DISSERTATION

A NOVEL APPROACH TO STATISTICAL PROBLEMS WITHOUT IDENTIFIABILITY

Submitted by

Addison D. Adams

Department of Statistics

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Summer 2024

Doctoral Committee:

Advisor: Haonan Wang
Co-Advisor: Tianjian Zhou

Piotr Kokoszka
Ben Shaby
Indrakshi Ray

ABSTRACT

A NOVEL APPROACH TO STATISTICAL PROBLEMS WITHOUT IDENTIFIABILITY

In this dissertation, we propose novel approaches to random coefficient regression (RCR) and the recovery of mixing distributions under nonidentifiable scenarios.

The RCR model is an extension of the classical linear regression model that accounts for individual variation by treating the regression coefficients as random variables. A major interest lies in the estimation of the joint probability distribution of these random coefficients based on the observable samples of the outcome variable evaluated for different values of the explanatory variables. In Chapter 2, we consider fixed-design RCR models, under which the coefficient distribution is not identifiable. To tackle the challenges of nonidentifiability, we consider an equivalence class, in which each element is a plausible coefficient distribution that, for each value of the explanatory variables, yields the same distribution for the outcome variable. In particular, we formulate the approximations of the coefficient distributions as a collection of stochastic inverse problems, allowing for a more flexible nonparametric approach with minimal assumptions. An iterative approach is proposed to approximate the elements by incorporating an initial guess of a solution called the global ansatz. We further study its convergence and demonstrate its performance through simulation studies. The proposed approach is applied to a real data set from an acupuncture clinical trial.

In Chapter 3, we consider the problem of recovering a mixing distribution, given a component distribution family and observations from a compound distribution. Most existing methods are restricted in scope in that they are developed for certain component distribution families or continuity structures of mixing distributions. We propose a new, flexible nonparametric approach with minimal assumptions. Our proposed method iteratively steps closer to the desired mixing distribution, starting from a user-specified distribution, and we further establish its convergence

properties. Simulation studies are conducted to examine the performance of our proposed method. In addition, we demonstrate the utility of our proposed method through its application to two sets of real-world data, including prostate cancer data and Shakespeare's canon word count.

ACKNOWLEDGEMENTS

I express my gratitude to my advisor Haonan Wang for his encouragement, patience, and dedication to my success. He provided me with valuable feedback and insights. I value the time we spent discussing research and life in general. I would like to thank Tianjian Zhou, my co-advisor, for being such a joy to work with. I appreciate his hard work ethic, intelligence, and dedication to my success.

I thank the members of my committee, Piotr Kokoszka, Ben Shaby, and Indrakshi Ray for their time, feedback, and encouragement.

To my fellow graduate students, and in particular, to Troy Wixson and Gray Stanton, I express my gratitude. It was enjoyable and helpful to discuss our different research subjects and to bounce ideas off each other. I value our developed friendship that helped make the journey of graduate school a pleasure.

To my parents, Duane and Natalie Adams, and to my parents-in-law, Chuck and Tina Schwab, I give my sincere thanks. Thank you for your advice, assistance, and support. Your sacrifices and generosity helped make this journey possible.

To my daughters, thank you for helping me to see the brighter side of life and to stay positive. You elevate my perspective and help me to prioritize what matters most.

To Kelsey, I express my sincere gratitude and love. The pursuit of this dream would simply not have been possible without you. I treasure your faith in me and the confidence you inspire in me. Your dedication, sacrifices, and encouragement were and are invaluable. I am grateful to have shared the graduate school experience with you, and I am eager to continue the rest of life with you. I love you.

DEDICATION

I dedicate this dissertation to Kelsey.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

Identifiability and estimation are intrinsically linked, where identifiability is a crucial concept to estimation theory. Without identifiability, inference may be problematic (Koopmans and Reiersol, 1950; Rothenberg, 1971). Koopmans and Reiersol (1950) emphasized that nonidentifiability arises in a variety of disciplines. In this dissertation we present novel nonparametric approaches to models that often suffer from nonidentifiability and propose a perspective to handle nonidentifiability. In particular, we focus on random coefficient regression (RCR) and the mixing distribution recovery problem. We begin by providing general background for estimation theory, identifiable models, and nonparametric statistics, before introducing RCR models and the mixing distribution recovery problem.

Estimation theory in statistics has been well-studied, where the goal is to approximate some unknown quantity from a sample of data drawn from a population. In classical inference and decision theory, it is assumed that the observed data come from a distribution which belongs to a class of distributions, $\mathcal{P} = \{P(\cdot|\theta) : \theta \in \mathcal{T}\}$, indexed by a parameter θ, and where estimation is to determine a plausible value for θ from the data. This is known as point estimation (Lehmann and Casella, 1998; Casella and Berger, 2002). More generally, one may specify a subset of $\mathcal{T}$ of plausible values for θ, known as interval estimation. This would include the well-known confidence intervals, used in frequentist settings, and credible intervals, used in Bayesian analysis.

The set of distributions $\mathcal{P}$ is known as a parametric family. However, often no parametric family can be reasonably assumed. The term parameter may then take a more general meaning. That is, the parameter is not only a value or vector that specifies a distribution in $\mathcal{P}$, but a characteristic of the population. Examples of parameters are quantiles, means, modes, or probability distributions. Without assuming a family of distributions, this type of estimation, or inference more generally, is known as nonparametric; see Sprent and Smeeton (2007) and Kendall et al. (1994) for discussions on the classification of nonparametric.

Commonly, the parameter of interest, is the probability distribution of the population itself, such as in kernel density estimation (Rosenblatt, 1956; Parzen, 1962; Bowman, 1984; Silverman, 1986). In this work, we investigate the nonparametric approximation of probability distributions, under two models that will be described presently. That is, our parameter of interest is a probability distribution, particularly, a distribution of unobserved variables. Admittedly, the term nonparametric may have different meanings, so to be precise, we take nonparametric approximation of a probability distribution to mean that there is no assumption that the unknown distribution belongs to a parametric family.

A careful reader may notice that, in describing our focus, we omitted the terms "estimate" and "estimation". This is because we feel the term estimation should be reserved for identifiable models, and the models we consider in Chapters 2 and 3 suffer from nonidentifiability. In general, if there is a one-to-one correspondence between the parameter of interest and the distribution of the observable data then the parameter, or equivalently the distribution of the data, is said to be identifiable (Koopmans and Reiersol, 1950; Rothenberg, 1971; Durlauf et al., 2020). Note, the term identifiable is also applied to the model when the parameter is identifiable. This is because the identifiability of a parameter is not a consequence of the form of an estimator, but rather of the model and assumptions (Casella and Berger, 2002). For example, returning to the classical case, if there is a one-to-one correspondence between θ and $P(\cdot|\theta) \in \mathcal{P}$, then θ, or $\mathcal{P}$, is said to be identifiable.

Identifiability, or the lack thereof, has been dealt with in a few ways. Often, identifiability can be ensured by reparameterization, e.g., using a cell-means model for ANOVA rather than the overparameterized model (Casella and Berger, 2002). In nonparametric estimation, identifiability is occasionally ensured by some independence assumption (Beran and Hall, 1992; Durlauf et al., 2020). Generally speaking, assumptions, many of which are unverifiable, are imposed on a model so that the parameter is identifiable; see Beran et al. (1996), Hoderlein et al. (2010), and Teicher (1963) for a few examples. Although some assumptions are unverifiable, they may be justifiable, e.g., scientific or domain knowledge may justify an assumption. However, if the only justifica-

tion of an assumption is to ensure identifiability, the resulting inference or conclusion does not inspire confidence. Neath and Samaniego (1997) discussed the advantages and disadvantages of using Bayesian methods for nonidentifiable parameters. Wechsler et al. (2013) illustrates using a Bayesian method in a simple nonidentifiable model. Similar to unverifiable assumptions, if there is justification of a prior, which influences the resulting inference, then it may be advantageous to use. If the prior is not justifiable, then the concluding inference may be met with high levels of skepticism.

In both models investigated in this work, we deal with nonidentifiability by considering approximating distributions within an equivalence class. The equivalence classes are defined as the sets of distributions of the unobserved variables that map to the same distribution, or set of distributions, of the observed data. In both Chapters 2 and 3, we propose methods for approximating elements within the equivalence class, each starting with a user-specified distribution.

Here, we describe the modeling scenarios within this dissertation. We begin with the RCR models of Chapter 2. Consider the following RCR model

$$Y_i = A_i + B_i x_i$$

for $i = 1, 2, \ldots, n$. We consider Y_i and x_i to be the observed response variable of interest and the observed design variable for subject i, respectively. The pairs (A_i, B_i) are unobserved random coefficients drawn for an unknown joint coefficient distribution $F_{A,B}$.

The RCR model is an extension of the linear regression model, but where (A_i, B_i) is considered random, which accounts for the heterogeneity of subjects. For example, if Y_i is the observed response of subject i to treatment x_i, B_i is the individual treatment effect for subject i.

The goal in the RCR model is to recover the coefficient distribution $F_{A,B}$ from the observed data (Y_i, x_i) without assuming $F_{A,B}$ comes from a particular parametric family. In recovering the coefficient distribution, many methods in the literature impose assumptions to ensure the identifiability of $F_{A,B}$. For example, Beran and Hall (1992) assumed A_i and B_i are mutually independent for each i. Others assumed the design variable to be random with a distribution that has support

over the whole real line (Beran et al., 1996; Hoderlein et al., 2010). Of course, in many scenarios, such as in randomized clinical trials, where $x_i = 1$ for treatment and 0 for control, assuming the distribution for x_i has full support is unreasonable.

In contrast, we consider the design variables to come from a finite discrete lattice $\mathcal{X} = \{x_1^*, x_2^*, \ldots, x_J^*\}$ and consider each x_i to be fixed. We prove that under such a case, the coefficient distribution is nonidentifiable.

From a population perspective, that is assuming the distributions of Y for each x are known, solving for a coefficient distribution $F_{A,B}$ can be framed as a collection of *stochastic inverse problems* (SIP; Breidt et al., 2011; Butler et al., 2012, 2014, 2018). In essence, the SIP is to find a distribution whose push-forward measure for a given map is equal to a specified distribution. Specifically, in the context of RCR, let $F_{A,B}$ be a measure on $(\mathbb{R}^2, \mathcal{B}_{\mathbb{R}^2})$ and $Q_x : \mathbb{R}^2 \to \mathbb{R}$ be a measurable map defined as $Q_x(a, b) = a + bx$. Then Q_x induces a measure F_{Y_x} on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ through

$$F_{Y_x}(C) = F_{A,B}(\{(a, b) \in \mathbb{R}^2 : Q_x(a, b) = a + bx \in C\})$$

for all $C \in \mathcal{B}_{\mathbb{R}}$. Calculating the image measure of $F_{A,B}$ for map Q_x, is a special case of a *stochastic forward problem* (SFP). Then, given the image measure F_{Y_x} and map Q_x, the SIP is to find a distribution $\widetilde{F}_{A,B}$ on $(\mathbb{R}^2, \mathcal{B}_{\mathbb{R}^2})$ such that

$$F_{Y_x}(C) = \widetilde{F}_{A,B}(\{(a, b) \in \mathbb{R}^2 : Q_x(a, b) = a + bx \in C\})$$

for all $C \in \mathcal{B}_{\mathbb{R}}$.

Then, recovering coefficient distributions $\widetilde{F}_{A,B}$ can be viewed as a collection of SIPs in the following manner. Let

$$\mathcal{E}_x = \{\widetilde{F}_{A,B} \text{ on } (\mathbb{R}^2, \mathcal{B}_{\mathbb{R}^2}) : Q_x \widetilde{F}_{A,B} = F_{Y_x}\}$$

where $Q_x\widetilde{F}_{A,B}$ denotes the image measure of $\widetilde{F}_{A,B}$ for map Q_x. We seek a distribution $\widetilde{F}_{A,B}$ that simultaneously belongs to each $\mathcal{E}_x$, i.e., $\widetilde{F}_{A,B} \in \mathcal{E} = \bigcap_{x\in\mathcal{X}} \mathcal{E}_x$. The collection of SIPs is illustrated in Figure 1.1.

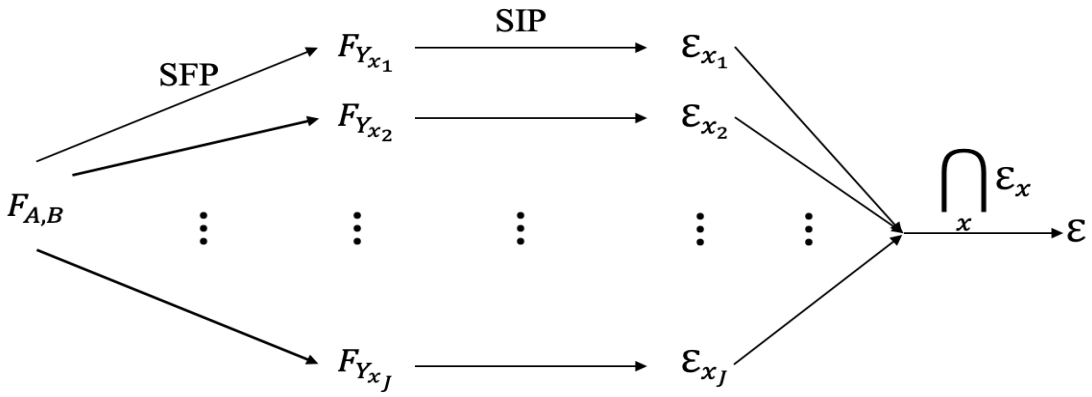

Figure 1.1: An illustration of the collection of SIPs that characterizes the RCR model from a population view.

