymbol{\mu}}^{j^T}_{k|T}\right) \left(\sum_{k=1}^{T} \left(P^j_{k|T} + \hat{\boldsymbol{\mu}}^j_{k|T} \hat{\boldsymbol{\mu}}^{j^T}_{k|T}\right)\right)^{-1}\,.
\end{aligned} \tag{3.9}$$

- Output noise covariance:

$$\begin{aligned}
\frac{\partial \mathcal{Q}(\theta^j, \hat{\theta}^j_n)}{\partial R^{j-1}} &= \frac{T}{2} R - \sum_{k=1}^{T} \left(\frac{1}{2} \mathbf{x}^j_k \mathbf{x}^{j^T}_k - C^j \hat{\boldsymbol{\mu}}^j_{k|T} \mathbf{x}^{j^T}_k + \frac{1}{2} C \left(P^j_{k|T} + \hat{\boldsymbol{\mu}}^j_{k|T} \hat{\boldsymbol{\mu}}^{j^T}_{k|T}\right) C^T\right) = 0 \\
\hat{R}^j_{n+1} &= \frac{1}{T} \sum_{k=1}^{T} \left(\mathbf{x}^j_k \mathbf{x}^{j^T}_k - \hat{C}^j_{n+1} \hat{\boldsymbol{\mu}}^j_{k|T} \mathbf{x}^{j^T}_k\right)\,.
\end{aligned} \tag{3.10}$$

- State dynamics matrix:

$$
\begin{aligned}
\frac{\partial \mathcal{Q}(\theta^j, \hat{\theta}^j_n)}{\partial A^j} &= -\sum_{k=2}^{T} Q^{j^{-1}} \left(P^j_{k,k-1|T} + \hat{\boldsymbol{\mu}}^j_{k|T} \hat{\boldsymbol{\mu}}^{j^T}_{k-1|T} \right) + \sum_{k=2}^{T} Q^{j^{-1}} A^j \left(P^j_{k-1|T} + \hat{\boldsymbol{\mu}}^j_{k-1|T} \hat{\boldsymbol{\mu}}^{j^T}_{k-1|T} \right) = 0 \\
\hat{A}^j_{n+1} &= \left(\sum_{k=2}^{T} \left(P^j_{k,k-1|T} + \hat{\boldsymbol{\mu}}^j_{k|T} \hat{\boldsymbol{\mu}}^{j^T}_{k-1|T} \right) \right) \left(\sum_{k=2}^{T} \left(P^j_{k-1|T} + \hat{\boldsymbol{\mu}}^j_{k-1|T} \hat{\boldsymbol{\mu}}^{j^T}_{k-1|T} \right) \right)^{-1} . \qquad (3.11)
\end{aligned}
$$

- State noise covariance:

$$
\begin{aligned}
\frac{\partial \mathcal{Q}(\theta^j, \hat{\theta}^j_n)}{\partial Q^{j^{-1}}} &= \frac{T-1}{2} Q - \frac{1}{2} \left(\sum_{k=2}^{T} \left(P^j_{k|T} + \hat{\boldsymbol{\mu}}^j_{k|T} \hat{\boldsymbol{\mu}}^{j^T}_{k|T} \right) - \hat{A}^j_{n+1} \sum_{k=1}^{T} \left(P^j_{k,k-1|T} + \hat{\boldsymbol{\mu}}^j_{k|T} \hat{\boldsymbol{\mu}}^{j^T}_{k-1|T} \right)^T \right) = 0 \\
\hat{Q}^j_{n+1} &= \frac{1}{T-1} \left(\sum_{k=2}^{T} \left(P^j_{k|T} + \hat{\boldsymbol{\mu}}^j_{k|T} \hat{\boldsymbol{\mu}}^{j^T}_{k|T} \right) - \hat{A}^j_{n+1} \sum_{k=1}^{T} \left(P^j_{k,k-1|T} + \hat{\boldsymbol{\mu}}^j_{k|T} \hat{\boldsymbol{\mu}}^{j^T}_{k-1|T} \right)^T \right) . \qquad (3.12)
\end{aligned}
$$

- Initial state mean:

$$
\begin{aligned}
\frac{\partial \mathcal{Q}(\theta^j, \hat{\theta}^j_n)}{\partial \boldsymbol{\pi}^j_1} &= \left(\hat{\boldsymbol{\mu}}^j_1 - \boldsymbol{\pi}_1 \right) V_1^{j^{-1}} = 0 \\
\hat{\boldsymbol{\pi}}_1{}^j_{n+1} &= \hat{\boldsymbol{\mu}}^j_1 . \qquad (3.13)
\end{aligned}
$$

- Initial state covariance:

$$
\begin{aligned}
\frac{\partial \mathcal{Q}(\theta^j, \hat{\theta}^j_n)}{\partial V_1^{j^{-1}}} &= \frac{1}{2} V^j_1 - \frac{1}{2} \left(\left(P^j_{1|T} + \hat{\boldsymbol{\mu}}^j_{1|T} \hat{\boldsymbol{\mu}}^{j^T}_{1|T} \right) - \hat{\boldsymbol{\mu}}^j_1 \boldsymbol{\pi}^{j^T}_1 - \boldsymbol{\pi}^j_1 \hat{\boldsymbol{\mu}}^{j^T}_1 + \hat{\boldsymbol{\mu}}^j_1 \hat{\boldsymbol{\mu}}^{j^T}_1 \right) = 0 \\
\hat{V}_1{}^j_{n+1} &= \left(P^j_{1|T} + \hat{\boldsymbol{\mu}}^j_{1|T} \hat{\boldsymbol{\mu}}^{j^T}_{1|T} \right) - \hat{\boldsymbol{\mu}}^j_1 \hat{\boldsymbol{\mu}}^{j^T}_1 . \qquad (3.14)
\end{aligned}
$$

The process continues until the maximum absolute difference between the estimated parameters in two consecutive steps gets smaller than the prespecified value ϵ. After stopping the EM algorithm for both classes, the Kalman Smoother introduced in previous section can be run tuned to the inferred θ^j parameters to estimate the means of classes at different time steps.

3.3.2 GMM-based Kalman Smoother

In this subsection, a way of simultaneous estimation of measurement time-labels and parameters of the class conditional distributions are provided. Let θ^j be the set of unknown parameters of class j, where θ^j might contain the initial mean or covariance matrix, dynamics of state or measurement process or even the statistics of noises. This parameter vector is assumed to take its values from the space Θ^j. Letting $p(\theta^j)$ be the prior information about the parameter vector θ^j, the maximum a posteriori estimate (MAP) of joint labels and parameters can be computed as follows:

$$(\theta^*, \{t_1^{*,j}, ..., t_n^{*,j}\}) = \underset{\hat{\theta} \in \Theta^j, \{t_1, ..., t_n\} \in (\mathbb{N}^+)^n}{\text{argmin}} \quad p(\hat{\theta}, \{t_1, ..., t_n\} \mid S_n^j) \tag{3.15}$$

where $\mathbb{N}^+ = \{1, 2, ...\}$. Finding solution for minimization in (3.15) can be so challenging.

In this case, the combination of the well-known mixture Gaussian model (GMM) and Kalman filter (KF) are used for joint estimation of time labels and parameters in equation (3.15). Let $\hat{\theta}$ be a vector containing parameters of the class conditional distribution. Assuming the time labels are independent of measurements, the joint distribution of parameters and time labels can be approximated as:

$$\{t_1^{\hat{\theta},j}, ..., t_n^{\hat{\theta},j}\} \quad = \underset{\{t_1, ..., t_n\} \in (\mathbb{N}^+)^n}{\text{argmin}} \quad P(\{t_1, ..., t_n\} \mid \hat{\theta})\,. \tag{3.16}$$

The minimization in equation (3.16) can be computed efficiently using the Gaussian mixture model(GMM) [53]. Using the linear structure for class conditional distribution, it is easy to show that the projection of the initial mean and covariance matrix of the class conditional distribution represented by parameter vector $\hat{\theta}$ will result in the following means and covariance matrices under the GMM clustering algorithm:

$$\begin{aligned} \boldsymbol{\mu}_{k,\hat{\theta}}^{\text{GMM},j} &= A_{\hat{\theta}}^{j}\, \boldsymbol{\mu}_{k-1,\hat{\theta}}^{\text{GMM},j}, \\ \Sigma_{k,\hat{\theta}}^{\text{GMM},j} &= A_{\hat{\theta}}^{j}\, \Sigma_{k-1,\hat{\theta}}^{\text{GMM},j}\, (A_{\hat{\theta}}^{j})^T + Q_{\hat{\theta}}^{j} + R_{\hat{\theta}}^{j}, \end{aligned} \tag{3.17}$$

for $k = 1, 2, ...$; where $\mathcal{N}(\boldsymbol{\mu}_{k,\hat{\theta}}^{\mathrm{GMM},j}, \Sigma_{k,\hat{\theta}}^{\mathrm{GMM},j})$ specifies the Gaussian distribution over the kth cluster, $A_{\hat{\theta}}^{j}, Q_{\hat{\theta}}^{j}, R_{\hat{\theta}}^{j}$ are parameters of state space model represented by a parameter vector $\hat{\theta}$, and $\boldsymbol{\mu}_{0,\hat{\theta}}^{\mathrm{GMM},j} = \boldsymbol{\mu}_{0,\hat{\theta}}^{j}, \Sigma_{0,\hat{\theta}}^{\mathrm{GMM},j} = P_{0,\hat{\theta}}^{j}$. Since, the GMM is the soft clustering technique, we assign each measurement to a cluster with the highest probability. The time assigned to $\mathbf{X}_j \in S_n^j$ is as follows:

$$t_i^{j,\hat{\theta}} = \arg\max_{k=1,2,...} \mathcal{N}(\mathbf{X}_j; \boldsymbol{\mu}_{k,\hat{\theta}}^{\mathrm{GMM},j}, \Sigma_{k,\hat{\theta}}^{\mathrm{GMM},j}) \tag{3.18}$$

for $i = 1, ..., n$; where $\mathcal{N}(\mathbf{x}; \boldsymbol{\mu}, \Sigma)$ specifies the probability of sample $\mathbf{x}$ in Gaussian distribution with mean $\boldsymbol{\mu}$ and covariance matrix Σ.