In Chapter 2, an iterative method is proposed to approximate elements in $\mathcal{E}$ by sequentially solving each SIP, indexed by x, by using the disintegration theorem (Chang and Pollard, 1997; Yang, 2018). The set solutions to each SIP are averaged to step closer to $\mathcal{E}$. This process is initiated by a user-specified distribution on $(\mathbb{R}^2, \mathcal{B}_{\mathbb{R}^2})$, called a *global ansatz*. By changing the global ansatz, different elements in $\mathcal{E}$ can be approximated.

In Chapter 3, we shift focus to the mixing distribution recovery problem. Similar to the RCR models, the goal is to approximate distributions of unobserved variables. Suppose $(X_i, \Theta_i), i = 1, 2, \ldots n$, to be a random sample from a joint distribution $F_{X,\Theta}$ on $(\mathcal{X} \times \mathcal{T}, \mathcal{B}_{\mathcal{X}} \otimes \mathcal{B}_{\mathcal{T}})$. The joint distribution is partially known in the sense that we assume the family of conditional distributions $\mathcal{P} = \{P(\cdot|\theta) : \theta \in \mathcal{T}\}$, called the *component distribution family*, is known. For each i, X_i is observed, but Θ_i is not. That is, we can consider X_i, for $i = 1, 2, \ldots, n$, to be a random sample

from the *compound distribution* F. The compound distribution is defined as

$$F(A) = \int_{\mathcal{T}} P(A|\theta) \mathrm{d}G(A)$$

for all $A \in \mathcal{B}_{\mathcal{X}}$, and where G is the Θ-marginal of $F_{X,\Theta}$, known as the *mixing distribution*. It is desired, from the sample $\{X_i\}$, to recover the unknown mixing distribution G.

Although the mixing distribution recovery problem has many applications, it is particularly applicable to empirical Bayesian methods, where the observed data is used to estimate the prior (Robbins, 1964; Efron, 2010, 2014, 2016; Zhai and Jiang, 2023). Under the empirical Bayesian framework, the mixing distribution is the prior and the component distribution family is the likelihood.

Teicher (1960, 1963); Tallis (1969); Yakowitz (1969); Atienza et al. (2006) studied the identifiability of the mixing distribution. In Section 3.2, the identifiability is summarized. Here, it is sufficient for the reader to know that the mixing distribution G may or may not be identifiable, depending on the component distribution family $\mathcal{P}$.

From the population view, the recovery of the mixing distribution is to solve for G, given the compound distribution F and component distribution $\mathcal{P}$. We define the solution set to the problem as

$$\mathcal{E}_{F,\mathcal{P}} = \left\{ \widetilde{G} \text{ on } (\mathcal{T}, \mathcal{B}_{\mathcal{T}}) : F(A) = \int_{\mathcal{T}} P(A|\theta) \mathrm{d}\widetilde{G}(\theta) \text{ for all } A \in \mathcal{B}_{\mathcal{X}} \right\}.$$

To find distributions in $\mathcal{E}_{F,\mathcal{P}}$, we propose another iterative method, in which disintegration and integration are used to step closer to distributions within the solution set. Again, the iterative method is initiated by a distribution, which we call an ansatz, and the ansatz can be changed to approximated different elements in $\mathcal{E}_{F,\mathcal{P}}$ when more than one element exists.

Although the mixing distribution may be identifiable for some component distribution families $\mathcal{P}$, the mixing distribution recovery problem may still suffer from *instability*. That is, small differences in the compound distribution may lead to large differences in the resulting mixing distributions, making the recovery of the mixing distribution challenging. A careful choice of an

ansatz may help to stabilize the recovery. Without knowing which ansatz to choose, however, a practitioner may leverage the flexibility of the proposed method to stabilize the recovery in other ways; see Section 3.4 for details.

For both methods in Chapters 2 and 3, the convergence properties of each iterative method are studied. Empirical versions of the methods are suggested and validated through simulation studies which are compared to existing methods. Specifically, for the RCR models, the image measures for each map Q_x are estimated from the data (Y_i, x_i), and the resulting set of estimated distributions $\{\hat{F}_{Y_x} : x \in \mathcal{X}\}$ are used in place of $\{F_{Y_x} : x \in \mathcal{X}\}$. Similarly, for the mixing distribution recovery, the sample $\{X_i\}$ is used to estimate the compound distribution. The estimate $\hat{F}$ is used in place of the unknown compound distribution F. Finally, for each modeling scenario, both methods are demonstrated on real-life applications. In Chapter 4 we summarize the connection of both methods and discuss future work for nonidentifiable methods.

Chapter 2

A New Approach to Random Coefficient Regression without Identifiability

2.1 Introduction

We consider a random coefficient regression (RCR) model of the form

$$Y_i = A_i + B_i x_i. \tag{2.1}$$

For subject i, x_i and Y_i represent the explanatory and outcome variables of interest, respectively. Assume that x_i is a design point from $\mathcal{X} = \{x_j^*, j = 1, \ldots, J\}$. The relationship between Y_i and x_i is modeled through a linear relationship with unobserved random regression coefficients A_i and B_i. We further assume that $\{(A_i, B_i) : i = 1, \ldots, n\}$ is independently drawn from a joint distribution $F_{A,B}$, which does not vary across x_i. Let (x_i, Y_i) be an independent sample of observations. The goal is to approximate a coefficient distribution from which, for each distinct design point x, the induced distribution for Y, the *output distribution* and denoted by F_{Y_x}, can plausibly generate the sample. In this chapter, Y_i, x_i, A_i, and B_i are scalars, and we will suppress the subscript i for ease of discussion.

Applications of RCR models with such a setup are abundant in medical and econometric studies. For instance, the RCR model has been used to quantify treatment effects in randomized clinical trials, characterize the dose-response relationships in dose-response studies, and study time trends in pooled cross-section analyses. In those examples, the design variables are the treatment assignment indicator, dose level, and time.

RCR models are also known as stochastic coefficient regression models (Johnson, 1977) or regression models of the second kind (Fisk, 1967; Nelder, 1968). The use of random rather than fixed coefficient regression models has been advocated by many authors. For example, Hildreth

and Houck (1968) and Beran and Hall (1992) commented that RCR models account for unobserved individual heterogeneity and incorporate heteroscedastic errors. Fixed coefficient regression models are a special case of RCR models where $F_{A,B}$ is taken to be degenerate.

In general, the coefficient distribution in RCR models is *nonidentifiable*. That is, there may exist multiple coefficient distributions that will induce the same distributions for the output variable for all available design points in $\mathcal{X}$. However, certain characteristics are identifiable. For instance, Hermann and Holzmann (2021) showed that the mixed moments of regression coefficients are identifiable when x takes on a sufficiently large number of design values. Early works on RCR models (e.g., Hildreth and Houck, 1968; Froehlich, 1973) focus on the estimation of the mean and covariance structure of (A, B). More recently, Dunker et al. (2019) and Hermann and Holzmann (2021) studied the problem of testing specific characteristics of the joint distribution and identifying higher-order mixed moments.

In many applications, however, it is of interest to study other characteristics of the coefficient distribution that are not generally identifiable, e.g., quantiles, modes, and density support. Thus, the estimation and inference of those quantities becomes problematic. A remedy to nonidentifiability is to impose unverified assumptions under which the coefficient distribution can be uniquely identified, and hence be estimated. Under such assumptions, existing works have been devoted to the estimation of the unique coefficient distribution.

Nonparametric identifiability of $F_{A,B}$ have been investigated by Beran and Hall (1992), Hoderlein et al. (2010), Masten (2018), and Holzmann and Meister (2020), among others. For example, Beran and Hall (1992) assumed that A and B are independent, and further showed that $F_{A,B}$ is identifiable if the design point x follows a nondegenerate distribution, say F_X, with at least one of the points 0, $+\infty$, or $-\infty$ in the support, or if the distribution of B is uniquely determined by its moments and the support of F_X contains an open interval. Masten (2018) further discussed and summarized the identifiability of the coefficient distribution without the independence assumption. Specifically, $F_{A,B}$ is identifiable if either F_X has full support or the support of F_X contains an open interval, and the joint distribution of (A, B) is uniquely determined by its moments.

Under the desired identifiability, various methods have been proposed to estimate the unique coefficient distribution, $F_{A,B}$. Beran and Hall (1992) proposed a moment-matching method for approximating $F_{A,B}$, assuming A and B are independent. Beran and Millar (1994) proposed an estimation approach that seeks to minimize the distance between the induced distribution and the empirical distribution of (x, Y) under the RCR models. Beran et al. (1996) and Hoderlein et al. (2010) assumed full support of x and that its second moment is not finite (e.g., Cauchy), which allowed the inversion of a Radon transform to estimate the distribution of (A, B). Holzmann and Meister (2020) estimated $F_{A,B}$ when the design variables have a Lebesgue density with polynomial tail behavior, by use of an empirical version of the conditional characteristic function, and provide optimal pointwise convergence rates. Dunker et al. (2021) proposed a quasi-maximum likelihood method while relaxing the assumption that the design variable density has heavy tails. Gaillac and Gautier (2022) pointed out that the assumption that x has full support is rarely true in practice, and proposed an adaptive estimation method for when the design variable has limited variation.

The traditional approaches in RCR modeling share two common features, a set of assumptions on the coefficient distribution and the design to ensure identifiability and a procedure to estimate the coefficient distribution. In particular, the design point x is required to take values over an open interval, which is rather restrictive for practical purposes. As will be shown in Section 2.2.2 when the cardinality of $\mathcal{X}$ is finite, the coefficient distribution lacks identifiability; in fact, there is a collection of distributions, including $F_{A,B}$ itself, leading to the same output distributions F_{Y_x} for all $x \in \mathcal{X}$. From that perspective, they are all equivalent and form an equivalence class.

In contrast to existing works, we assume that the cardinality of $\mathcal{X}$ is finite and adopt a different approach without enforcing identifiability. Instead of estimating an identifiable, hence unique, coefficient distribution, which essentially is determined by the assumptions imposed, we develop a procedure that yields multiple plausible coefficient distributions that are equivalent in terms of inducing the same output distributions for each $x \in \mathcal{X}$. Our proposed method starts with an arbitrary initial guess for the distribution of (A, B), referred to as a *global ansatz*. The global ansatz contains available prior information on the structure of a coefficient distribution that may

not be learned from the output distributions, facilitating the identification of a specific member within the equivalence class. The method proceeds by recursively updating the global ansatz. We show that each update moves the global ansatz closer to the desired equivalence class in terms of Kullback–Leibler (KL) divergence.

An alternative viewpoint of our proposed method arises from its connection to the *stochastic inverse problem* (SIP; Butler et al., 2014). Those authors studied the problem of finding a probability measure on the input space for a given probability measure on the output space under a measurable map. They further proposed an algorithm based on the disintegration theorem (Chang and Pollard, 1997), which provides a framework to define conditional probability. In our context, for a pre-specified design point x and the corresponding output distribution F_{Y_x}, finding a coefficient distribution that induces F_{Y_x} through (2.1) is indeed a SIP. At each iteration, our proposed approach solves a collection of SIPs. Of course, the output distributions F_{Y_x} are not observed, rather, we can estimate each output distribution from the observed data. The SIP with the estimated output distributions substituted for the output distributions can be considered as an *empirical* SIP (Bingham et al., 2024).

The remainder of this chapter is organized as follows. In Section 2.2, we introduce the problem setup and discuss the nonidentifiability of the coefficient distribution. In Section 2.3, we develop an iterative method based on recursively updating a global ansatz. In Section 2.4, we validate our proposed approach through simulation studies. In Section 2.5, the proposed method is applied to a randomized-controlled clinical trial of acupuncture (Vickers et al., 2004). We conclude with a discussion in Section 2.6.

2.2 Problem Formulation and Nonidentifiability

2.2.1 Random Coefficient Regression: an SIP Perspective

Consider linear maps from $\mathbb{R}^2$ to $\mathbb{R}$,

$$Q_x(a, b) = a + bx,$$

which are indexed by $x \in \mathcal{X}$. The RCR model in (2.1) can be written as $Y = Q_x(A, B)$.

Each map Q_x gives rise to two types of problems, *stochastic forward problem* and *stochastic inverse problem*. A stochastic forward problem intends to find the distribution F_{Y_x} of Y induced by Q_x for a given distribution $F_{A,B}$. Here, $F_{A,B}$ is referred to as the *generating distribution*, and F_{Y_x} is referred to as an *output distribution* (Butler et al., 2014; Chi, 2021; Bingham et al., 2024).

Conversely, for a stochastic inverse problem, the output distribution F_{Y_x} is known, and the goal is to identify a distribution, whose output distribution is also F_{Y_x}. Each distribution is called a solution to the SIP. Define the solution set $\mathcal{E}_x$, for $x \in \mathcal{X}$,

$$\mathcal{E}_x := \left\{ \widetilde{F}_{A,B} : \widetilde{F}_{A,B} \text{ induces } F_{Y_x} \text{ via } Q_x \right\}.$$

Clearly, the generating distribution is an element in the solution set; however, for a given output distribution F_{Y_x}, the solution set generally contains more than one element. For instance, given the distribution of $A + B$, corresponding to $x = 1$, we can find many solutions through deconvolution.

The RCR model can be viewed as a collection of SIPs, indexed by $x \in \mathcal{X}$. In particular, given a set of output distributions $\mathcal{F}_{\mathcal{X}} = \{F_{Y_x} : x \in \mathcal{X}\}$, the aim of an RCR model is to obtain a coefficient distribution that can act as the generating distribution for all F_{Y_x}. Alternatively, we are seeking a coefficient distribution that lies in the intersection of all solution sets $\mathcal{E}_x$ for $x \in \mathcal{X}$, denoted by $\mathcal{E}$. For a general discussion, see Chi (2021). The author commented that $\mathcal{E}$ is a convex set.

For each $x \in \mathcal{X}$, $\mathcal{E}_x$ is determined by the corresponding output distribution F_{Y_x}. The solution set $\mathcal{E}$ is characterized by the collection of output distributions. All elements in $\mathcal{E}$ are *equivalent* as they induce the same collection of output distributions. Under such equivalence relation, the solution set $\mathcal{E}$ is an equivalence class.

There are important considerations inherent with the use of RCR models. One of which, is that the model, $Y = Q_x(A, B)$, is correctly specified. If so, $\mathcal{E}$ is non-empty as the generating distribution $F_{A,B} \in \mathcal{E}$. Under model misspecification, however, there is no guarantee that there exists a distribution $\widetilde{F}_{A,B}$ that induces the same set of output distributions $\mathcal{F}_{\mathcal{X}}$ through (2.1), i.e., $\mathcal{E}$ may be empty.

Additionally, it is common to assume that the generating distribution has a Lebesgue density, as it can be beneficial to work with the densities directly. Consequently, in this chapter, it is assumed that the Lebesgue density of $F_{A,B}$ exists, denoted as $f_{A,B}$, and is Riemann integrable. Under such assumptions, the solution to the stochastic forward problem can be expressed, in terms of densities, as the well-known *Radon transform* (Deans, 1983),

$$f_{Y_x}(y) = \int_{\mathbb{R}} f_{A,B}(y - bx, b)\ \mathrm{d}b, \tag{2.2}$$

where f_{Y_x} is the density of the output distribution F_{Y_x}. Then, for design point x, any density $\widetilde{f}_{A,B}$ whose Radon transform is equal to f_{Y_x}, defines a distribution $\widetilde{F}_{A,B} \in \mathcal{E}_x$, i.e., $\widetilde{F}_{A,B}$ is a solution to the SIP specified by Q_x.

In practical applications of RCR models, the output distributions (i.e., the true distributions of Y at the specified design points) are unknown. Instead, we are presented with observations of Y from the output distribution F_{Y_x} for each $x \in \mathcal{X}$. This leads to a collection of *empirical* SIPs (Bingham et al., 2024), in which the output distributions are replaced by their empirical analogs. Common choices are histograms and kernel density estimators.

2.2.2 Nonidentifiability and a Solution Set

In this section, we will show that, under some mild conditions, the solution set $\mathcal{E}$ contains infinitely many elements. As a result, these elements, i.e., distributions, are indistinguishable in terms of generating the same collection of output distributions.

To provide an illustration, we present an example that commonly arises in randomized clinical trials. Let x be the indicator of an individual's treatment assignment, with $x = 0$ for control and 1 for treatment; hence, $\mathcal{X} = \{0, 1\}$. In this context, A represents the individual's potential outcome under control, and B represents the individual treatment effect (ITE), i.e., the contrast between the individual's potential outcomes under treatment and control. Most of the time, the primary interest lies in the population average (or mean) treatment effect (PATE), $\mathrm{E}(B)$. Here, we consider more generally the joint distribution of (A, B), which characterizes the distribution of the ITE in the

population. Suppose the generating distribution $F_{A,B}$ is a standard bivariate normal distribution with density

$$f_{A,B}(a,b) = \frac{1}{2\pi}\exp\left\{-\frac{1}{2}(a^2+b^2)\right\},$$

which induces the output distribution, $\mathrm{N}(0, x^2+1)$, for $x \in \mathcal{X} = \{0,1\}$. Note that, for any $x \in \mathcal{X}$, the following two densities also give rise to the same output distributions

$$f^1_{A,B}(a,b) = \frac{1}{\sqrt{7}\pi}\exp\left\{-\frac{2}{7}(2a^2+b^2+ab)\right\},$$
$$f^2_{A,B}(a,b) = \frac{1}{2\pi}\exp\left\{-\frac{1}{2}(a^2+b^2)\right\}\left(1+ab(a^2-b^2)\exp\left\{-\frac{1}{2}(a^2+b^2)\right\}\right).$$

The three densities $f_{A,B}$, $f^1_{A,B}$, and $f^2_{A,B}$ are depicted by heatmaps shown in Panels A, B, and C of Figure 2.1, respectively. Here, $f^1_{A,B}$ is a bivariate normal density with covariance $-1/2$ between A and B, and the variance of A and B are 1 and 2, respectively. Additionally, $f^2_{A,B}$ is a non-normal density such that A and B are uncorrelated, and the densities of A, B, $A+B$, and $A-B$ are all normal (Hamedani and Tata, 1975). Interestingly, $f_{A,B}$ and $f^2_{A,B}$ yield the same output distribution for $x = -1 \notin \mathcal{X}$.

For this example, the solution set $\mathcal{E}$ contains, at least, all three distributions as well as their convex combinations. Thus, it consists of infinitely many elements. This phenomenon is not specific to the situation where the design variable is binary, or the output distributions are normal. Rather, it is a general issue when the cardinality of $\mathcal{X}$ is finite. This result is formally presented in Theorem 1.

Theorem 1. Let $\mathcal{X}$ be a finite collection of design points. Let $F_{A,B}$ be a generating distribution of (A,B), whose Lebesgue density is Riemann integrable, and $\mathcal{F}_{\mathcal{X}}$ be the collection of output distributions. The solution set $\mathcal{E}$ contains infinitely many elements that are absolutely continuous with respect to the Lebesgue measure.

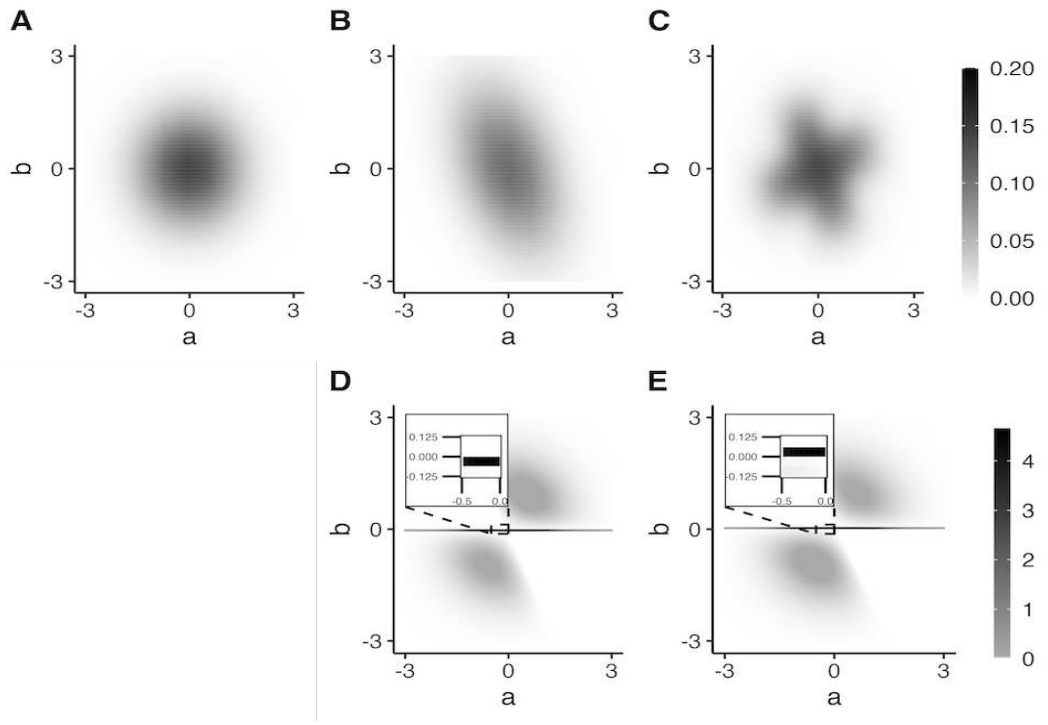

Figure 2.1: Heatmaps of five different densities of (A, B) that induce the same $\mathrm{N}(0, x^2 + 1)$ output distributions for Y at $x \in \mathcal{X} = \{0, 1\}$. Panels A-E show $\widetilde{f}_{A,B}$, $f^1_{A,B}$, $f^2_{A,B}$, $f^3_{A,B}$ and $f^4_{A,B}$, respectively. In Panels D and E, the dark density band is beneath and above the horizontal line $b = 0$, respectively.

The proof of Theorem 1 is included in the Appendix A.0.1, which is based on a construction method. The method creates a new density $\widetilde{f}_{A,B}$ by perturbing the density function $f_{A,B}$ in such a way to preserve the output densities for each $x \in \mathcal{X}$. The construction method is also used in Section 2.5 to generate more empirical coefficient distributions that are equivalent in terms of the output distributions.

Although the generating distribution $F_{A,B}$ is nonidentifiable, some features of $\mathcal{E}$ may still be identifiable. For example, when $\mathcal{X}$ contains at least two points, $\mathrm{E}(A)$ and $\mathrm{E}(B)$ are identifiable, given that $\mathrm{E}(A)$ and $\mathrm{E}(B)$ exist. More generally, Hermann and Holzmann (2021) showed that if the first k moments of A and B exist and x takes on $k+1$ different values, then all mixed moments $\mathrm{E}[A^i B^j]$, $i, j \geq 0$, $i + j \leq k$, are identifiable.

In general, certain important features of $\mathcal{E}$ remain nonidentifiable, such as the probability of an event of (A, B). As an illustration, let us revisit the aforementioned randomized clinical trial example. Recall that the true generating distribution of (A, B) is standard bivariate normal, which

leads to a PATE of $\mathrm{E}(B) = 0$. Now, suppose a positive value of B is indicative of a therapeutic effect. Consider the problem of identifying the population proportion of patients who benefit more from the treatment, i.e., $F_{A,B}(B > 0)$. The three distributions, $F_{A,B}$, $F^1_{A,B}$, and $F^2_{A,B}$, assign the same probability of 0.5 to $\{B > 0\}$. However, other distributions in $\mathcal{E}$ lead to different probabilities for this event.

For example, Panels D and E of Figure 2.1 depict the densities $f^3_{A,B}$ and $f^4_{A,B}$ of two distributions $F^3_{A,B}$ and $F^4_{A,B}$ in $\mathcal{E}$, obtained via the construction method described in the proof of Theorem 1. The functions $f^3_{A,B}$ and $f^4_{A,B}$ yield high density within a narrow band that runs horizontal near the line $b = 0$. For $f^3_{A,B}$ ($f^4_{A,B}$), the band lies under (above) the horizontal line, $b = 0$, as illustrated in Figure 2.1. Consequently, we have $F^3_{A,B}(B > 0) = 0.15$ and $F^4_{A,B}(B > 0) = 0.85$, respectively. The implication is that, while the PATE is zero, the probability that a randomly selected individual benefits more from the treatment than control has a considerably wide range. The analytic expressions of $f^3_{A,B}$ and $f^4_{A,B}$ are given in the Appendix A.0.2.

It may be of interest to the reader to know that not all distributions in $\mathcal{E}$, from this example, have Lebesgue densities. An example of such a distribution is given in the Appendix A.0.2, and it assigns a probability of 0.145 to event $B > 0$. This implies that, although the unknown generating distribution $\widetilde{F}_{A,B}$ is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^2$, its corresponding solution set $\mathcal{E}$ may contain distributions that are not absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^2$. In this chapter, we focus our attention on the restriction of $\mathcal{E}$ to those distributions with a Lebesgue density.

In Section 2.3, we develop an iterative method to approximate elements of $\mathcal{E}$ by incorporating prior information through a global ansatz. This provides a starting point for making inferences about nonidentifiable features of $\mathcal{E}$ that are of interest. For sensitivity analysis, it may be desirable to evaluate how inference on events may vary across the solution set. Different elements of $\mathcal{E}$ (or their approximants) can be obtained through (i) the construction method outlined in the proof of Theorem 1, (ii) the iterative method with different global ansatzes, or (iii) a combination of

both (i) and (ii). An illustration of our proposed method and the sensitivity analysis scheme on an acupuncture dataset is provided in Section 2.5.

2.3 Methodology

2.3.1 Finding a Local Solution in $\mathcal{E}_x$

Let $\{f_{Y_x} : x \in \mathcal{X}\}$ be a set of output densities associated with $\mathcal{F}_{\mathcal{X}}$. Recall that, as described in Section 2.2.1, $\mathcal{E}_x$ is the solution set to the SIP determined by the map Q_x and its corresponding output distribution F_{Y_x}. Furthermore, $\mathcal{E}$ is the solution set formed as the intersection of $\mathcal{E}_x$ for all $x \in \mathcal{X}$.

To begin, we characterize the support of all elements $\mathcal{E}$ in the following lemma.