Now, one needs to compute the posterior probability of the parameter vector $\hat{\theta}$ under the estimated time labels by the GMM method. This can be written as:

$$p(\hat{\theta} \mid \{t_1^{\hat{\theta}}, ..., t_n^{\hat{\theta}}\}, S_n^j) \propto p(\theta^j = \hat{\theta})\, p(S_n^j \mid \hat{\theta}, \{t_1^{\hat{\theta}}, ..., t_n^{\hat{\theta}}\}), \tag{3.19}$$

The second term in the right hand side of above equation specifies the likelihood of the measurement set given the estimated parameter vector $\hat{\theta}$ and the assigned labels. It should be noted that there might be more than one sample at any given time point. This suggests that the prediction step of Kalman filter should be followed by several update steps to estimate the likelihood function of model represented by the parameter vector $\hat{\theta}$.

There are several ways, such as Markov Chain Monte Carlo (MCMC) methods, for finding the parameter $\hat{\theta} \in \Theta$ which maximizes the posterior distribution in (3.19) [32, 34]. Without loss of generality, letting $\Theta = \{\hat{\theta}^1, ..., \hat{\theta}^1\}$ be the set of estimated parameters, the maximum aposterior estimate of θ and time labels can be approximated as:

$$\hat{\theta}^{*,j} = \arg\max_{\hat{\theta} \in \Theta} \log p(\theta^j = \hat{\theta}) + \log p(S_n^j \mid \hat{\theta}, \{t_1^{\hat{\theta},j}, ..., t_n^{\hat{\theta},j}\}). \tag{3.20}$$

Notice that the second term in the right hand side of above equation specifies the log-likelihood of the Kalman filter tuned to vector $\hat{\theta}$ obtained based on the assigned time labels $\{t_1^{\hat{\theta},j}, ..., t_n^{\hat{\theta},j}\}$. For more information see [32].

After estimating the best set of parameters and their associated time labels, one need to estimate the means and covariance matrix of class conditional distributions at different time points. This can be done by performing the backward process of Kalman smoother (KS) based on the results computed by the KF tuned to the parameter vector $\hat{\theta}^{*,j}$ and estimated time labels $\{t_1^{\hat{\theta},j}, ..., t_n^{\hat{\theta},j}\}$. The output of the KS specifies the mean vectors and covariance matrices of class conditional distributions over time.

3.4 EM-based and GMM-based Nonstationary Linear Discriminant Analysis

After computation of the class conditional distributions at various time steps, one can use nonstationary linear discriminant analysis for classification at each time step. Letting $\Pi_k^0 \sim \mathcal{N}\left(\boldsymbol{\mu}^0_{k|T,\hat{\theta}^*}, P^0_{k|T,\hat{\theta}^*}\right)$ and $\Pi_k^1 \sim \mathcal{N}\left(\boldsymbol{\mu}^1_{k|T,\hat{\theta}^*}, P^1_{k|T,\hat{\theta}^*}\right)$ be the class conditional distributions at time step k, where $\boldsymbol{\mu}^i_{k|T,\hat{\theta}^*}$ and $P^i_{k|T,\hat{\theta}^*}$ are the mean vector and covariance matrix computed by the KS at time step k tuned to $\hat{\theta}^{*,i}$. The optimal quadratic discriminant is [29]

$$D_k\left(\mathbf{x}\right) = \mathbf{x}^T E_k \mathbf{x} + F_k^T \mathbf{x} + G_k\,, \tag{3.21}$$

where

$$\begin{aligned} E_k &= -\frac{1}{2}\left((\Sigma^1_{k|T} + R^1)^{-1} - (\Sigma^0_{k|T} + R^0)^{-1}\right) \\ F_k &= (\Sigma^1_{k|T} + R^1)^{-1}\hat{\boldsymbol{\mu}}^1_{k|T} - (\Sigma^0_{k|T} + R^0)^{-1}\hat{\boldsymbol{\mu}}^0_{k|T} \\ G_k &= -\frac{1}{2}(\hat{\boldsymbol{\mu}}^1_{k|T})^T(\Sigma^1_{k|T} + R^1)^{-1}\hat{\boldsymbol{\mu}}^1_{k|T} \\ &\quad + \frac{1}{2}(\hat{\boldsymbol{\mu}}^0)^T(\Sigma^0_{k|T} + R^0)^{-1}(\hat{\boldsymbol{\mu}}^0_{k|T})^T - \frac{1}{2}\log\frac{|\Sigma^1_{k|T} + R^1|}{|\Sigma^0_{k|T} + R^0|}, \end{aligned} \tag{3.22}$$

with hyper-quadratic optimal decision boundary $D_k = P\left(Y_k = 1\right)/P\left(Y_k = 0\right)$, for $k = 1, 2, ...$

The designed Nonstationary Linear Discriminant Analysis (NSLDA) classifier is defined by substituting the estimates for the unknown parameters in the equation (3.22):

$$\psi_k(\mathbf{x}) = \begin{cases} 1, & D_k(\mathbf{x}) \leq 0 \\ 0, & \text{otherwise} \end{cases}, \tag{3.23}$$

The whole process of the proposed EM-based and GMM-based Nonstationary Linear Discriminant Analysis is summarized in Algorithm 1 and 2, respectively.

Algorithm 1 EM-based Linear Non-Stationary Discriminant Analysis

```
**For** $j = \{0, 1, ...\}$, do:
    - EM process:
        - Initial guess: $\hat{\theta}_0^j$.
        - $n = 0$.
          **Repeat**
            - E-Step: Run Kalman smoother tuned to $\hat{\theta}_n^j$,
            - M-Step: Update estimation of $\theta^j$ using (3.9) - (3.14)
            - $n = n + 1$.
          **Until** $\left(\max|\hat{\theta}_n^j - \hat{\theta}_{n-1}^i|\right) < \epsilon$
    - Set $\hat{\theta}^j = \hat{\theta}_n^j$.
    - Run Kalman smoother tuned to $\hat{\theta}^i$ and obtain $\hat{\boldsymbol{\mu}}_k^i$ and $P_k^j$.
  **EndFor**
- NSLDA: plug $\hat{\boldsymbol{\mu}}_k^j$ and $P_k^j$ into (3.23), to obtain the classifier $\psi_k(\mathbf{x})$ for $k = 1, ...T$.
```

Algorithm 2 GMM-based Non-Stationary Linear Discriminant Analysis

```
for $\hat{\theta} \in \{\hat{\theta}^1, ..., \hat{\theta}^M\}$ do
    for $i = 0, 1$ do
        $\boldsymbol{\mu}_{0,\hat{\theta}}^{\mathrm{GMM},i} = \boldsymbol{\mu}_{0,\hat{\theta}},\ \Sigma_{0,\hat{\theta}}^{\mathrm{GMM},i} = P_{0,\hat{\theta}}$
        for $k = 1, 2, ..., n$ do
            $\boldsymbol{\mu}_{k,\hat{\theta}}^{\mathrm{GMM},i} = A_{\hat{\theta}}^{i}\, \boldsymbol{\mu}_{k-1,\hat{\theta}}^{\mathrm{GMM},i}$.
            $\Sigma_{k,\hat{\theta}}^{\mathrm{GMM},i} = A_{\hat{\theta}}^{i}\, \Sigma_{k-1,\hat{\theta}}^{\mathrm{GMM},i}\, (A_{\hat{\theta}}^{i})^T + Q_{\hat{\theta}}^{i} + R_{\hat{\theta}}^{i}$.
        end for
        $t_j^{i,\hat{\theta}} = \underset{k=1,2,...,n}{\arg\max}\, \mathcal{N}(\mathbf{X}_j; \boldsymbol{\mu}_{k,\hat{\theta}}^{\mathrm{GMM},i}, \Sigma_{k,\hat{\theta}}^{\mathrm{GMM},i}),\ j = 1, 2, ..., n$
    end for
    $T^{\hat{\theta}} = \max(t_1^{0,\hat{\theta}}, ..., t_1^{0,\hat{\theta}}, t_1^{1,\hat{\theta}}, ..., t_1^{1,\hat{\theta}})$
    Run Kalman Filter tuned to $\hat{\theta}$
    $\{\hat{\boldsymbol{\mu}}_{k|k,\hat{\theta}}^{i}, P_{k|k,\hat{\theta}}^{i}\}_{k=0}^{T^{\hat{\theta}}}, L^i(\hat{\theta}) \leftarrow \mathrm{KF}(\hat{\theta}, \{t_1^{i,\hat{\theta}}, ..., t_n^{i,\hat{\theta}}\}, S_n^i), i = 0, 1$
end for
Time Label and Parameter Estimation, for $i = 0, 1$:
    $$\hat{\theta}^{*,i} = \underset{\hat{\theta} \in \{\hat{\theta}^1, ..., \hat{\theta}^M\}}{\arg\max} \log p(\theta^i = \hat{\theta}) + L^i(\hat{\theta})$$
$T = T^{\hat{\theta}^*}$.
Run a Kalman Smoother (KS) tuned to $\hat{\theta}^*$, for $i = 0, 1$:
    $\{\hat{\boldsymbol{\mu}}_{k|T,\hat{\theta}}^{i}, P_{k|T,\hat{\theta}}^{i}\}_{k=0}^{T} \leftarrow \mathrm{KS}(\hat{\theta}^{*,i}, \{t_1^{i,\hat{\theta}^{*,i}}, ..., t_n^{i,\hat{\theta}^{*,i}}\}, S_n^i)$.
NSLDA:
    $$\psi_k(\mathbf{x}) = \begin{cases} 1, & \left(\mathbf{x} - \frac{\hat{\boldsymbol{\mu}}_{k|T}^{0} + \hat{\boldsymbol{\mu}}_{k|T}^{1}}{2}\right)^T \Sigma_k^{-1} \left(\hat{\boldsymbol{\mu}}_{k|T}^{0} - \hat{\boldsymbol{\mu}}_{k|T}^{1}\right) \leq 0 \\ 0, & \text{otherwise} \end{cases}$$
```

3.5 Simulation Results and Discussion

To study the performance of our proposed state-space models to nonstationary discriminant analysis, we designed several numeric experiments in linear drift models. All simulation results compare the error rates of the “naive” discriminant analysis that did not consider population drift between our proposed nonstationary discriminant analysis.