Lemma 1. Let $\widetilde{F}_{A,B} \in \mathcal{E}$ and $\widetilde{f}_{A,B}$ be its density. Then

$$\text{supp}(\widetilde{f}_{A,B}) \subseteq \bigcap_{x \in \mathcal{X}} Q_x^{-1}(\text{supp}(f_{Y_x})) := \Lambda_0.$$

Here, $\text{supp}(f)$ denotes the support of a density function f, which is the smallest closed set Λ such that $\int_\Lambda f(a,b)\mathrm{d}(a,b) = 1$.

The proof of Lemma 1 is given in the Appendix A.0.3. Note that Λ_0 is a superset of the supports of all solutions in $\mathcal{E}$. It can be readily obtained based on the supports of the output distributions.

Now, consider an arbitrary distribution of (A, B), denoted by $F^{(0)}_{A,B}$, which is not necessarily an element of $\mathcal{E}$. We refer to this initial guess as a *global ansatz*, the meaning of which will be clear later. As a direct consequence of Lemma 1, we should start with a global ansatz such that $\text{supp}(f^{(0)}_{A,B}) \supseteq \Lambda_0$, where $f^{(0)}_{A,B}$ denotes the density of $F^{(0)}_{A,B}$.

For any design point $x \in \mathcal{X}$, $f^{(0)}_{A,B}$ induces a density of Y under the linear map Q_x,

$$f^{(0)}_{Y_x}(y) = \int_{\mathbb{R}} f^{(0)}_{A,B}(y - bx, b)\ \mathrm{d}b.$$

The induced density $f^{(0)}_{Y_x}$ does not necessarily match the target output density f_{Y_x} for any $x \in \mathcal{X}$. However, for each $x \in \mathcal{X}$, the following update of $f^{(0)}_{A,B}$,

$$f^{(1)}_{A,B,x}(a,b) = f^{(0)}_{A,B}(a,b)\frac{f_{Y_x}(a+bx)}{f^{(0)}_{Y_x}(a+bx)}, \tag{2.3}$$

gives a valid density of (A, B) that yields the target output density f_{Y_x} under the linear map Q_x. Note that we define $f^{(1)}_{A,B,x}(a,b) = 0$ whenever $f^{(0)}_{Y_x}(a,b) = 0$. The proof and necessary derivation are included in the Appendix A.0.4. A measure-theoretic justification of the update through the disintegration theorem is deferred to Section 2.3.3.

The additional subscript x in $f^{(1)}_{A,B,x}$ emphasizes the dependence of the update on x. In general, $f^{(1)}_{A,B,x}$ does not induce the target output density $f_{Y_{x'}}$ under $Q_{x'}$ for $x' \neq x$. Thus, its distribution $F^{(1)}_{A,B,x}$ is an element in $\mathcal{E}_x$ and is referred to as a *local* solution at the specific design point x. Additionally, it is shown in the Appendix A.0.5 that (2.3) follows the principle of minimum discrimination information. That is,

$$F^{(1)}_{A,B,x} = \underset{\widetilde{F}_{A,B}\in\mathcal{E}_x}{\arg\min}\, \mathcal{D}_{KL}(\widetilde{F}_{A,B}||F^{(0)}_{A,B}),$$

where $\mathcal{D}_{KL}$ denotes the Kullback-Leibler (KL) divergence. Then $F^{(1)}_{A,B,x}$ is the unique Bergman projection, in terms of the KL-divergence, of $F^{(0)}_{A,B}$ onto the solution set $\mathcal{E}_x$. Note that the uniqueness of the minimum is a direct consequence of the convexity of $\mathcal{E}_x$.

The choice of $F^{(0)}_{A,B}$ is arbitrary, with the only constraints being $\text{supp}(f^{(0)}_{A,B}) \supseteq \Lambda_0$ to avoid dividing any non-zero number by zero in (2.3) and it has a Lebesgue density. Finding a local solution is analogous to reconstructing a joint density from a marginal, in which the solution is not unique and requires an additional assumption on the conditional density. Specifically, the component $f^{(0)}_{A,B}(a,b)/f^{(0)}_{Y_x}(a+bx)$ in (2.3) represents a prior assumption about the structure of the local solution that cannot be learned from the target output density. Such an assumption is referred to as an *ansatz* (Butler et al., 2014); see Section 2.3.3 for more details. We refer to $F^{(0)}_{A,B}$ as a *global ansatz*, as it induces the individual ansatzes at different design points.

The role of an ansatz in the SIP with non-unique solutions is analogous to that of a prior in Bayesian inference with nonidentifiable parameters (Neath and Samaniego, 1997; Wechsler et al., 2013). If subject-matter knowledge is available, such information can be used to set up the global ansatz. Otherwise, for standard noninformative choices, the global ansatz can be chosen to be a diffuse bivariate normal distribution (when Λ_0 is unbounded) or a uniform distribution (when Λ_0 is bounded).

2.3.2 Approximating a Global Solution in $\mathcal{E}$

A local solution yields the target output density for a specific x, but our goal is to identify (or approximate) an element of $\mathcal{E}$ which induces the target output densities globally for all $x \in \mathcal{X}$. To be more precise in terminology, we occasionally refer to an element of $\mathcal{E}$ as a *global* solution. In this section, we propose a new method to approximate a global solution based on local solutions.

As discussed in Section 2.3.1, for each $x \in \mathcal{X}$, can find a local solution $f^{(1)}_{A,B,x}$. A natural first step is to remove the effect of the design point x by averaging all local solutions over $\mathcal{X}$,

$$f^{(1)}_{A,B}(a,b) = \frac{1}{J}\sum_{x\in\mathcal{X}} f^{(1)}_{A,B,x}(a,b), \tag{2.4}$$

where J is the cardinality of $\mathcal{X}$. The resulting density $f^{(1)}_{A,B}$ may not be a global solution, and in fact, it may not be a local solution in any $\mathcal{E}_x$. However, as will be shown in Theorem 2, we have

$$\mathcal{D}_{KL}(\widetilde{F}_{A,B}||F^{(0)}_{A,B}) \geq \mathcal{D}_{KL}(\widetilde{F}_{A,B}||F^{(1)}_{A,B}),$$

where $\widetilde{F}_{A,B}$ is any global solution in $\mathcal{E}$, and $F^{(0)}_{A,B}$ and $F^{(1)}_{A,B}$ are the distribution functions corresponding to $f^{(0)}_{A,B}$ and $f^{(1)}_{A,B}$, respectively. Therefore, the update from $f^{(0)}_{A,B}$ to $f^{(1)}_{A,B}$ can be seen as a step closer to a global solution.

Note that $f_{A,B}^{(1)}$ is independent of the design point x and can be used as an updated global ansatz. We can repeat the process above through the following recursive equation,

$$f_{A,B}^{(\ell+1)}(a,b) = \frac{1}{J}\sum_{x\in\mathcal{X}} f_{A,B,x}^{(\ell+1)}(a,b) = f_{A,B}^{(\ell)}(a,b)\frac{1}{J}\sum_{x\in\mathcal{X}}\frac{f_{Y_x}(a+bx)}{f_{Y_x}^{(\ell)}(a+bx)}, \quad \text{for } \ell = 0,1,\ldots, \tag{2.5}$$

where, for $x \in \mathcal{X}$,

$$f_{Y_x}^{(\ell)}(y) = \int_{\mathbb{R}} f_{A,B}^{(\ell)}(y-bx,b)\ \mathrm{d}b$$

is the output distribution under the ℓth updated global ansatz.

The rationale behind the recursive update is as follows. At each iteration, the local solutions are averaged, merging their distinct features. The average then serves as the prior to guide the next iteration, resulting in a new set of local solutions. This iterative process promotes similarity among the local solutions over iterations, driving them towards a global solution.

Next, Theorem 2 provides the convergence results in terms of the KL-divergence.

Theorem 2. Let $F_{A,B}^{(0)}$ be a global ansatz with the density $f_{A,B}^{(0)}$ satisfying $\mathrm{supp}(f_{A,B}^{(0)}) \supseteq \Lambda_0$. Further, let $\{F_{A,B}^{(\ell)} : \ell = 0,1,\ldots\}$ be a sequence of distributions defined by the density functions in the recursive equation (2.5). For any global solution $\widetilde{F}_{A,B} \in \mathcal{E}$, if $\mathcal{D}_{KL}(\widetilde{F}_{A,B}||F_{A,B}^{(0)}) < \infty$, then

(i) the sequence $\mathcal{D}_{KL}(\widetilde{F}_{A,B}||F_{A,B}^{(\ell)})$ is non-increasing and converges;

(ii) the sequence $\mathcal{D}_{KL}(F_{A,B}^{(\ell+1)}||F_{A,B}^{(\ell)})$ converges to zero; and

(iii) the sequence $\mathcal{D}_{KL}(F_{Y_x}||F_{Y_x}^{(\ell)})$ converges to zero for each $F_{Y_x} \in \mathcal{F}_{\mathcal{X}}$.

The proof of Theorem 2 is deferred to the Appendix A.0.6. Theorem 2 suggests that the recursive update is a viable approach to moving closer to $\mathcal{E}$ from an arbitrary global ansatz, thereby producing an improved approximation of a global SIP solution. Evaluating (2.5) over many iterations may not be analytically tractable. In practice, we consider a computational method that numerically approximates (2.5) with discrete approximations and numeric integration. For $\ell \geq 10$, $f_{A,B}^{(\ell)}$ typically yields $f_{Y_x}^{(\ell)}$ that closely resembles the target output density f_{Y_x} for all $x \in \mathcal{X}$.

As an example, Figure 2.2 illustrates the recursive update for the example in Section 2.2.2. Recall that the output distributions are $\mathrm{N}(0, x^2 + 1)$ for $x \in \mathcal{X} = \{0, 1\}$. Let the global ansatz be a bivariate normal distribution with mean $(2, 1)^\top$ and covariance matrix I_2 (the 2×2 identity matrix), shown in Panel A. Panels B and C visualize the local solutions at $x = 0$ and 1 after the first update, respectively, and Panel D shows their average. Next, starting with the distribution in Panel D, the process is repeated, producing the local solutions in Panels E and F and their average in Panel G. Lastly, Panels H, I and J show the local solutions and their average after 15 updates, which exhibit a noticeable similarity.

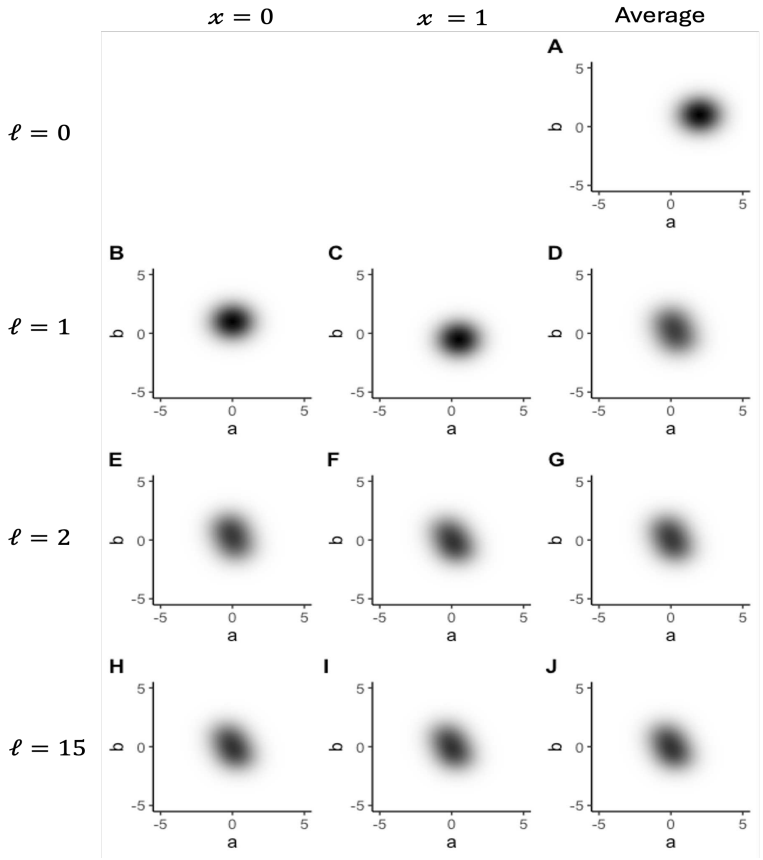

Figure 2.2: Illustration of the recursive update for approximating a global solution. The output distributions are $\mathrm{N}(0, x^2 + 1)$ for $x \in \mathcal{X} = \{0, 1\}$. Panel A shows a bivariate normal (mean = $(2, 1)^\top$, covariance matrix = I_2) distribution as the global ansatz. The second row visualizes the first update, where Panels B, C and D represent the local solutions at $x = 0$ and 1 as well as their average, respectively. Similarly, the third row visualizes the second update, and the last row visualizes the 15th update. The scale of the densities ranges from 0 to 0.16.

To demonstrate that the distribution $F_{A,B}^{(15)}$ in Panel J of Figure 2.2 yields output distributions that closely resemble the target $\mathrm{N}(0, x^2 + 1)$ for both $x = 0$ and 1, Figure 2.3 plots the induced output distributions against F_{Y_x} for $x = 0, 1$. In Panel A, $F_{Y_0}^{(\ell)}$ (horizontal axis) is plotted against F_{Y_0} (vertical axis) for $\ell = 0$ (dashed curve), $\ell = 1$ (dotted curve), $\ell = 2$ (dash-dotted curve), and $\ell = 15$ (solid curve). Similarly, Panel B plots $F_{Y_1}^{(\ell)}$ (horizontal axis) against F_{Y_0} (vertical axis) for $\ell = 0, 1, 2,$ and 15. As ℓ increases, the curves become closer to a straight 45 degree line. In particular, the solid curves, as seen in both Panels A and B, suggest that $F_{Y_0}^{(15)}$ and $F_{Y_1}^{(15)}$ approximate the respective output distributions, F_{Y_0} and F_{Y_1}, reasonably well.

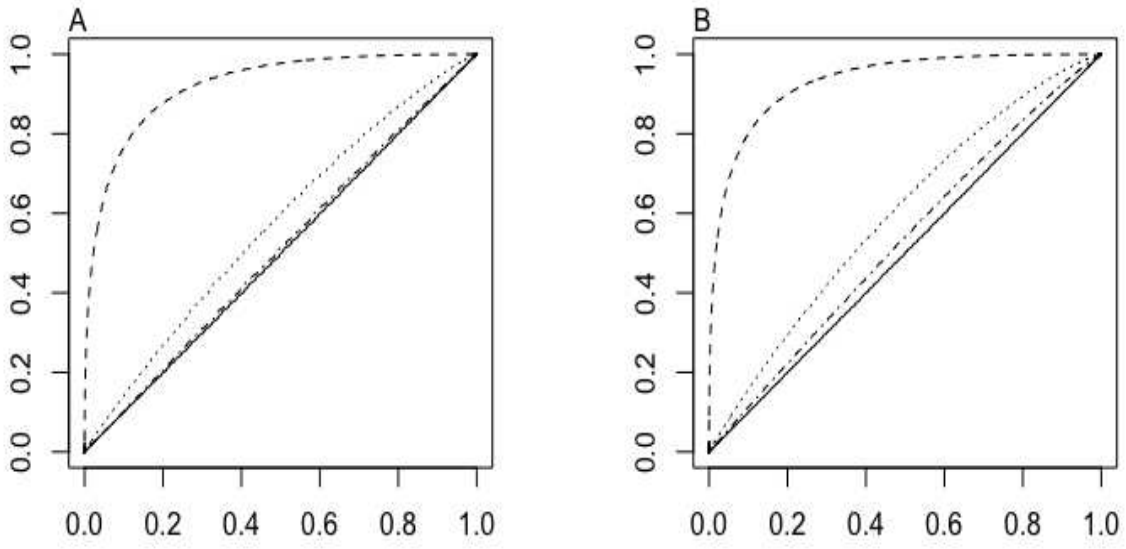

Figure 2.3: Induced output distributions of the iterative procedure by an off-centered normal distribution against the output distributions of $\mathcal{F}_{\mathcal{X}}$. Panel A plots $F_{Y_0}^{(\ell)}$ (horizontal axis) against F_{Y_0} (vertical axis) for $\ell = 0, 1, 2,$ and 15. Similarly, Panel B plots $F_{Y_0}^{(\ell)}$ against F_{Y_0}. In both panels, the dashed lines correspond with $\ell = 0$, the dotted lines correspond with $\ell = 1$, the dashed-dotted lines correspond with $\ell = 2$, and the solid lines correspond with $\ell = 15$.

In the RCR model, we are often presented with observations of Y at different design points, which can be used to generate a set of estimated output densities $\{\hat{f}_{Y_x} : x \in \mathcal{X}\}$ through his-

tograms or kernel density estimators. The recursive update in (2.5) can be applied directly to the estimated output densities to produce an approximation $\hat{F}_{A,B}$ of the coefficient distribution. Additional goodness-of-fit checks, such as QQ plots and Kolmogorov–Smirnov (KS) tests, can be used to assess whether $\hat{F}_{A,B}$ could have plausibly given rise to the observed data. Our approach is supported by simulation studies, as shown later in Section 2.4.

2.3.3 A Measure-Theoretic Framework

In this section, we provide a measure-theoretic justification for our proposed method. Particularly, we will highlight that the essence of the local solution for a given x at each iteration is to solve an SIP through a disintegration. Butler et al. (2014, 2012); Breidt et al. (2011) proposed a theoretical framework for solving SIPs by means of disintegration. Chi (2021) expanded this work by leveraging an experimental variable x.

The general setup is as follows. Let $(\Lambda, \mathcal{B}_\Lambda, \mathbb{P}_\Lambda)$ denote a probability space, where $\Lambda \subseteq \mathbb{R}^k$, $\mathcal{B}_\Lambda$ is the Borel σ-algebra on Λ, and $\mathbb{P}_\Lambda$ is a probability measure on $(\Lambda, \mathcal{B}_\Lambda)$. Furthermore, for each $x \in \mathcal{X}$, let Q_x be a measurable map from $(\Lambda, \mathcal{B}_\Lambda)$ to $(\Gamma_x, \mathcal{B}_{\Gamma_x})$, where $\Gamma_x \subseteq \mathbb{R}^m$, and $\mathcal{B}_{\Gamma_x}$ is the Borel σ-algebra on Γ_x. It is assumed that Q_x is surjective, continuously differentiable, and the Jacobian of Q_x has full rank. Note that a probability measure $\mathbb{P}_{\Gamma_x}$ on $(\Gamma_x, \mathcal{B}_{\Gamma_x})$ is induced by $\mathbb{P}_\Lambda$ via

$$\mathbb{P}_{\Gamma_x}(D) = \mathbb{P}_\Lambda(Q_x^{-1}(D)) \quad \text{for } D \in \mathcal{B}_{\Gamma_x}, \tag{2.6}$$

which is referred to as the output probability measure. Here, Q_x^{-1} is the inverse map, defined by

$$Q_x^{-1}(D) = \{\lambda \in \Lambda : Q_x(\lambda) \in D\}.$$

Finding $\mathbb{P}_{\Gamma_x}$ through (2.6) defines a stochastic forward problem. On the other hand, for each x, given an output probability measures $\mathbb{P}_{\Gamma_x}$ on $(\Gamma_x, \mathcal{B}_{\Gamma_x})$, finding a probability measure $\mathbb{P}_\Lambda$ on $(\Lambda, \mathcal{B}_\Lambda)$ that induces $\mathbb{P}_{\Gamma_x}$ is called a SIP.

Under the RCR model in (2.1), Λ is the domain of (A, B), $Q_x(\lambda) = a + bx$ is a linear map for $\lambda = (a, b) \in \Lambda$, Γ_x is the domain of Y for any given x, and the probability measures $\mathbb{P}_\Lambda$ and $\mathbb{P}_{\Gamma_x}$ are respectively the distributions of $F_{A,B}$ and F_{Y_x}. In this chapter, we assume that $\Lambda = \mathbb{R}^2$, and hence $\Gamma_x = \mathbb{R}$. Our goal is to find probability measures $\widetilde{\mathbb{P}}_\Lambda$ on $(\Gamma_x, \mathcal{B}_{\Gamma_x})$ that induce the set of output measures $\{\mathbb{P}_{\Gamma_x} : x \in \mathcal{X}\}$ via the set of maps $\{Q_x : x \in \mathcal{X}\}$. That is, $\widetilde{\mathbb{P}}_\Lambda$ is a solution to each SIP defined by the maps Q_x, and hence, is a global solution as defined in Section 2.2.1.

The approach of Butler et al. (2014), Yang (2018), and Chi (2021) is based on the disintegration theorem (Chang and Pollard, 1997). Consider a σ-finite measure Ψ_Λ on $(\Lambda, \mathcal{B}_\Lambda)$. Let $Q_x\Psi_\Lambda$ denote the *image measure* of Ψ_Λ under the map Q_x, defined as $Q_x\Psi_\Lambda(D) = \Psi_\Lambda(Q_x^{-1}(D))$ for $D \in \mathcal{B}_{\Gamma_x}$. The disintegration theorem states that, under mild conditions, Ψ_Λ can be disintegrated as

$$\Psi_\Lambda(C) = \int_{y \in \Gamma_x} \int_{\lambda \in Q_x^{-1}(y) \cap C} \mathrm{d}\Psi_{Q_x^{-1}(y)}(\lambda) \mathrm{d}\nu_{\Gamma_x}(y), \ \forall C \in \mathcal{B}_\Lambda. \tag{2.7}$$

See Yang (2018). Here, ν_{Γ_x} is a σ-finite measure on $(\Gamma_x, \mathcal{B}_{\Gamma_x})$ that dominates the image measure $Q_x\Psi_\Lambda$ and is called the *mixing measure*. Furthermore, for each y, $\Psi_{Q_x^{-1}(y)}$ is a σ-finite measure on $(\Lambda, \mathcal{B}_\Lambda)$ concentrated on the set $Q_x^{-1}(y) = \{\lambda \in \Lambda : Q(\lambda) = y\}$, often referred to as the *generalized contour* (Butler et al., 2014). That is, $\Psi_{Q_x^{-1}(y)}\big(\Lambda \setminus Q_x^{-1}(y)\big) = 0$ for ν_{Γ_x}-almost all y. The measures $\{\Psi_{Q_x^{-1}(y)} : y \in \Gamma_x\}$ are called *disintegrating measures* or a (Q_x, ν_{Γ_x})-*disintegration* of Ψ_Λ. The disintegrating measures are uniquely determined up to an almost sure equivalence: if $\{\Psi^*_{Q_x^{-1}(y)}\}$ is another (Q_x, ν_{Γ_x})-disintegration of Ψ_Λ, then $\nu_{\Gamma_x}\left(y \in \Gamma_x : \Psi_{Q_x^{-1}(y)} \neq \Psi^*_{Q_x^{-1}(y)}\right) = 0$.

The disintegration theorem can be viewed as representing Ψ_Λ as an iterated integral of a *conditional* measure $\Psi_{Q_x^{-1}(y)}$ along each generalized contour and a *marginal* measure ν_{Γ_x} over the set of generalized contours. In the context of the RCR model, Figure 2.4 illustrates the disintegration theorem for $x = 1$. Each generalized contour, $Q_x^{-1}(y) = \{\lambda = (a, b) : b = -a + y\}$, is a straight line with slope -1, and a disintegrating measure $\Psi_{Q_x^{-1}(y)}$ concentrates on such a line. The mixing measure ν_{Γ_x} defines a measure over these parallel lines.

Now, let us revisit our iterative approach for RCR models. We begin by selecting *a priori* a global ansatz $\mathbb{P}_\Lambda^{(0)}$, under mild constraints. The global ansatz is disintegrated with respect to the

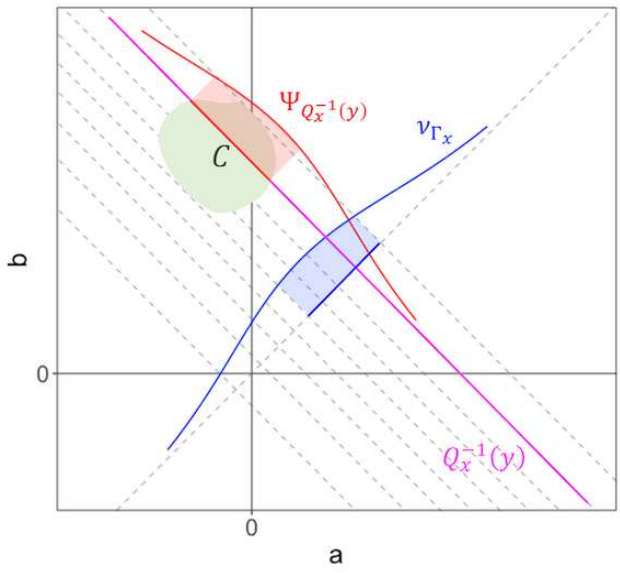

Figure 2.4: The disintegration of a measure Ψ_Λ for $x = 1$. The contours are lines with slopes -1, a few of which are depicted as parallel dashed lines. The density of a disintegrating measure $\Psi_{Q_x^{-1}(y)}$, concentrated on $Q_x^{-1}(y)$ for $y \in \Gamma_x$, is shown as a red solid line. The shaded region under the red curve, represents the $\Psi_{Q_x^{-1}(y)}$ measure of set C. The density of the mixing measure ν_{Γ_x} over the contours is shown as a blue solid line. The Ψ_Λ measure of set C, can then be viewed as the integral of the $\Psi_{Q_x^{-1}(y)}$ measures of set C with respect to the mixing measure ν_{Γ_x}.

image measure $\mathbb{P}^{(0)}_{\Gamma_x} = Q_x \mathbb{P}^{(0)}_\Lambda$ as

$$\mathbb{P}^{(0)}_\Lambda(C) = \int_{y\in\Gamma_x} \int_{\lambda \in Q_x^{-1}(y)\cap C} \mathrm{d}\mathbb{P}^{(0)}_{Q_x^{-1}(y)}(\lambda)\mathrm{d}\mathbb{P}^{(0)}_{\Gamma_x}(y), \quad \forall C \in \mathcal{B}_\Lambda.$$

The resulting set of disintegration measures $\{\mathbb{P}^{(0)}_{Q_x^{-1}(y)} : y \in \Gamma_x\}$ provides an initial guess of the distributions along the collection of generalized contours.