3.5.1 Systems with Fully-Known Dynamics

When all matrices in both state and measurement models of two classes are assumed to be known, the parameters settings are shown in Table 3.1. Fig. 3.1 shows the average error for naive LDA and NSLDA for different values of T. We can see that both classification rules perform better as the number of time points increases, but NSLDA overall performs much better than naive LDA.

Parameters	Values
Total time points	T = 3, 4, 5, ..., 18, 19, 20
Sample size	n = 10
Dimensionality	d = 2
Initial means	$\mu_1^0 = \begin{bmatrix}1\\1\end{bmatrix}, \mu_1^1 = \begin{bmatrix}1.5\\1.5\end{bmatrix}$
Initial covariances	$P_1^0 = 0.5I_d, P_1^1 = 0.5I_d$
Evolution matrix	$A^0 = \begin{bmatrix}1 & 0.1\\0.1 & 1\end{bmatrix}, A^1 = \begin{bmatrix}0.99 & 0\\0 & 0.99\end{bmatrix}$ $C^0 = I_d, C^1 = I_d, B^0 = I_d, B^1 = I_d, D^0 = I_d, D^1 = I_d$
Noise	$Q^0 = 0.1I_d, Q^1 = 0.1I_d, R^1 = 0.2I_d, R^1 = 0.2I_d$

Table 3.1: Parameter settings for systems with fully-known dynamics

3.5.2 Systems with Partially-Known Dynamics

In this subsection, we showed two case studies for linear-drift numerical experiments, where the parameter settings in Eq (3.1) and Eq (3.2) are shown in Table 3.2.

- *Case One:* initial states $\boldsymbol{\mu}_0$, matrices A in Eq (3.1) are unknown.
- *Case Two:* initial states $\boldsymbol{\mu}_0$, matrices A in Eq (3.1) and corresponding time labels of available measurements in Eq (3.2) are unknown.

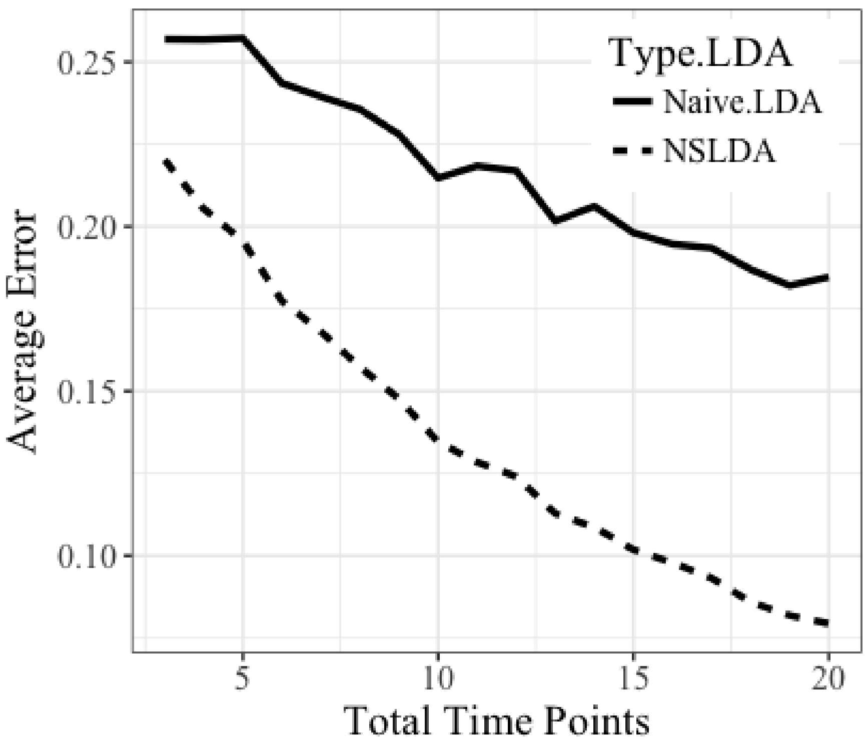


Figure 3.1: Average errors for naive LDA and NSLDA for a fully-known system.

To estimate the unknown parameters and states, for the first case, we used the EM-based Kalman Smoother (showed in Section 3.3.1); while for the second case, we applied the GMM-based Kalman Smoother (showed in Section 3.3.2). To examine the performance of our modified Kalman Smoother, we defined the Mean Square Error (MSE) at time k as the sum of diagonal elements of the error covariance matrix $E\left[\left(\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}}_{k|T,\hat{\theta}^\star}\right)^T \times \left(\boldsymbol{\mu}_k - \hat{\boldsymbol{\mu}}_{k|T,\hat{\theta}^\star}\right)\right]$.

Last, we plug the estimated states from the modified Kalman Smoother into our proposed NSLDA (showed in Section 3.4), and obtained the average error estimates, and compare the “naive” LDA using pooled sample means and covariances of measurements. For above steps, both NSLDA and naive LDA, 1000 Monte Carlo simulated data sets are employed.

3.5.2.1 Case One Discussion

While initial states $\boldsymbol{\mu}_0^c$, matrices A^c, and in model (3.1) are unknown, as shown in the Algorithm 1, we initialize values of these unknown parameters below in Table 3.3:

Parameters	Value
Total time points	T = 4, 6, 8, 10, 12
Sample size	n = 20
Dimensionality	d = 2
Initial means	$\mu_0^0 = \begin{bmatrix} 1.2 \\ 0.5 \end{bmatrix}, \mu_0^1 = \begin{bmatrix} 0.2 \\ 1.5 \end{bmatrix}$
Initial covariances	$P_0^0 = 0.1 I_d, P_0^1 = 0.1 I_d$
Evolution matrix	$A^0 = \begin{bmatrix} 1.3 & 0.2 \\ 0.9 & 0.6 \end{bmatrix}, A^1 = \begin{bmatrix} 0.9 & 0.8 \\ 1 & 0.5 \end{bmatrix}$ $C^0 = I_d, C^1 = I_d, B^0 = I_d,$ $B^1 = I_d, D^0 = I_d, D^1 = I_d$
Noise	$Q^0 = 0.1 I_d, Q^1 = 0.1 I_d,$ $R^1 = 0.2 I_d, R^1 = 0.2 I_d$

Table 3.2: Parameter settings for case in 3.5.2

Parameters	Value
Initial means	$\mu_{0,\text{guess}}^0 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \mu_{0,\text{guess}}^1 = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$
Matrix A	$A_{\text{guess}}^0 = \begin{bmatrix} 1 & 1.5 \\ 0.8 & 1.2 \end{bmatrix}, A_{\text{guess}}^1 = \begin{bmatrix} 0.8 & 1.3 \\ 0.5 & 2 \end{bmatrix}$

Table 3.3: Initial parameter settings for case in 3.5.2

By repeating the E-steps and M-steps in the Algorithm 1, the unknown parameters got converged. In the meanwhile, estimated states using these converged parameters are very close to true states.

3.5.2.2 Case Two Discussion

In this numerical experiment, in contrast to *Case One*, we assume state model noise covariance Q is known, but time labels of measurements are unknown. Therefore to estimate time labels, unknown parameters initial states $\boldsymbol{\mu}_0^c$, matrices A^c, as well as states in the state space model, GMM-based Kalman Smoother was applied. As shown in the Algorithm 2, we also initialize values of these unknown parameters in Table 3.3, but have Q^c known. It turns out that the time labels can be estimated correctly by our algorithm, and the estimated states using the estimated parameters and time labels are very close to true states.

The last step for these two case studies is nonstationary discriminant analysis. The estimated states from our modified Kalman Smoother provided the information of conditional population of the dataset, therefore can be used in our proposed NSLDA in Section 3.4. On contrast, the "naive" LDA use estimates the means by pooling all the data. Fig. 3.2 and 3.3 show the average errors of different classifiers for different values of T. From the figure of error estimate, nonstationary discriminant analysis classifiers performs better than naive LDA and SVM. The NSQDA perform better when the distributions overlap in complicated ways. As time goes, the distributions are more linearly separable, then you won't see much difference between these two.

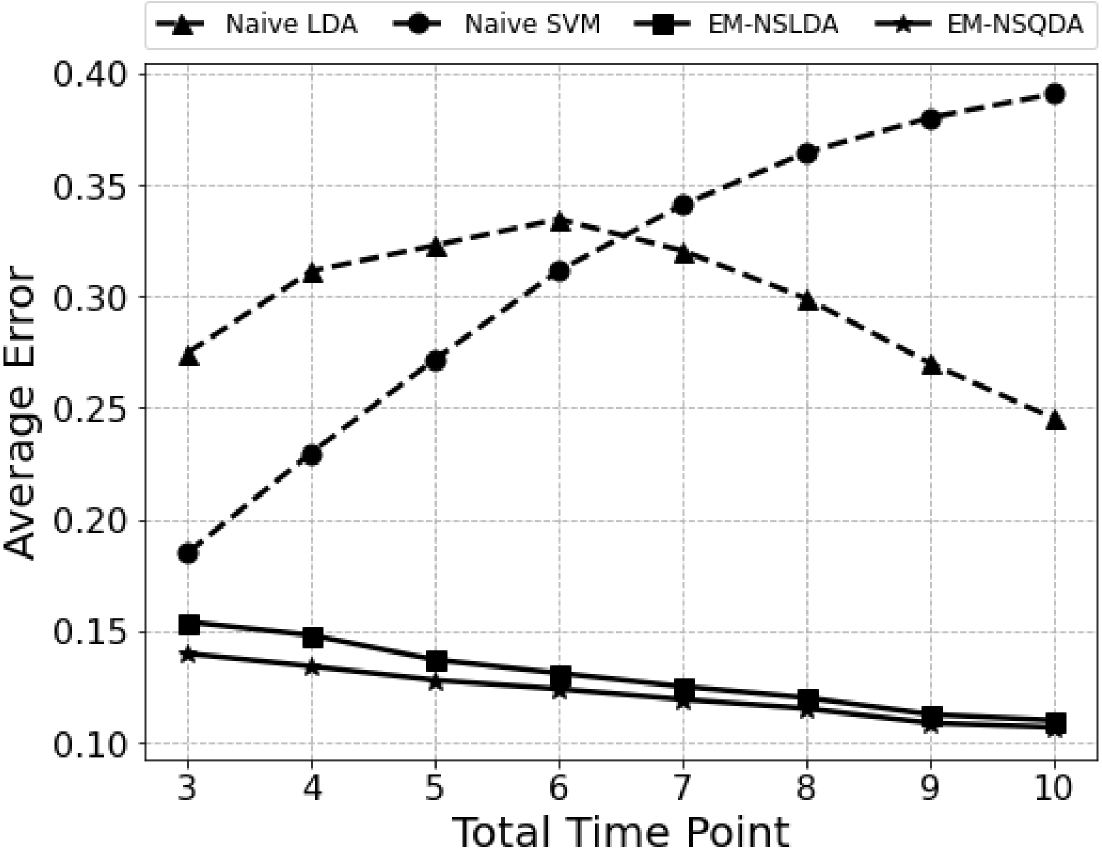


Figure 3.2: Average classification errors for linear drift numeric experiment in Section 3.5.2 *Case One*. The dash line marked by ▲ represent the "naive" LDA error rate and dash line marked by • shows the "naive" support vector machine error rate. Two solid lines are for nonstationary classifier results, where ■ represents EM-Based NSLDA ; ★ presents EM-Based NSQDA.

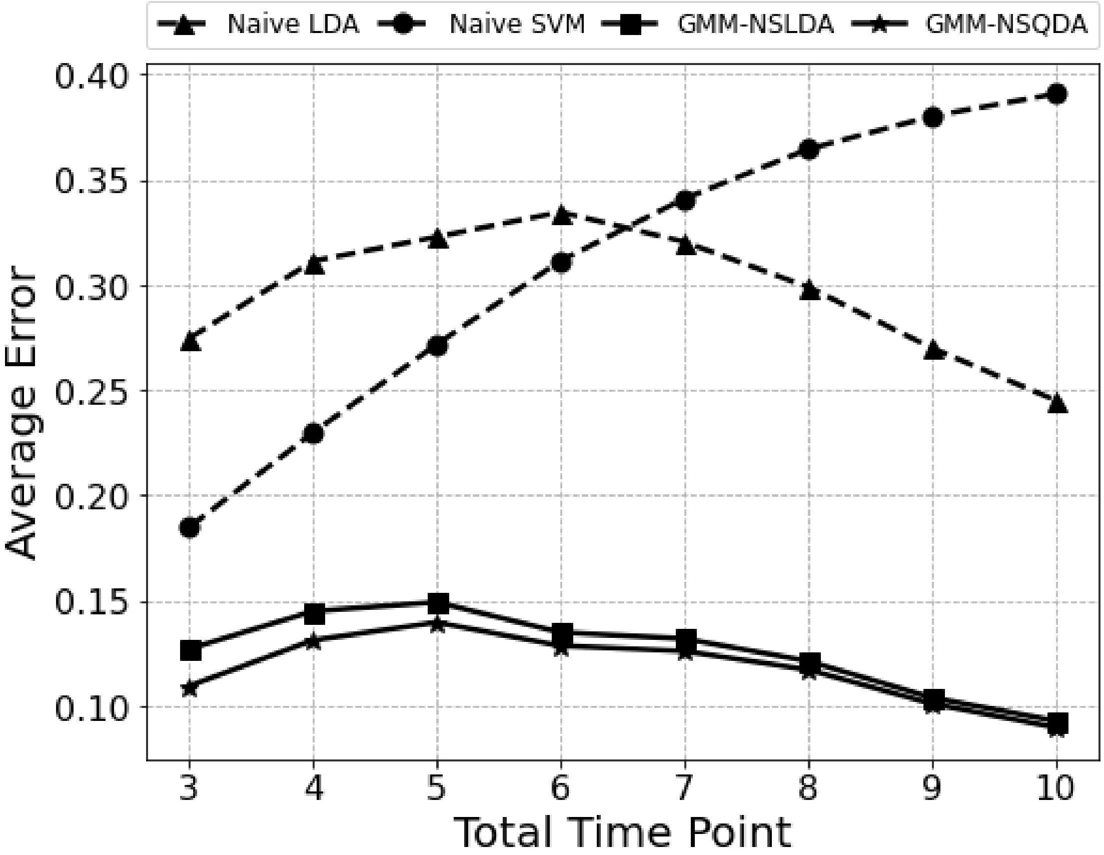


Figure 3.3: Average classification errors for linear drift numeric experiment in Section 3.5.2 *Case Two*. The dash line marked by ▲ represent the "naive" LDA error rate and dash line marked by • shows the "naive" support vector machine error rate. Two solid lines are for nonstationary classifier results, where ■ represents GMM-Based NSLDA ; ★ presents GMM-Based NSQDA.

4. NONLINEAR STATE SPACE MODELS TO NONSTATIONARY DISCRIMINANT ANALYSIS

4.1 Nonlinear Drift Model

In the previous two chapters, the restrictive assumptions such as linearity of the state-space model or Gaussianity of the noise process, are omitted by modeling the evolution of the class conditional distributions using general state-space models. Training classifiers at any given time point using the available data from various classes might not be practical or lead to poor classification performance. This is due to factors such as data limitation, missing data, or large noise in the data. To overcome these difficulties, we propose using sequential Monte-Carlo (SMC) techniques [36, 37] for efficient estimation of the class-conditional distributions via a finite set of particles. This is achieved by the use of a particle smoother technique [54], modified here to handle multiple data at each time point. Upon representing the underlying process of the class-conditional distributions, any discriminant analysis classifiers can be employed for decision making using the sets of particles at different time points. We need to define the nonlinear drifts in class-conditional distributions.

In a multiclass nonstationary problem with c classes and T time points, we assume that the centroid of each class is a latent variable that evolves in time according to the following nonlinear model:

$$\mathbf{z}_k^j = \mathbf{f}_k^j\left(\mathbf{z}_{k-1}^j, \mathbf{w}_k^j\right), \tag{4.1}$$

for $j = 0, 1, \ldots, c-1$ and $k = 1, \ldots, T$, where $\mathbf{f}_k^j$ is an arbitrary nonlinear function governing the evolution of class j and $\mathbf{w}_k^j$ defines an i.i.d. transition noise process, which is independent of the $\mathbf{z}_k^j$ process. The initial states $\mathbf{z}_0^j$ are generated from given starting "prior" distributions.

For notational simplicity, we partition the training data into $c \times T$ subsamples

$$S_k^j = \{\mathbf{x}_{k,1}^j, \ldots, \mathbf{x}_{k,n_k^j}^j\}, \tag{4.2}$$

for $j = 0, \ldots, c-1$ and $k = 1, \ldots, T$, where n_k^j are the sample sizes for each class j at time k, adding up to the total sample size n. Nothing is assumed in this paper about the sampling mechanism; e.g., for fixed k, n_k^j could be a random variable or a fixed experimental design parameter (future work will examine the sampling issue). The data are assumed to satisfy the following general observation model:

$$\mathbf{x}_{k,i}^j = \mathbf{h}_k^j(\mathbf{z}_k^j, \mathbf{v}_{k,i}^j)\,, \tag{4.3}$$

for $i = 0, 1, \ldots, n_k^j$, $j = 0, 1, \ldots, c-1$ and $k = 1, \ldots, T$, where $\mathbf{h}_k^j$ is an arbitrary nonlinear function mapping the latent variables to the observable data and $\mathbf{v}_{k,i}^j$ defines an i.i.d. observation noise process, which is independent of the $\mathbf{z}_k^j$ process.

4.2 Particle Smoother for Nonstationary Classification Model Inference

Our ultimate goal is to developed a framework for nonstationary classification based on the model described in the previous section (for short, the NCS model). For that purpose, the primary task is to estimate the latent variables $\mathbf{z}_k^j$, but it may also be necessary to estimate the noise parameters. In this paper, we will assume that all noise parameters are known and focus on the estimation of the latent variables given the data. In the next section, we build a classification rule using the results of this section.

Several inference methods exist in the literature that can handle the nonlinearity and non-Gaussianity of the NSC model [37, 55]. In this paper, we propose a method that is based on the particle smoother in [54], with a suitable modification to handle multiple independent data at each time point. For each class, the smoother consists of forward and backward processes. In the forward process a particle filter algorithm is run to compute the forward particles and weights characterizing the filtering distributions. The approximate smoothed distribution is computed by running a backward process for correcting the filtering weights. We describe each of these steps next.

4.2.1 Forward Process

The auxiliary particle filter (APF) [56] is a sequential Monte-Carlo (SMC) method, which efficiently predicts the location of "particles" with high probability at time step k using information up to time step $k-1$ via an auxiliary variable ζ_k. Let

$$S^j_{1:k} = S^j_1 \cup \cdots \cup S^j_1 \tag{4.4}$$

be the the data available for class j *up to* time k. The method first draws a sample (the "particles") from the joint distribution $p(\mathbf{z}^j_k, \zeta_k \mid S^j_{1:k})$, then drops the auxiliary variable to obtain particles from $p(\mathbf{z}^j_k \mid S^j_{1:k})$, for $k = 1, \ldots, T$.