Now, combing with the target output probability measure $\mathbb{P}_{\Gamma_x}$, we obtain a probability measure on Λ through an integration

$$\mathbb{P}^{(1)}_{\Lambda,x}(C) = \int_{y\in\Gamma_x} \int_{\lambda \in Q_x^{-1}(y)\cap C} \mathrm{d}\mathbb{P}^{(0)}_{Q_x^{-1}(y)}(\lambda)\mathrm{d}\mathbb{P}_{\Gamma_x}(y), \quad \forall C \in \mathcal{B}_\Lambda. \tag{2.8}$$

Note that the image measure of $\mathbb{P}^{(1)}_{\Lambda,x}$ is the desired output distribution $\mathbb{P}_{\Gamma_x}$, and hence is a solution to the SIP under Q_x.

Now, for the RCR models, we will show that $f^{(1)}_{A,B,x}$ in (2.3) is the Lebesgue density of $\mathbb{P}^{(1)}_{\Lambda,x}$ in (2.8). Let μ_1 and μ_2 denote the Lebesgue measures on $(\mathbb{R}, \mathcal{B}_{\mathbb{R}})$ and $(\mathbb{R}^2, \mathcal{B}_{\mathbb{R}^2})$, respectively.

First, the (Q_x, μ_1)-disintegration of μ_2 yields the set of disintegrating measures, denoted as $\{\mu_{Q_x^{-1}(y)} : y \in \mathbb{R}\}$. For a given $y \in \mathbb{R}$ and a measurable function $f : \mathbb{R}^2 \to \mathbb{R}$, we show in the Appendix A.0.7 that each disintegrating measure can be obtained through the Radon transform; that is,

$$\mu_{Q_x^{-1}(y)}(f) := \int f \mathrm{d}\mu_{Q_x^{-1}(y)} = \int f(y - sx, s)\mathrm{d}\mu_1(s). \tag{2.9}$$

The notation $\nu(f) = \int f \mathrm{d}\nu$ follows the convention in Chang and Pollard (1997).

Next, assume that the global ansatz $\mathbb{P}^{(0)}_\Lambda$ is absolutely continuous with respect to μ_2 with Radon-Nikodym (RN) derivative $\mathrm{d}\mathbb{P}^{(0)}_\Lambda/\mathrm{d}\mu_2 = f^{(0)}_{A,B}$. According to Theorem 3 of Chang and Pollard (1997), we can express the disintegrating measure of $\mathbb{P}^{(0)}_\Lambda$ for $y \in \Gamma_x$ as

$$\mathbb{P}^{(0)}_{Q_x^{-1}(y)}(C) = \frac{\mu_{Q_x^{-1}(y)}\left(f^{(0)}_{A,B} \cdot \mathbb{I}_C\right)}{\mu_{Q_x^{-1}(y)}(f^{(0)}_{A,B})}, \quad \forall C \in \mathcal{B}_{\mathbb{R}^2}, \tag{2.10}$$

where $\mathbb{I}_C$ is the indicator function of set C. Combining (2.9) and (2.10), the RN derivative of $\mathbb{P}^{(0)}_{Q_x^{-1}(y)}$ with respect to $\mu_{Q_x^{-1}(y)}$ for a given $y \in \Gamma_x$ can be written as

$$\frac{\mathrm{d}\mathbb{P}^{(0)}_{Q_x^{-1}(y)}}{\mathrm{d}\mu_{Q_x^{-1}(y)}}(\lambda) = \frac{f^{(0)}_{A,B}(\lambda) \cdot \mathbb{I}_{Q_x^{-1}(y)}(\lambda)}{f^{(0)}_{Y_x}\left(Q_x(\lambda)\right)}, \tag{2.11}$$

where $f^{(0)}_{Y_x}(y) = \int f^{(0)}_{A,B}(y - sx, s)\mathrm{d}s$ is the line integral of $f^{(0)}_{A,B}$ over $Q_x^{-1}(y)$. The form of the disintegrating density has an intuitive interpretation as the ratio of the joint density of (A, B) to the output density, i.e., as a conditional density.

Lastly, assume that the output probability measure $\mathbb{P}_{\Gamma_x}$ is absolutely continuous with respect to $\mathbb{P}^{(0)}_{\Gamma_x}$ and hence to μ_1 also, with RN derivative $\mathrm{d}\mathbb{P}_{\Gamma_x}/\mathrm{d}\mu_1 = f_{Y_x}$. Plugging-in (2.11), (2.8) can be

expressed as

$$\begin{aligned}\mathbb{P}^{(1)}_{\Lambda,x}(C) &= \int_{y\in\Gamma_x}\left(\int_{\lambda\in Q_x^{-1}(y)\cap C}\frac{f^{(0)}_{A,B}(\lambda)}{f^{(0)}_{Y_x}(y)}\mathrm{d}\mu_{Q_x^{-1}(y)}(\lambda)\right) f_{Y_x}(y)\mathrm{d}\mu_1(y)\\ &= \iint_{\mathbb{R}^2}\frac{f^{(0)}_{A,B}(y-sx,s)}{f^{(0)}_{Y_x}(y)}\cdot\mathbb{I}_C(y-sx,s)\cdot f_{Y_x}(y)\ \mathrm{d}s\mathrm{d}y\\ &= \iint_{\mathbb{R}^2}\frac{f^{(0)}_{A,B}(a,b)}{f^{(0)}_{Y_x}(a+bx)}\cdot f_{Y_x}(a+bx)\cdot\mathbb{I}_C(a,b)\ \mathrm{d}a\mathrm{d}b,\end{aligned}$$

for all $C\in\mathcal{B}_{\mathbb{R}^2}$.

The probability measures $\mathbb{P}^{(1)}_{\Lambda,x}$ are averaged over $\mathcal{X}$ to obtain a probability measure $\mathbb{P}^{(1)}_{\Lambda}$, which does not depend on the design point x. The resulting measure $\mathbb{P}^{(1)}_{\Lambda}$ can then be used as an ansatz, and the procedure repeats. The iterative procedure can be summarized by the following recursive equation,

$$\mathbb{P}^{(\ell+1)}_{\Lambda}(C)=\frac{1}{J}\sum_{x\in\mathcal{X}}\int_{y\in\Gamma_x}\int_{\lambda\in Q_x^{-1}(y)\cap C}\mathrm{d}\mathbb{P}^{(\ell)}_{Q_x^{-1}(y)}(\lambda)\mathrm{d}\mathbb{P}_{\Gamma_x}(y), \tag{2.12}$$

for $\ell=0,1,\ldots$ and $C\in\mathcal{B}_{\mathbb{R}^2}$, where $\{\mathbb{P}^{(\ell)}_{Q_x^{-1}(y)}\}$ is the $(Q_x,\mathbb{P}^{(\ell)}_{\Gamma_x})$-disintegration of $\mathbb{P}^{(\ell)}_{\Lambda}$. The probability measure $\mathbb{P}^{(\ell+1)}_{\Lambda}$ has an RN derivative with respect to μ_2 equal to $f^{(\ell)}_{A,B}$ in (2.5). Thus (2.12) is the measure-theoretic form of (2.5), in terms of measures rather than density functions.

Moreover, (2.5) can be viewed as a special case of the iterative procedure proposed by Chi (2021) for linear maps, in terms of density functions and under the RCR modeling framework, but with a key difference. Chi (2021) implemented ansatzes on individual generalized contours for each map Q_x to initiate the procedure, whereas we adopt a global ansatz. The global ansatz provides a tractable way for a practitioner to supply prior information. For RCR models, our approach with the global ansatz leads to the readily available iterative process, which attains desirable convergence features, such as those described in Theorem 2.

2.4 Simulation

We assess the proposed method through simulation studies. Assume a sample size of $n = 450$ and a balanced experimental design, with design points $\mathcal{X} = \{-1, 0, 1\}$. That is, the number of observations for each distinct design point is 150. The random regression coefficients (A_i, B_i) are drawn independently from a generating distribution $F_{A,B}$ for $i = 1, \ldots, 450$, and for each x_i, Y_i is calculated as $Y_i = Q_{x_i}(A_i, B_i)$.

We consider two simulation scenarios. For the first scenario, $F_{A,B}$ is the standard bivariate normal distribution. For the other, $F_{A,B}$ is a bivariate uniform distribution on $[-5, 5]^2$. To account for sampling variability, 100 datasets are generated under each simulation scenario.

To estimate the output densities based on the simulated data, both histograms and kernel density estimation (KDE) techniques are used. For KDE, an Epanechnikov kernel is used. For both simulation scenarios, two global ansatzes are employed: an off-centered normal distribution (mean = $(2, 1)^\top$, covariance matrix = I_2) and a bivariate distribution made up of two independent univariate logistic distributions with location 0 and scale $\sqrt{3}/\pi$. With two types of density estimators and two global ansatzes, four variations of the iterative method are run on each of the 100 simulated data sets. The iterative process as summarized by (2.5) is applied to the estimated output densities for 10 iterations. The result is used as an approximation of the coefficient distribution, $\hat{F}_{A,B} = F^{(10)}_{A,B}$.

For comparison, we implement the moment-matching estimator (MME) proposed by Beran and Hall (1992) and the Radon transform estimator (RTE) proposed by Hoderlein et al. (2010). Additionally, we consider a nonparametric Bayesian method by modeling the coefficient distribution as a Dirichlet process mixture (DPM) of bivariate normal distributions. The MME, RTE and DPM methods are run on the same simulated datasets.

For each simulated dataset and each method, an approximation of a plausible coefficient distribution is obtained, denoted by $\hat{F}_{A,B}$. Due to nonidentifiability, we do not directly compare $\hat{F}_{A,B}$ with the generating distribution $F_{A,B}$. Rather, the performance of each method is evaluated by comparing the output distributions induced by $\hat{F}_{A,B}$ with those induced by the generating distri-

bution $F_{A,B}$. Specifically, the Kolmogorov–Smirnov (KS) distance is calculated at each design point.

Table 2.1 summarizes the simulation results. The values shown are averages (and standard deviations) of the KS distances over the 100 simulated datasets. As benchmarks, the KS distances between the true output distributions and those estimated by histogram and KDE techniques are also included. The results produced by the four variations of our proposed method are similar, demonstrating its robustness to the choice of the estimator for the output densities and the specification of the global ansatz. Notably, the proposed method yields KS distances comparable to or shorter than the histogram and KDE benchmarks. Both the MME and RTE result in greater KS distances than the proposed method. The DPM has a slight advantage over the proposed method when the generating coefficient distribution is normal, likely because it matches the modeling assumption of DPM. However, when the generating distribution is uniform, the proposed method leads to shorter KS distances than the DPM.

In summary, the proposed method outperforms the alternatives in the sense that the approximated coefficient distributions provide better recovery of the true output distributions, as compared to other methods. Since the proposed method is fully nonparametric, the parametric form of the true generating distribution has little impact on its performance. Additionally, by varying the global ansatz, the proposed method can produce different plausible approximations of the coefficient distribution, whereas the other three methods produce only one approximation for a provided dataset. The simulation results indicate that the proposed method is preferable for approximating coefficient distributions in a RCR model when x is considered fixed on a discrete lattice.

2.5 Real Data: Acupuncture Example

In this section, we apply our iterative method to a dataset from a randomized controlled clinical trial investigating the use of acupuncture to mitigate headaches (Vickers et al., 2004). The patient-level data from this trial were made available by Vickers (2006). The trial consists of 401 patients with chronic headaches, who were assigned to one of two treatment groups based on randomized

Table 2.1: KS distances between the output distributions induced by $\hat{F}_{A,B}$ and $F_{A,B}$. Each row corresponds to a method and each column corresponds to a design point. Values shown are averages (and standard deviations) of the KS distances over the 100 simulated datasets.

Method	$x = -1$	$x = 0$	$x = 1$
Scenario 1: $F_{A,B}$ is standard bivariate normal			
Histogram benchmark	0.0677 (0.0252)	0.0695 (0.0230)	0.0678 (0.0201)
KDE benchmark	0.0440 (0.0228)	0.0468 (0.0202)	0.0448 (0.0179)
Histogram, normal ansatz	0.0495 (0.0189)	0.0472 (0.0173)	0.0491 (0.0158)
Histogram, logistic ansatz	0.0485 (0.0189)	0.0474 (0.0174)	0.0486 (0.0163)
KDE, normal ansatz	0.0434 (0.0198)	0.0348 (0.0150)	0.0428 (0.0158)
KDE, logistic ansatz	0.0421 (0.0195)	0.0378 (0.0151)	0.0413 (0.0155)
MME	0.0999 (0.0110)	0.1599 (0.0329)	0.1707 (0.0095)
RTE	0.1207 (0.0096)	0.1470 (0.0083)	0.1210 (0.0097)
DPM	0.0338 (0.0187)	0.0322 (0.0151)	0.0332 (0.0155)
Scenario 2: $F_{A,B}$ is uniform on $[-5, 5]^2$			
Histogram benchmark	0.0682 (0.0206)	0.0678 (0.0203)	0.0685 (0.0213)
KDE benchmark	0.0462 (0.0191)	0.0498 (0.0154)	0.0456 (0.0208)
Histogram, normal ansatz	0.0519 (0.0169)	0.0492 (0.0151)	0.0511 (0.0171)
Histogram, logistic ansatz	0.0511 (0.0164)	0.0485 (0.0149)	0.0508 (0.0165)
KDE, normal ansatz	0.0445 (0.0173)	0.0481 (0.0123)	0.0434 (0.0171)
KDE, logistic ansatz	0.0439 (0.0170)	0.0467 (0.0123)	0.0431 (0.0170)
MME	0.3056 (0.0207)	0.3599 (0.0132)	0.3155 (0.0266)
RTE	0.0654 (0.0255)	0.0601 (0.0192)	0.0644 (0.0247)
DPM	0.0588 (0.0165)	0.0965 (0.0139)	0.0598 (0.0165)

minimization. One group received acupuncture treatments, and the other received a control intervention. Several outcome measures, such as the headache score on a 0 to 100 scale, were assessed at baseline as well as 3 and 12 months after treatment. For the purpose of illustration, we focus on the decrease in headache score at 12 months compared to that at baseline, which is the trial's primary endpoint. A positive value of the outcome indicates an improvement. There were 301 patients who completed the study, including 161 in the acupuncture group and 140 in the control group. For simplicity, we restrict the analysis to the complete cases.