Let $\{\tilde{\mathbf{z}}^j_{k-1,i}, w^j_{k-1,i}\}^N_{i=1}$ be N particles and their associated weights at time $k-1$, which approximate $p(\mathbf{z}^j_{k-1} \mid S^j_{1:k-1})$. The process consists of two stages. The *first stage* weights are computed as:

$$v^j_{k,i} = p(S^j_k \mid \nu^j_{k,i}) w^j_{k-1,i}\,, \tag{4.5}$$

for $i = 1, \ldots, N$; where $\nu^j_{k,i}$ is a characteristic of $\mathbf{z}^j_k$ given $\tilde{\mathbf{z}}^j_{k-1,i}$, which can be the mean, the mode or even a point sampled from $p(\mathbf{z}^j_k \mid \tilde{\mathbf{z}}^j_{k-1,i})$ [56]. The auxiliary variables $\{\zeta_{k,i}\}^N_{i=1}$ are sampled from the weights:

$$\{\zeta_{k,i}\}^N_{i=1} \sim \mathrm{Cat}(\{\tilde{v}^j_{k,i}\}^N_{i=1})\,, \tag{4.6}$$

where $\{\tilde{v}^j_{k,i}\}^N_{i=1}$ denotes the normalized first-stage weights, and Cat$(a_1, \ldots, a_N)$ is the discrete categorical distribution with probability mass function $p(i) = a_i$. Finally, the new particles are obtained via

$$\{\tilde{\mathbf{z}}^j_{k,i}\}^N_{i=1} \sim p(\mathbf{z}^j_k \mid \tilde{\mathbf{z}}^j_{k-1,\zeta_{k,i}})\,, \tag{4.7}$$

with associated *second-stage* weights

$$w^j_{k,i} = \frac{p(S^j_k \mid \tilde{\mathbf{z}}^j_{k,i})}{p(S^j_k \mid \nu^j_{k,\zeta_{k,i}})} = \prod_{j=1}^{n^j_k} \frac{p(\mathbf{x}^j_{k,j} \mid \tilde{\mathbf{z}}^j_{k,i})}{p(\mathbf{x}^j_{k,j} \mid \nu^j_{k,\zeta_{k,i}})}\,, \tag{4.8}$$

for $i = 1, \ldots, N$. Iterating the previous process from $k = 1$ to T leads to the full set of forward particles and weights $\{\tilde{\mathbf{z}}^j_{0:T,i}, w^j_{0:T,i}\}_{i=1}^N$.

4.2.2 Backward Process

The backward process is based on the following equation:

$$
\begin{aligned}
&p(\mathbf{z}^j_k \mid S^j_{1:T}) \\
&= \int_{\mathbf{z}^j_{k+1}} p(\mathbf{z}^j_k \mid \mathbf{z}^j_{k+1}, S^j_{1:T})\, p(\mathbf{z}^j_{k+1} \mid S^j_{1:T})\, d\mathbf{z}^j_{k+1} \\
&= \int_{\mathbf{z}^j_{k+1}} p(\mathbf{z}^j_k \mid \mathbf{z}^j_{k+1}, S^j_{1:k})\, p(\mathbf{z}^j_{k+1} \mid S^j_{1:T})\, d\mathbf{z}^j_{k+1} \\
&= \int_{\mathbf{z}^j_{k+1}} \frac{p(\mathbf{z}^j_{k+1} \mid \mathbf{z}^j_k) p(\mathbf{z}^j_k \mid S^j_{1:k}) p(\mathbf{z}^j_{k+1} \mid S^j_{1:T})}{p(\mathbf{z}^j_{k+1} \mid S^j_{1:k})}\, d\mathbf{z}^j_{k+1}\,,
\end{aligned}
$$

where $k < T$ and $p(\mathbf{z}^j_{k+1} \mid S^j_{1:T})$ is the smoothed distribution at time step $k + 1$. The smoothed weights $w^j_{T|T,i}$ are just the forward weights $w^j_{T,i}$ and the end of the time interval. The smoothed weights at time $k < T$ can be obtained recursively:

$$
w^j_{k|T,i} = w^j_{k,i} \sum_{i=1}^{N} \frac{p(\tilde{\mathbf{z}}^j_{k+1,i} \mid \tilde{\mathbf{z}}^j_{k,i})\, w^j_{k+1,i}}{\sum_{l=1}^N p(\tilde{\mathbf{z}}^j_{k+1,i} \mid \tilde{\mathbf{z}}^j_{k,i})\, w^j_{k,l}}\,. \tag{4.9}
$$

for $k = T - 1, \ldots, 1$, $i = 1, \ldots, N$.

4.3 SMC-Based Nonstationary Discriminant Analysis

The particles and weights calculated with the particle smoother in the previous section, together with the information about the observational model in (4.3) can be used to approximate the class-conditional densities $p(\mathbf{x}_k \mid y = j)$ for each class j at each time k, $j = 0, 1, \ldots, c{-}1$ and $k = 1, \ldots, T$.

In this paper, we will assume a specific case of model (4.3):

$$
\mathbf{x}^j_{k,i} = \mathbf{z}^j_k + \mathbf{v}^j_{k,i}\,, \tag{4.10}
$$

where $\mathbf{v}^j_{k,i} \sim \mathcal{N}(\mathbf{0}, R^j)$ is i.i.d. zero-mean Gaussian noise with known covariance matrix R^j, for

$i = 0, 1, \ldots, n_k^j$, $j = 0, 1, \ldots, c-1$ and $k = 1, \ldots, T$.

From (4.10), the first and second moments of $p(\mathbf{x}_k \mid y = j)$ are given by

$$\begin{aligned} \boldsymbol{\mu}_k^j &= E[\mathbf{x}_k \mid y = j] = E[\mathbf{z}_k^j] \\ \Sigma_k^j &= \Sigma_{\mathbf{z}_k}^j + R^j \end{aligned} \tag{4.11}$$

for $j = 0, 1, \ldots, c-1$ and $k = 1, \ldots, T$.

Using the particles and weights $\{\tilde{\mathbf{z}}_{k,i}^j, w_{k|T,i}^j\}_{i=1}^N$ calculated previously leads to the following approximations

$$\begin{aligned} \hat{\boldsymbol{\mu}}_k^j &= \sum_{i=1}^N \tilde{\mathbf{z}}_{k,i}^j \, w_{k|T,i}^j \,, \\ \hat{\Sigma}_k^j &= \frac{N}{N-1} \sum_{i=1}^N w_{k|T,i}^j \left(\tilde{\mathbf{z}}_{k,i}^j - \hat{\boldsymbol{\mu}}_k^j\right) \left(\tilde{\mathbf{z}}_{k,i}^j - \hat{\boldsymbol{\mu}}_k^j\right)^T + R^j \end{aligned} \tag{4.12}$$

for $j = 0, 1, \ldots, c-1$ and $k = 1, \ldots, T$.

Quadratic Discriminant Analysis (QDA) [21] relies on general estimators $\hat{\boldsymbol{\mu}}^j, \hat{\Sigma}^j$, for $j = 0, \ldots, c-1$, of the first and second moments of the class-conditional densities to obtain a classifier. Given the *discriminants* (derived from the theory of optimal classification with Gaussian class-conditional densities)

$$D_{\mathrm{QDA}}^j(\mathbf{x}) = \log \pi^j - \frac{1}{2} \log |\hat{\Sigma}^j| - \frac{1}{2} (\mathbf{x} - \hat{\boldsymbol{\mu}}^j)^T (\hat{\Sigma}^j)^{-1} (\mathbf{x} - \hat{\boldsymbol{\mu}}^j), \tag{4.13}$$

where $\pi^j = P(y = j)$ are the class prior probabilities or estimates of the same, the general QDA classifier is given by

$$\psi_{\mathrm{QDA}}(\mathbf{x}) = \underset{j=0,1,\ldots,c-1}{\operatorname{argmax}} \, D_{\mathrm{QDA}}^j(\mathbf{x}) \,. \tag{4.14}$$

On the other hand, *Linear Discriminant Analysis* (LDA) [21] is based on the discriminants:

$$D_{\mathrm{LDA}}^j(\mathbf{x}) = \log \pi^j - \frac{1}{2} (\mathbf{x} - \hat{\boldsymbol{\mu}}^j)^T \hat{\Sigma}^{-1} (\mathbf{x} - \hat{\boldsymbol{\mu}}^j), \tag{4.15}$$

where the *pooled* covariance matrix estimator is given by

$$\hat{\Sigma} = \frac{\sum_{j=0}^{c-1}(n^j - 1)\hat{\Sigma}^j}{n-2} . \tag{4.16}$$

The LDA classifier is then given by

$$\psi_{\text{LDA}}(\mathbf{x}) = \underset{j=0,1,\ldots,c-1}{\operatorname{argmax}} D^j_{\text{LDA}}(\mathbf{x}) . \tag{4.17}$$

The QDA and LDA decision boundaries are composed of pieces of hyperquadric surfaces and hyperplanes; see [21] for more details.

The *naive* QDA and LDA classification rules ignore the nonstationarity in the data and plug in the usual sample means and sample covariance matrix based on all the data in the previous formulas.

By contrast, we are in position to define (SMC-based) nonstationary LDA and QDA (NSLDA and NSQDA, for short) classifiers at each time point k, which *also* use the entire data, but are adapted to the state of the evolving distribution at time k. This is done by defining discriminants

$$D^j_{\text{QDA},k}(\mathbf{x}) = \log \pi^j - \frac{1}{2}\log|\hat{\Sigma}^j_k| - \frac{1}{2}(\mathbf{x} - \hat{\boldsymbol{\mu}}_k{}^j)^T(\hat{\Sigma}^j_k)^{-1}(\mathbf{x} - \hat{\boldsymbol{\mu}}^j_k), \tag{4.18}$$

and

$$D^j_{\text{LDA},k}(\mathbf{x}) = \log \pi^j - \frac{1}{2}(\mathbf{x} - \hat{\boldsymbol{\mu}}^j_k)^T\hat{\Sigma}^{-1}_k(\mathbf{x} - \hat{\boldsymbol{\mu}}^j_k), \tag{4.19}$$

with

$$\hat{\Sigma}_k = \frac{\sum_{j=0}^{c-1}(n^j_k - 1)\hat{\Sigma}^j_k}{n_k - 2} , \tag{4.20}$$

and define the classifiers

$$\begin{aligned} \psi_{\text{QDA},k}(\mathbf{x}) &= \underset{j=0,1,\ldots,c-1}{\operatorname{argmax}} D^j_{\text{QDA},k}(\mathbf{x}) , \\ \psi_{\text{LDA},k}(\mathbf{x}) &= \underset{j=0,1,\ldots,c-1}{\operatorname{argmax}} D^j_{\text{LDA},k}(\mathbf{x}) . \end{aligned} \tag{4.21}$$

for $k = 1, \ldots, T$.

The entire process for the proposed SMC-based nonstationary discriminant analysis, for both the QDA and LDA cases, is summarized in Fig. 4.1.

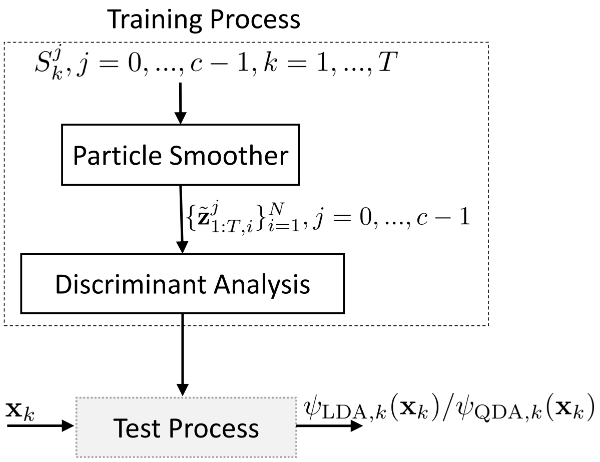

Training Process
$S_k^j, j = 0, ..., c-1, k = 1, ..., T$
Particle Smoother
$\{\tilde{\mathbf{z}}^j_{1:T,i}\}_{i=1}^N, j = 0, ..., c-1$
Discriminant Analysis
$\mathbf{x}_k$
Test Process
$\psi_{\mathrm{LDA},k}(\mathbf{x}_k)/\psi_{\mathrm{QDA},k}(\mathbf{x}_k)$

Figure 4.1: Proposed SMC-based Nonstationary Discriminant Analysis classification algorithm.

4.4 Simulation Results and Discussion

In this section, we showed several studies for non-linear drift numerical experiments when the non-linear drift system is either fully-known or partially know (i.e. some parameters in Eq. (4.1) and/or Eq. (4.3) are unknown). To estimate the states, in fully-known nonlinear system, we used the particle smoother (showed in Section 4.2.1 and 4.2.2). We plug the estimated states from the particle smoother into our proposed SMC-based nonstationary discriminant analysis (showed in Section 4.3), and obtained the average error estimates.

4.4.1 Fully-known Non-linear System

the following example of two-class, two-dimensional nonlinear centroid evolution (4.1) is adopted:

$$\begin{bmatrix} z_k^0(1) \\ z_k^0(2) \end{bmatrix} = \begin{bmatrix} z_{k-1}^0(1) + 2 \\ -0.5 z_{k-1}^0(2) + 0.4 z_{k-1}^0(1) * \sin\left(0.8 z_{k-1}^0(1)\right) + 2 \end{bmatrix}$$
$$\begin{bmatrix} z_k^1(1) \\ z_k^1(2) \end{bmatrix} = \begin{bmatrix} z_{k-1}^1(1) + 2 \\ -0.7 z_{k-1}^1(2) + \cos\left(z_{k-1}^1(1)\right) + 4 \end{bmatrix} \qquad (4.22)$$

for $k = 1, \ldots, T$. We therefore consider no transition noise in this example. The starting points for the evolution have the following Gaussian distributions: $z_0^0 \sim \mathcal{N}([1,2]^T, I_2)$ and $z_0^1 \sim \mathcal{N}([1,3]^T, I_2)$. In the first experiment, we compare the results of the SMC-based NSLDA and the previously described naive LDA. The time horizon is $T = 6$, with equal class probability $\pi^0 = \pi^1 = 1/2$. Four different scenarios are considered:

Case 1: $n_k^0 = n_k^1 = 10$, for $k = 1, \ldots, 6$, and low observation noise: $R^0 = R^1 = 0.01\, I_2$.

Case 2: (noisy data) Same sample sizes as in Case 1, and high observation noise: $R^0 = R^1 = 0.1\, I_2$.

Case 3: (missing data) Same sample sizes and observation noise as in Case 1, except that $n_4^0 = n_4^1 = 1$.

Case 4: (unbalanced data) $n_k^0 = 10$, $n_k^1 = 3$, for $k = 1, \ldots, 6$, and same observation noise as Case 1.

The classifier decision boundaries obtained in all four cases for both naive LDA and NSLDA are displayed in Fig. 4.2. The arrows specifies the direction of class 0. In case 1, the naive LDA is clearly underfitted, with a large rate of misclassified training data, while NSLDA can separate the data. This is due to the nonlinear evolution of the class conditional densities. In Case 2, under more noise, naive LDA continues to do a poor job, but NSLDA still performs reasonably well, given the confusion between the classes. The robustness against noise is a consequence of using the entire data set to fit the decision boundaries at each time point. In Case 3, we can see that the even with a single training point per class at $k = 4$, NSLDA is still able to find an accurate decision boundary

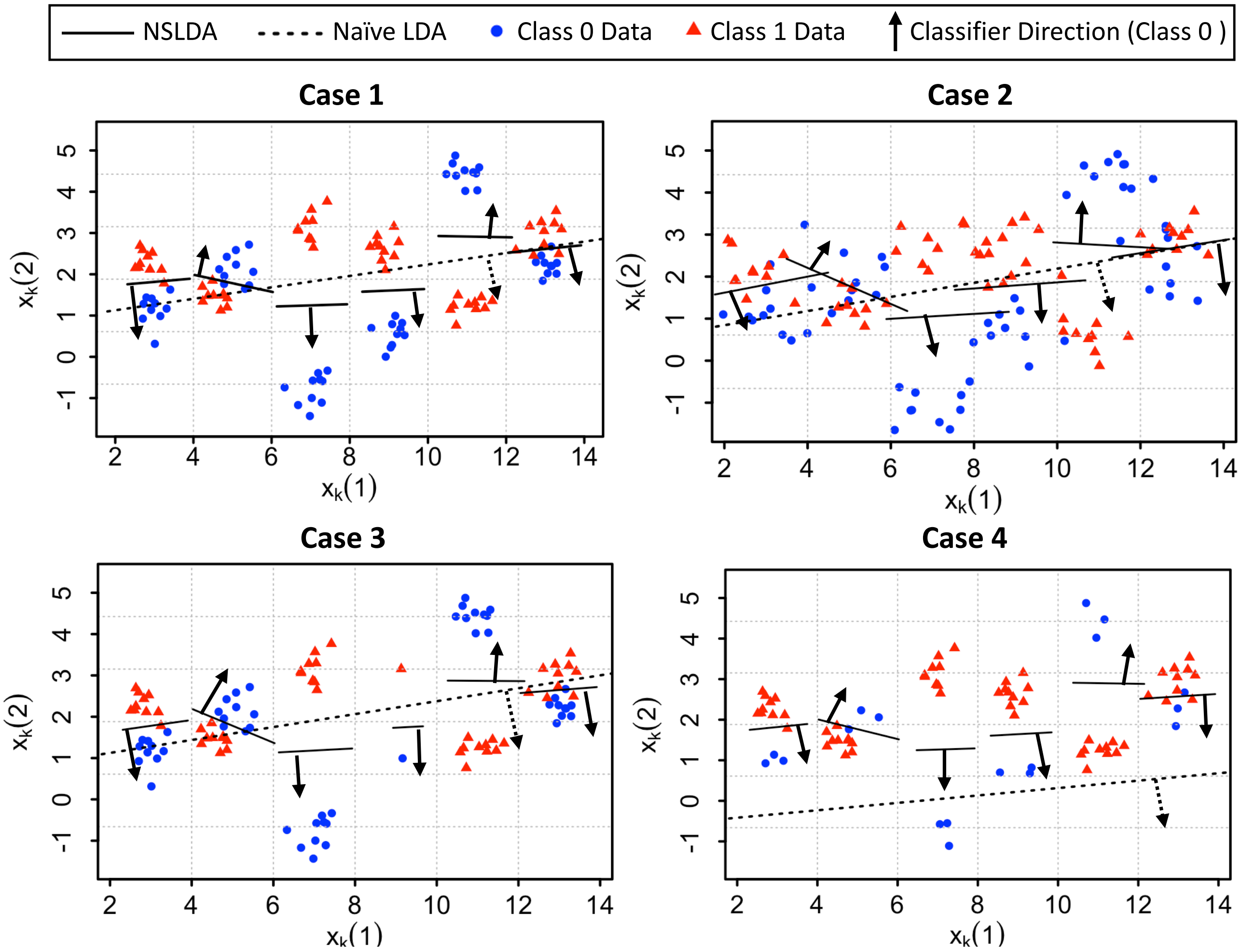


Figure 4.2: Decision boundaries produced by the naive LDA and SMC-based nonstationary LDA in four different cases of fully-known non-linear systems.

(compare this with the corresponding decision boundary at $k = 4$ in Case 1), which shows again the ability of NSLDA to "borrow" information from the other time points. Finally, in Case 4, one can see that the naive LDA decision boundary has shifted significantly (and erroneously), which is a well-known issue arising from unbalanced data. However, the unbalanced data does not affect the NSLDA results.

In the second experiment, the average classification error (estimated with large synthetic test sets) for naive LDA, NSLDA and a naive nonlinear radial-basis function Support Vector Machine (SVM) classification rule, which uses the entire data for training while ignoring nonstationarity.