Suppose that the treatment effect of acupuncture is of interest. A standard analysis involves calculating the sample means of the headache score reductions in the control and acupuncture groups, which are given by 4.367 and 8.329, respectively. The PATE of acupuncture as compared

to control is estimated to be 3.962. No obvious violation of the normality assumption is seen. Therefore, a 95% Welch's confidence interval is calculated for the PATE, which is $(1.339, 6.585)$. A one-sided test against the null hypothesis that the PATE is less than or equal to 0 results in a p-value of 0.002. A limitation of such an analysis, however, is that it does not provide information about how the effect of acupuncture differs among individuals.

The RCR model, as in (2.1), allows inference about the ITE of acupuncture. Let x be the treatment assignment indicator (1 for acupuncture and 0 for control). Accordingly, A is the potential outcome under control, and B is the ITE of acupuncture as compared to control. By approximating the joint distribution of (A, B), a plausible probability of a positive ITE, $F_{A,B}(B > 0)$, can be obtained.

We fit the RCR model to the data and obtain an approximation of a coefficient distribution using our proposed method. A bivariate normal distribution with mean 0 and covariance matrix $25^2 \cdot I_2$ is used as the global ansatz. Next, we use histograms of the actual data to estimate the output densities of Y at $x = 0$ and 1 and implement the update in (2.5). The result after 10 iterations is shown in Panel A of Figure 2.5 and is used as an approximation of a joint distribution of (A, B), $\hat{F}_{A,B} = F^{(10)}_{A,B}$. Based on this approximation, the probability of a positive ITE under $\hat{F}_{A,B}$ is 0.600.

To assess the goodness of fit of the RCR model and the approximated coefficient distribution produced by the iterative method, we carry out two one-sample KS tests. Each KS test compares the empirical distribution of Y with the output distribution induced by $\hat{F}_{A,B}$ at each design point. For $x = 0$ and 1, the KS statistics are 0.051 and 0.050, respectively, with corresponding p-values of 0.856 and 0.808. These results suggest that the observed outcome measures in both treatment groups can reasonably be produced by $\hat{F}_{A,B}$.

As discussed in Section 2.2.2, there exists an infinite set of distributions that yield identical output distributions as $\hat{F}_{A,B}$ on $\mathcal{X} = \{0, 1\}$. For this reason, the probability assigned to the event $\{B > 0\}$ is not uniquely identified. For sensitivity analysis, we employ the construction method used in the proof of Theorem 1 to obtain different coefficient distributions that are equivalent to $\hat{F}_{A,B}$ in terms of output distributions for $x = 0, 1$. Specifically, Panels B and C of Figure 2.5

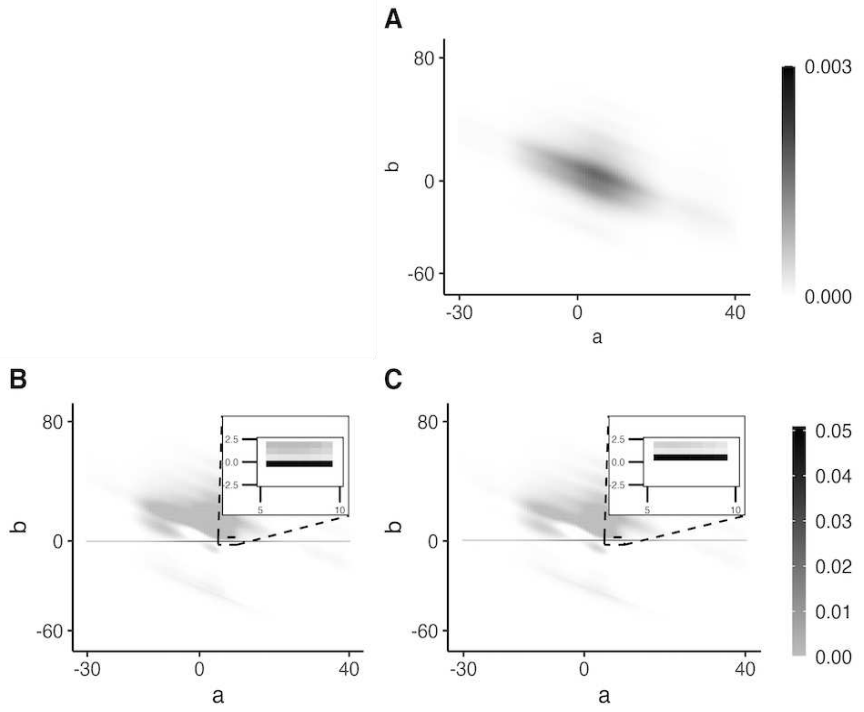

Figure 2.5: Application of the iterative method to the acupuncture trial data. Panel A shows the density heatmap of the approximated coefficient distribution $\hat{F}_{A,B}$ using a $\mathrm{N}(0, 25^2 \cdot I_2)$ global ansatz. Panels B and C show two constructed distributions which yield identical output distributions as $\hat{F}_{A,B}$ on $\mathcal{X} = \{0, 1\}$. In Panels B and C, the dark density band is beneath and above the horizontal line $b = 0$, respectively.

visualize the densities of two such distributions with the corresponding probabilities assigned to $\{B > 0\}$ as 0.223 and 0.987, respectively.

In summary, although the PATE of acupuncture as compared to control is estimated to be positive, the chance that a randomly selected individual benefits more from acupuncture than control may be as low as 22.3%. Incorporating random coefficients allows for inference beyond population means.

2.6 Discussion

RCR models have been widely used in a range of applications which allow for heterogeneity of individuals. When the design variable is fixed on a discrete lattice, we have shown in Theorem 1 that the joint distribution of the regression coefficients is nonidentifiable. Viewing the approximation of coefficient distributions as solving a collection of SIPs, we have developed an iterative

method that leverages prior assumptions on the structure of the coefficient distribution through a global ansatz. We have shown in Theorem 2 that each recursive update moves the global ansatz closer to an element in the solution set $\mathcal{E}$ in terms of KL-divergence, and that the output distributions converge in total variation to $\mathcal{F}_{\mathcal{X}}$. Through simulation studies, we have demonstrated that the proposed method outperforms existing RCR estimation methods. Another advantage of the proposed method is that the use of different global ansatzes and the construction method of Theorem 1 allows for exploration of the solution class, which can serve as a sensitivity analysis.

In this work, our focus is on simple RCR models with a single, fixed explanatory variable. From this setup, many natural extensions of the proposed method can be pursued. One such extension involves multiple RCR models incorporating several explanatory variables, where both x and B in model 2.1 become vectors. While the overall theoretical framework remains largely unchanged, implementing this extension requires a tailored computational method for efficient numerical approximation of the recursive update in (2.5) over many iterations. Another extension pertains to a random explanatory variable. In this case, a possible modification of (2.5) is

$$f_{A,B}^{(\ell+1)}(a,b) = f_{A,B}^{(\ell)}(a,b) \int_{\mathcal{X}} \frac{f_{Y|X}(a+bx \mid x)}{f_{Y|X}^{(\ell)}(a+bx \mid x)} f_X(x)\,\mathrm{d}x, \quad \text{for } \ell = 0, 1, \ldots.$$

Presented with observations of (Y, X), one can estimate $f_{Y|X}$ and f_X and apply the modified recursive update to $\hat{f}_{Y|X}$ and $\hat{f}_X$. This modified approach is left as a future direction.

In many applications, the probability of an event E of the random coefficients is of interest, which in general cannot be uniquely identified. Here, we advocate considering all elements in the solution set $\mathcal{E}$ rather than focusing on a solution based on a set of unverified assumptions.

Following similar notion as Dempster (1967) and Chi (2021), one may define lower and upper probabilities of E as

$$F_*(E) = \inf_{F_{A,B} \in \mathcal{E}} F_{A,B}(E) \qquad \text{and} \qquad F^*(E) = \sup_{F_{A,B} \in \mathcal{E}} F_{A,B}(E).$$

Given the convexity of $\mathcal{E}$, the interval $(F_*(E), F^*(E))$ delineates the plausible range of probabilities for event E. While the proposed method, coupled with the construction method outlined in the proof of Theorem 1, can be used to explore the elements of this probability interval, principled approaches for computing the infimum and supremum of the interval are worth further investigation.

The theoretical properties of the recursive update in (2.5) can be explored in more detail. Theorem 2 establishes the convergence of the sequence of KL-divergences of an arbitrary element of $\mathcal{E}$ from $f_{A,B}^{(\ell)}$. It is of interest to determine, possibly under additional regulatory conditions, the convergence of the sequence $f_{A,B}^{(\ell)}$ itself. It is also of interest to theoretically study the finite-sample and asymptotic behaviors of the recursive update when the true output densities $\{f_{Y_x} : x \in \mathcal{X}\}$ are replaced by the estimated output densities $\{\hat{f}_{Y_x} : x \in \mathcal{X}\}$. Lastly, bootstrap methods (Efron and Tibshirani, 1994) may be used to characterize the sampling variability of the proposed estimator of the coefficient distribution.

Chapter 3

Nonparametric Recovery of Mixing Distributions

3.1 Introduction

Let (X_i, Θ_i), $i = 1, 2, \ldots, n$, be a random sample from an unknown joint distribution. For each i, X_i is observed, and Θ_i is unobserved. The latent variables Θ_i can be viewed as iid realizations from an unknown marginal distribution G, known as a *mixing distribution*. We further assume that, conditioned on Θ_i, X_i has a probability distribution $P(\cdot|\Theta_i)$. The set of conditional distributions or the parametric family, $P(\cdot|\theta)$, indexed by θ, is known as the *component distribution family*. The *compound distribution* F can be expressed as

$$F(A) = \int P(A|\theta)\mathrm{d}G(\theta), \tag{3.1}$$

for each measurable set A. Here, F is a probability measure, and $F((-\infty, x])$ will yield the standard distribution function of X_i. Our goal is to estimate the mixing distribution G based on a sample of observations, $X_1, \ldots, X_n$.

The recovery of the mixing distribution has been widely used in many applications, such as information theory (Patrick and Hancock, 1966), actuarial applications (Karlis and Xekalaki, 2005), biomedical studies (Zhai and Jiang, 2023), and empirical Bayesian estimation (Yakowitz, 1969; Robbins, 1964). Feller (1943) discussed many applications for compound Poisson distributions, including entomology, bacteriology, and telephone traffic.

Many existing approaches assumed a parametric family for the mixing distributions, hence focused on parameter estimation. For instance, the well-known beta-binomial distribution has been used for modeling count data, which can be viewed as a compound distribution. Researchers proposed the maximum likelihood approach to estimate parameters in the beta distribution (Skellam, 1948; Barnard, 1954; Crowder, 1978). Similarly, a Lomax distribution can be viewed as a compound distribution, with an exponential component family mixed over the rate parameter according

to a gamma mixing distribution. Then modeling data with a Lomax distribution and estimating its shape and scale parameters, as done by Lomax (1954) for business failure data, can be viewed as estimating a parametric gamma mixing distribution.

Nonparametric approaches are common in the literature, however, in general the proposed approaches are designed for specific data scenarios. Laird (1978) proposed a nonparametric maximum likelihood estimation procedure, where G is a discrete distribution. Zhang (1990) developed a kernel estimation approach using Fourier methods. Other relevant nonparametric works include, Carroll and Hall (1988), Liu and Taylor (1989), Stefanski and Carroll (1990), Fan (1991), Hall and Meister (2007) and Butucea and Comte (2009). Those methods used a deconvolution of the kernel density estimator. Goutis (1997) pointed out that the deconvolution methods are essentially only applicable in the errors-in-variables model, where the differences $\Theta_i - X_i$ are independent. Consequently, it is often assumed that $X_i = \Theta_i + \epsilon_i$, where the ϵ_i's form a random sample from a location family, e.g., normal, and ϵ_i is independent of Θ_i for each i. The term deconvolution arises naturally since (3.1) can be written, when the component distribution family is a location family, in terms of densities with respect to appropriate dominating measures as

$$f(x) = \int p(x - \theta) \mathrm{d}G(\theta),$$

where $p(x - \theta)$ is the density of the a component distribution $P(\cdot|\theta)$. In particular, when the component distribution family is normal, the inference of unknown distribution G of the means is referred to as the Normal Means problem. For a review of recent developments under this specific scenario, see Jiang and Zhang (2009), Sun (2020), and Liu (2023).