The averages are based on 1000 runs, with $T = 4, 6, 8, 10, 12$, $n_k^0 = n_k^1 = 10$, for $k = 1, \ldots, 12$, and $R^0 = R^1 = 0.01\, I_2$. For nonstationary classifiers, the errors at each time points are averaged across the interval. The classification error rates for each of the three classification rules are plotted in Fig. 4.3. It can be seen that naive LDA displays a very poor accuracy, while the naive nonlinear SVM does a much better job, given its ability to fit the nonlinearly-separable data. However, nonstationary classifiers are the best classification rule, displaying a remarkably low classification error throughout.

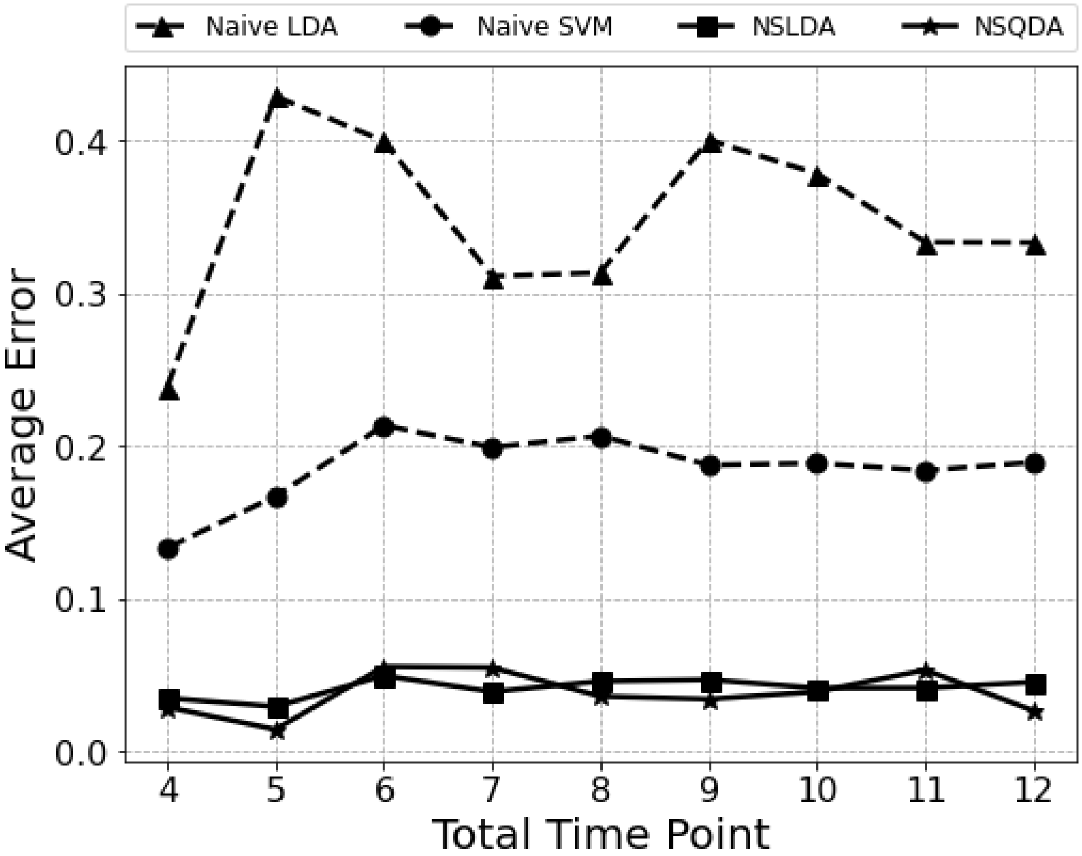


Figure 4.3: Average classification errors for naive LDA, naive nonlinear SVM, NSLDA and NSQDA classification rules versus time in fully-known non-linear systems. The dash line marked by ▲ represent the "naive" LDA error rate and dash line marked by • shows the "naive" support vector machine error rate. Two solid lines are for nonstationary classifier results, where ■ represents SMC-Based NSLDA ; ★ presents SMC-Based NSQDA.

5. CONCLUSIONS

This dissertation proposes methodologies for designing classifiers 1) under a restricted sampling method and 2) in nonstationary environment.

In Chapter 2, we focus on the bias of precision estimation under separate sampling. Accuracy and reproducibility in observational studies is critical to the progress of biomedicine, in particular, in the discovery of reliable biomarkers for disease diagnosis and prognosis. In this study, theoretical results confirmed by numerical experiments show that the usual estimator of precision can be severely biased under the typical separate sampling scenario in observational case-control studies. This will be true especially if the true disease prevalence differs significantly from the apparent prevalence in the data. If knowledge of the true disease prevalence is available, or can even be approximately ascertained, then it can be used to define a modified precision estimator, which is nearly unbiased at moderate sample sizes. In all the results using real data sets, we observed that the usual precision estimator produces values that are larger, i.e., more optimistic, than the modified one using the true prevalence, which agrees with the results obtained with the synthetic data. Absence of knowledge about the true prevalence means simply that the precision cannot be reliably estimated in observational case-control studies and its use should be discouraged. Finally, we note that in our experiments we considered the case where the prevalence is between 0.1 and 0.9, not without reason. If the prevalence is significantly under 0.1, as is the case in some rare diseases, then neither the precision, nor in fact the classification error, should be used as a criterion of performance, but rather the sensitivity and specificity need to be considered separately — otherwise, a large precision and small classification error can be achieved by biasing the classification rule to produce false positive rates close to zero while ignoring the false negative rate.

The rest of this dissertation, we discuss how state-space models work for nonstationary discriminant analysis. In Chapter 3, we showed nonstationary data in linear drift model. To address this problem, we first considered the simple case where the class-conditional densities evolve linearly under Gaussian observation noise, and applied standard methods of linear filtering to obtain a nonstationary linear discriminant analysis (NSLDA) classification rule. Later on, we address the case where parameters in linear state model are unknown, and/or time points of the measurements are unknown, by using a combination of Expectation Maximization (EM) or Gaussian mixture models (GMM) and classical Kalman Smoother (KS) to estimate the time labels, and then using these values with a Kalman smoother (KS) for estimating the unknown parameters. We are working on an extension of linear state-space model approach for nonstationary discriminant analysis, by Bayesian method. In the model, there are uncertain noise statistics, and a Bayesian robust Kalman smoothing framework is used to estimate the noise statistics and states of the model.

In Chapter 4, we proposed a general nonlinear, non-Gaussian model in a fully-known system for nonstationary data, which allowed us to derive non-stationary discriminant analysis classification rules capable of producing classifiers tuned to the state of the distribution at each time point, while borrowing information from all time points. The proposed framework uses the sequential Monte- Carlo (SMC) estimation of the class conditional density centroids at all time points using all available data. The high accuracy of the proposed nonstationary classification rules and its ability in handling missing or unbalanced data is demonstrated in a series of numerical experiments.

REFERENCES

[1] B. Gnedenko and A. Kolmogorov, *Independent Random Variables*. Cambridge, Massachusetts: Addison-Wesley, 1954.

[2] P. Gänssler and W. Stute, "Empirical processes: a survey of results for independent and identically distributed random variables," *The Annals of Probability*, pp. 193–243, 1979.

[3] P. Martin-Löf, "The definition of random sequences," *Information and control*, vol. 9, no. 6, pp. 602–619, 1966.

[4] S. M. Pincus, "Approximate entropy as a measure of system complexity.," *Proceedings of the National Academy of Sciences*, vol. 88, no. 6, pp. 2297–2301, 1991.

[5] T. W. Anderson, T. W. Anderson, T. W. Anderson, T. W. Anderson, and E.-U. Mathématicien, *An introduction to multivariate statistical analysis*, vol. 2. Wiley New York, 1958.

[6] T. Deisboeck and J. Y. Kresh, *Complex systems science in biomedicine*. Springer Science & Business Media, 2007.

[7] N. E. Breslow, "Statistics in epidemiology: the case-control study," *Journal of the American Statistical Association*, vol. 91, no. 433, pp. 14–28, 1996.

[8] P. M. Gy, "Introduction to the theory of sampling i. heterogeneity of a population of uncorrelated units," *TrAC Trends in Analytical Chemistry*, vol. 14, no. 2, pp. 67–76, 1995.

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

[10] M. Schena, D. Shalon, R. W. Davis, and P. O. Brown, "Quantitative monitoring of gene expression patterns with a complementary dna microarray," *Science*, vol. 270, no. 5235, pp. 467–470, 1995.

[11] D. J. Lockhart, H. Dong, M. C. Byrne, M. T. Follettie, M. V. Gallo, M. S. Chee, M. Mittmann, C. Wang, M. Kobayashi, H. Norton, *et al.*, "Expression monitoring by hybridization to high-density oligonucleotide arrays," *Nature biotechnology*, vol. 14, no. 13, p. 1675, 1996.

[12] A. Mortazavi, B. Williams, K. McCue, L. Schaeffer, and B. Wold, "Mapping and quantifying mammalian transcriptomes by rna-seq," *Nature Methods*, vol. 5, no. 7, pp. 621–628, 2008.

[13] R. Aebersold and M. Mann, "Mass spectrometry-based proteomics," *Nature*, vol. 422, no. 6928, pp. 198–207, 2003.

[14] U. Braga-Neto and E. Dougherty, "Is cross-validation valid for microarray classification?," *Bioinformatics*, vol. 20, no. 3, pp. 374–380, 2004.

[15] U. Braga-Neto and E. Dougherty, *Error Estimation for Pattern Recognition.* New York: Wiley, 2015.

[16] M.-S. Ong, F. Magrabi, and E. Coiera, "Automated categorisation of clinical incident reports using statistical text classification," *Quality and Safety in Health Care*, vol. 19, no. 6, pp. e55–e55, 2010.

[17] H. X. Dang and C. B. Lawrence, "Allerdictor: fast allergen prediction using text classification techniques," *Bioinformatics*, vol. 30, no. 8, pp. 1120–1128, 2014.

[18] S. Hassanpour, C. P. Langlotz, T. J. Amrhein, N. T. Befera, and M. P. Lungren, "Performance of a machine learning classifier of knee mri reports in two large academic radiology practices: A tool to estimate diagnostic yield," *American Journal of Roentgenology*, pp. 1–4, 2017.

[19] U. Braga-Neto, A. Zollanvari, and E. Dougherty, "Cross-validation under separate sampling: Strong bias and how to correct it," *Bioinformatics*, vol. 30, no. 23, 2014.

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

[21] R. Duda and P. Hart, *Pattern Classification.* New York: John Wiley & Sons, 2nd ed., 2001.

[22] C. Bishop, *Neural Networks for Pattern Recognition*. New York: Oxford University Press, 1995.

[23] G. Krempl and V. Hofer, "Classification in presence of drift and latency," in *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pp. 596–603, IEEE, 2011.

[24] M. G. Kelly, D. J. Hand, and N. M. Adams, "The impact of changing populations on classifier performance," in *Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 367–371, ACM, 1999.

[25] G. Hulten, L. Spencer, and P. Domingos, "Mining time-changing data streams," in *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 97–106, ACM, 2001.

[26] M. Sugiyama and M. Kawanabe, *Machine learning in non-stationary environments: Introduction to covariate shift adaptation*. MIT press, 2012.

[27] G. Ditzler, M. Roveri, C. Alippi, and R. Polikar, "Learning in nonstationary environments: A survey," *IEEE Computational Intelligence Magazine*, vol. 10, no. 4, pp. 12–25, 2015.

[28] S. S. Haykin *et al.*, *Kalman filtering and neural networks*. Wiley Online Library, 2001.

[29] G. McLachlan, *Discriminant analysis and statistical pattern recognition*, vol. 544. John Wiley & Sons, 2004.

[30] U. Braga-Neto, *Parametric Classification*, pp. 67–88. Springer International Publishing, 2020.

[31] H. E. Rauch, F. Tung, C. T. Striebel, *et al.*, "Maximum likelihood estimates of linear dynamic systems," *AIAA journal*, vol. 3, no. 8, pp. 1445–1450, 1965.

[32] Z. Ghahramani and G. E. Hinton, "Parameter estimation for linear dynamical systems," tech. rep., Technical Report CRG-TR-96-2, University of Totronto, Dept. of Computer Science, 1996.

[33] M. J. Beal, *Variational algorithms for approximate Bayesian inference*. University of London London, 2003.

[34] R. H. Shumway and D. S. Stoffer, “An approach to time series smoothing and forecasting using the em algorithm,” *Journal of time series analysis*, vol. 3, no. 4, pp. 253–264, 1982.

[35] S. Xie, M. Imani, E. R. Dougherty, and U. M. Braga-Neto, “Nonstationary linear discriminant analysis,” in *Signals, Systems, and Computers, 2017 51st Asilomar Conference on*, pp. 161–165, IEEE, 2017.

[36] D. Arnaud, N. de Freitas, and N. Gordon, *Sequential Monte Carlo methods in practice*. New York: Springer Science+ Business Media, Inc. LLC, 2001.

[37] N. Kantas, A. Doucet, S. S. Singh, J. Maciejowski, N. Chopin, *et al.*, “On particle methods for parameter estimation in state-space models,” *Statistical science*, vol. 30, no. 3, pp. 328–351, 2015.

[38] R. O. Duda, P. E. Hart, D. G. Stork, *et al.*, “Pattern classification. 2nd,” *Edition. New York*, vol. 55, 2001.

[39] R. Hewett and P. Kijsanayothin, “Tumor classification ranking from microarray data,” *BMC genomics*, vol. 9, no. 2, p. S21, 2008.

[40] H. N, N. AM, K. M, M. D, B. K, A. SF, K. CL, Y. M, R. J, T. Z, M. A, L. DR, C. HS, F. EJ, and C. K. (eds), “Seer cancer statistics review, 1975-2013, national cancer institute,” 2016. Bethesda, MD, http://seer.cancer.gov/csr/1975_2013/, based on November 2015 SEER data submission, posted to the SEER web site, April 2016.

[41] H. Asri, H. Mousannif, H. Al Moatassime, and T. Noel, “Using machine learning algorithms for breast cancer risk prediction and diagnosis,” *Procedia Computer Science*, vol. 83, pp. 1064–1069, 2016.

[42] D. Dua and C. Graff, “UCI machine learning repository,” 2017.

[43] W. H. Wolberg and O. L. Mangasarian, “Multisurface method of pattern separation for medical diagnosis applied to breast cytology.,” *Proceedings of the national academy of sciences*, vol. 87, no. 23, pp. 9193–9196, 1990.

[44] S. S. Shajahaan, S. Shanthi, and V. ManoChitra, “Application of data mining techniques to model breast cancer data,” *International Journal of Emerging Technology and Advanced Engineering*, vol. 3, no. 11, pp. 362–369, 2013.

[45] M. F. Akay, “Support vector machines combined with feature selection for breast cancer diagnosis,” *Expert systems with applications*, vol. 36, no. 2, pp. 3240–3247, 2009.

[46] L. Wilkins, *Interpreting Signs and Symptoms*. LWW medical book collection, Lippincott Williams & Wilkins, 2007.

[47] B. Ramana, M.S.P.Babu, and N.B.Venkateswarlu, “A critical study of selected classification algorithms for liver disease diagnosis,” *Journal of Database Management Systems*, vol. 3, no. 2, pp. 101–114, 2011.

[48] Z. Younossi, M. Stepanova, M. Afendy, Y. Fang, Y. Younossi, H. Mir, and M. Srishord, “Changes in the prevalence of the most common causes of chronic liver diseases in the united states from 1988 to 2008,” *Clinical Gastroenterology and Hepatology*, vol. 9, no. 6, pp. 524–530, 2011.

[49] G. Holmes, A. Donkin, and I. Witten, “Weka: A machine learning workbench,” 1994. Working paper 94/9, Hamilton, New Zealand: University of Waikato, Department of Computer Science.

[50] N. Friedman, D. Geiger, and M. Goldszmidt, “Bayesian network classifiers,” *Machine Learning*, vol. 29, pp. 131–163, 1997.

[51] T. Dietterich, “An experimental comparison of three methods for constructing ensembles of decision trees: Bagging, boosting, and randomization,” *Machine Learning*, vol. 40, no. 2, pp. 139–157, 2000.

[52] V. Digalakis, J. R. Rohlicek, and M. Ostendorf, “Ml estimation of a stochastic linear system with the em algorithm and its application to speech recognition,” *IEEE Transactions on speech and audio processing*, vol. 1, no. 4, pp. 431–442, 1993.

[53] C. A. Bouman, M. Shapiro, G. Cook, C. B. Atkins, and H. Cheng, “Cluster: An unsupervised algorithm for modeling gaussian mixtures.”

[54] M. Hürzeler and H. R. Künsch, “Monte Carlo approximations for general state-space models,” *Journal of Computational and graphical Statistics*, vol. 7, no. 2, pp. 175–193, 1998.

[55] G. Kitagawa, “Monte Carlo filter and smoother for non-gaussian nonlinear state space models,” *Journal of computational and graphical statistics*, vol. 5, no. 1, pp. 1–25, 1996.

[56] M. K. Pitt and N. Shephard, “Filtering via simulation: Auxiliary particle filters,” *Journal of the American statistical association*, vol. 94, no. 446, pp. 590–599, 1999.

APPENDIX A

ASYMPTOTIC APPROXIMATION FOR THE EXPECTATION OF A RATIO OF TWO RANDOM VARIABLES

Here we derive the asymptotic approximation for the expectation of a ratio of two random variables W and Z:

$$E\left[\frac{W}{Z}\right] \approx \frac{E[W]}{E[Z]}\,. \tag{A.1}$$

Proof:

If $f : \mathbb{R}^2 \to \mathbb{R}$ is infinitely differentiable at point (a, b) then it can be expanded by a bivariate Taylor series around (a, b) as:

$$\begin{aligned} f(x,y) \;=\; f(a,b) \;+\; \frac{\partial f(a,b)}{\partial x}\,(x-a) \;+\; \frac{\partial f(a,b)}{\partial y}\,(y-b) \\ +\; \text{second and higher order terms in } x-a \text{ and } y-b\,. \end{aligned} \tag{A.2}$$

Now let X_n and Y_n be sequences of random variables with means μ_X and μ_Y, with $\mu_Y \neq 0$. The ratio x/y is infinitely differentiable at (a, b) if $b \neq 0$, therefore we can apply the previous result and get

$$\begin{aligned} \frac{X_n}{Y_n} \;=\; \frac{\mu_X}{\mu_Y} \;+\; \frac{1}{\mu_Y}\,(X_n - \mu_X) \;-\; \frac{\mu_X}{\mu_Y^2}\,(Y_n - \mu_Y) \\ +\; \text{second and higher order terms in } X_n{-}\mu_X \text{ and } Y_n{-}\mu_Y\,. \end{aligned} \tag{A.3}$$

Taking expectations on both sides gives:

$$\begin{aligned} E\left[\frac{X_n}{Y_n}\right] \;=\; \frac{\mu_X}{\mu_Y} + E[\text{second and higher order terms} \\ \text{in } X_n{-}\mu_X \text{ and } Y_n{-}\mu_Y]\,. \end{aligned} \tag{A.4}$$

Except in pathological cases involving heavy-tailed distributions, the remainder in the previous equation becomes negligible as $X_n \to \mu_X$ and $Y_n \to \mu_Y$ in probability. Therefore, we write

$$E\left[\frac{X}{Y}\right] \approx \frac{E[X]}{E[Y]}, \tag{A.5}$$

as long as X and Y are around $E[X]$ and $E[Y]$ respectively (i.e., Var$[X]$ and Var$[Y]$ are small).