Unlike the deconvolution methods and those proposed in the Normal Means problem, Goutis (1997) took a different approach where the mixing distribution is assumed to have a Lebesgue density. In particular, the author integrated a kernel density estimator for the unknown mixing distribution with an expectation over the posterior distribution. Because the latent variables Θ_i's are unobserved, an iterative EM-type algorithm is proposed. Their method is computationally intensive since surrogate values for the Θ_is are sampled at each iteration.

Efron (2014, 2016) proposed a new approach to approximate the mixing distribution, given a known component distribution family, by modeling the mixing distribution with an exponential family. Efron's approach has been used in a variety of analyses, including two-sample tests in medical studies (Zhai and Jiang, 2023). Efron (2014) categorized most approaches for the recovery of a mixing distribution into two types: g-modeling and f-modeling. In particular, g-modeling approaches focus on the mixing distribution in (3.1) directly, while, f-modeling approaches model it indirectly through the compound distribution. The author proposed two f-modeling approaches, one parametric and one nonparametric, and the g-modeling approach mentioned before. Later, Efron (2016) compared the g-modeling approach with a nonparametric f-modeling approach by Stefanski and Carroll (1990), which demonstrated the advantage of modeling the mixing distribution with an exponential family.

Recently, Cui and Hannig (2022) proposed a fiducial alternative to Efron's g-modeling approach for discrete component distributions, e.g., a family of Poisson distributions indexed by a mixing random variable. Through simulations, it was demonstrated that the fiducial approach is a good alternative to g-modeling. However, the fiducial approach is not applicable to continuous component distributions, such as the normal family commonly assumed in the deconvolution methods.

In general, the problem of recovering the mixing distribution may be *ill-posed* or not *well-posed*. Mathematically speaking, a well-posed problem implies that the solution exists, is unique, and is stable; see Hadamard (1902), Tikhonov et al. (1995), and Strauss (2008) for more detailed discussions. If the solution is unique, the mixing distribution is said to be *identifiable*. The identifiability of the mixing distribution has been widely studied (Teicher, 1960, 1963; Tallis, 1969; Yakowitz, 1969; Atienza et al., 2006). Generally, the mixing distribution is nonidentifiable, but under certain scenarios, it is identifiable. In addition, the recovery problem may not be stable. To our knowledge, little attention has been given to the stability, or lack thereof, of the mixing distribution recovery problem. In Section 3.2.3, we demonstrate the instability of the recovery problem for a scenario where the solution exists and is unique.

In this paper, we present a novel framework to handle the ill-posedness of the mixing distribution recovery problem. In particular, we propose a nonparametric method to approximate elements within a set of mixing distributions, that induce the same compound distribution, according to (3.1), for a given set of component distributions. Our proposed approach, initiated by a user-specified distribution, namely, an *ansatz*, iteratively updates a distribution through integration and disintegration, in such a way that at each iteration, the distribution steps closer to the set of distributions of the aforementioned set.

Under scenarios where the recovery of the mixing distribution is well-posed, the selection of an ansatz matters little. For cases where the solution is unique, but the problem suffers from instability, a judicious selection of an ansatz acts to stabilize the problem. Finally, when the solution is not unique, different choices of ansatzes may be used to approximate different elements within the equivalence class with certain features, e.g., continuous or discrete distributions.

The paper is organized as follows. In Section 3.2, the background of the mixing distribution recovery problem and its ill-posed nature are discussed. In Section 3.3, we offer a new perspective to the ill-posed nature of the problem. Furthermore, we propose an iterative method and study its convergence properties. In Section 3.4, our proposed method is compared with Efron's *g*-modeling approach, when the mixing distribution is identifiable. In addition, we demonstrate the performance of our approach under a scenario in which the mixing distribution is nonidentifiable. Finally, our method is applied to two real applications in Section 3.5: a prostate cancer dataset (Singh et al., 2002; Efron, 2016) and the Shakespeare's word count dataset (Spevack, 1968; Narasimhan and Efron, 2020).

3.2 Mixing Distribution Recovery

3.2.1 Background

We will begin with a review of the measure-theoretic treatment of the mixing distributions. Consider the probability distribution $F_{X,\Theta}$ on the product space $(\mathcal{X} \times \mathcal{T}, \mathcal{B}_{\mathcal{X}} \otimes \mathcal{B}_{\mathcal{T}})$, where $\mathcal{X} \subset \mathbb{R}$,

$\mathcal{T} \subset \mathbb{R}^k$, for some integer $k \geq 1$, and where $\mathcal{B}_\mathcal{X}$ and $\mathcal{B}_\mathcal{T}$ are the Borel σ-algebras on $\mathcal{X}$ and $\mathcal{T}$, respectively.

Let G be the marginal distribution of Θ. Faden (1985) characterized the relationship between the measure $F_{X,\Theta}$ and the marginal distribution G through a kernel, the existence of which is guaranteed by the product regular conditional probability property of the product space $(\mathcal{X} \times \mathcal{T}, \mathcal{B}_\mathcal{X} \otimes \mathcal{B}_\mathcal{T})$. In particular, there exists a probability kernel $\nu : \mathcal{T} \times \mathcal{B}_\mathcal{X} \to [0, 1]$ such

$$F_{X,\Theta}(A, B) = \int_B \nu(\theta, A) \mathrm{d}G(\theta) \tag{3.2}$$

for $A \in \mathcal{B}_\mathcal{X}$ and $B \in \mathcal{B}_\mathcal{T}$. Taking $B = \mathcal{T}$, (3.1) is a special case of (3.2), in which $F(A) = F_{X,\Theta}(A, \mathcal{T})$. In addition, the component distributions $P(\cdot|\theta)$ can be defined by the probability kernel ν as $P(A|\theta) = \nu(\theta, A)$ for all $A \in \mathcal{B}_\mathcal{X}$. We denote the set of component distributions $\mathcal{P} = \{P(\cdot|\theta) : \theta \in \mathcal{T}\}$.

We further note that, in (3.2), for a given measure $F_{X,\Theta}$ on the product space and a probability kernel ν, the marginal distribution G can be uniquely identified. In contrast, our mixing distribution recovery problem is to find G for a given marginal distribution F, instead of $F_{X,\Theta}$, and a known probability kernel. Notably, a solution to our recovery problem may not exist. In this paper, we assume that there exists a joint distribution $F_{X,\Theta}$, and hence, the marginal distribution of Θ is a solution to our recovery problem.

In Sections 3.2.2 and 3.2.3 we will discuss the uniqueness and the stability of the mixing distribution recovery problem.

3.2.2 Nonidentifiability

The mixing distribution recovery problem may be ill-posed since that the solution may not be unique. In the context of our recovery problem, uniqueness of a solution is referred to as identifiability. In the literature, it is well known that the mixing distribution faces challenges of identifiability (Teicher, 1960, 1963; Tallis, 1969; Yakowitz, 1969; Atienza et al., 2006). We will begin our discussion with the following examples.

Example 3.2.1. *One-parameter case*. Let the component distribution family be the set of Bernoulli distributions with the success probability θ, $\theta \in [0, 1]$. A continuous mixing distribution G_1, a uniform distribution on $[0, 1]$, and a discrete mixing distribution G_2, with $G_2(\{0.25\}) = G_2(\{0.75\}) = 0.25$ and $G_2(\{0.5\}) = 0.5$, lead to the same compound distribution.

Example 3.2.2. *Two-parameter case*. Let the component distribution family be the set of normal distributions indexed by $\theta = (\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ is the mean and $\sigma^2 > 0$ is the variance. The mixing distribution functions G_3, a discrete distribution that assigns a probability mass of 0.25 to each of the four tuples in $\{-1, 1\} \times \{1, 4\}$, and G_4, defined by

$$G_4\left((-\infty, \mu], (-\infty, \sigma^2]\right) = \begin{cases} 0 & \sigma^2 < 0.5 \\ 0.25\left(\Phi(2\mu - 2) + \Phi(2\mu + 2)\right) & 0.5 \leq \sigma^2 < 3.5 \\ 0.5\left(\Phi(2\mu - 2) + \Phi(2\mu + 2)\right) & \sigma^2 \geq 3.5 \end{cases},$$

induce the same compound distribution. Here, Φ is the distribution function of the standard normal distribution. Straightforward calculation yields that the compound distribution has a density function defined by

$$F((-\infty, x]) = \frac{1}{4}\left[\Phi(x - 1) + \Phi(x + 1) + \Phi\left(\frac{x - 1}{2}\right) + \Phi\left(\frac{x + 1}{2}\right)\right].$$

Examples 1 and 2 demonstrate that the mixing distribution G cannot always be uniquely identified. Such a problem is often referred to as nonidentifiability. Here, we adopt the definition of identifiability of the mixing distribution as provided in the literature (Teicher, 1960, 1961, 1963; Tallis, 1969; Karlis and Xekalaki, 2005). Generally, the mixing distribution is identifiable, with respect to the component distribution family, if

$$\int P(A|\theta)\mathrm{d}G(\theta) = \int P(A|\theta)\mathrm{d}\widetilde{G}(\theta) \text{ for all } A \in \mathcal{B}_{\mathcal{X}}$$

implies that $G(B) = \widetilde{G}(B)$ for all $B \in \mathcal{B}_{\mathcal{T}}$. That is, for the component distribution family, there is a one-to-one correspondence between the mixing distribution G and its induced compound distribution F.

The identifiability of G depends on the component distribution family. Teicher (1961) showed that if the set of component distributions form a single-parameter, additively closed class, then the mixing distribution is identifiable. Examples include a family of normal distributions varying over the mean parameter and with fixed variance, a Poisson family varying over the mean, and a Binomial family varying of the number trials, but where the probability parameter is fixed. Evident from Example 3.2.1, if the Binomial family varies over the probability parameter with a fixed number of trials, the mixing distribution may not be identifiable. Teicher (1961) also gave a sufficient condition for identifiability when the component distributions form a scale family related to the Fourier transform of the generating distribution of the scale family.

Teicher (1963), Yakowitz (1969), and Atienza et al. (2006) showed sufficient conditions for identifiability of the compound distributions under finite mixtures. For a finite mixture, the mixing distribution has a cumulative distribution function (cdf) with finitely many jump discontinuities. In particular, Teicher (1963) showed that, for a normal component distribution family, the finite discrete mixing distribution is identifiable when mixing over both mean and variance parameters. As a direct consequence, in Example 3.2.2, G_3 is the only finite discrete mixing distribution that induces the compound distribution F. However, there exist other distributions, which are not finite discrete, that induce the same compound distribution, such as G_4.

Tallis (1969) gave necessary and sufficient conditions for countably infinite mixtures through characteristics of infinite systems of equations defined by the component distributions. Furthermore, the same author characterized identifiability for mixtures, under mild conditions of the component distributions, through infinite sums of the eigen values derived from a symmetric kernel defined by the component distributions, and completeness of the set of associated eigen vectors.

3.2.3 Instability

The mixing distribution recovery may also face the challenge of instability. That is, small perturbations in the compound distribution may lead to drastically different mixing distributions. We will elaborate on this challenge through a real-world application.

Singh et al. (2002) conducted a microarray expression analysis to identify genes that may be associated with prostate cancer. The dataset consists of 6033 t-statistics, $\{T_1, T_2, \ldots, T_{6033}\}$, where T_i is calculated for gene i between the groups of 52 men with cancer and 50 men without cancer. Efron (2010, 2016) transformed T_i as $X_i = \Phi^{-1}(F_{100}(T_i))$, where F_{100} is the cdf of the t distribution with 100 degrees of freedom. Therefore, each X_i is approximately normally distributed with mean Θ_i and variance 1. The unobserved value Θ_i quantifies the difference in expressions of gene i between those two groups; in particular, $\Theta_i = 0$ indicates that there is no expression difference for gene i (a null-gene) between the two groups of men.

Following Efron (2010, 2016), it is assumed each Θ_i follows the same unknown distribution G. To capture the proportion of null-genes, it is assumed that the cdf of G is continuous except potentially at 0 where a jump discontinuity would imply a positive probability assigned to the event $\{\Theta_i = 0\}$.

The unknown distribution G is the mixing distribution of interest, the component distributions form a normal family with parameter θ and a fixed variance equal to 1, and F is the compound distribution, according to which each X_i is distributed. Note that the component distribution family meets identifiability criteria, i.e, G is identifiable. We will return to this example in Section 3.5, but for now we demonstrate that estimating G, given its cdf has a jump discontinuity at $\theta = 0$, is unstable.

Consider a set of distributions $\{G_\pi\}$ for $\pi \in \{0.80, 0.82, \ldots, 0.90\}$, where

$$G_\pi((-\infty, \theta]) = \pi \mathbb{I}(\theta \geq 0) + (1 - \pi)\Phi(\theta).$$

Note that π represents the probability assigned to $\{\Theta = 0\}$. Panel A of Figure 3.1 visualizes the cdfs of the distributions G_π. Clearly, the cdfs differ, in particular, around the jump discontinuity.

For the same normal component distributions as in the prostate cancer example, for each π, the compound distribution F_π induced by G_π is

$$F_\pi\left((-\infty, x]\right) = \pi\Phi(x) + (1-\pi)\Phi\left(\frac{x}{\sqrt{2}}\right).$$

The cdfs of the compound distributions are shown in Panel B of Figure 3.1. In contrast to the mixing distributions, the cdfs of the compound distributions are visually indistinguishable.

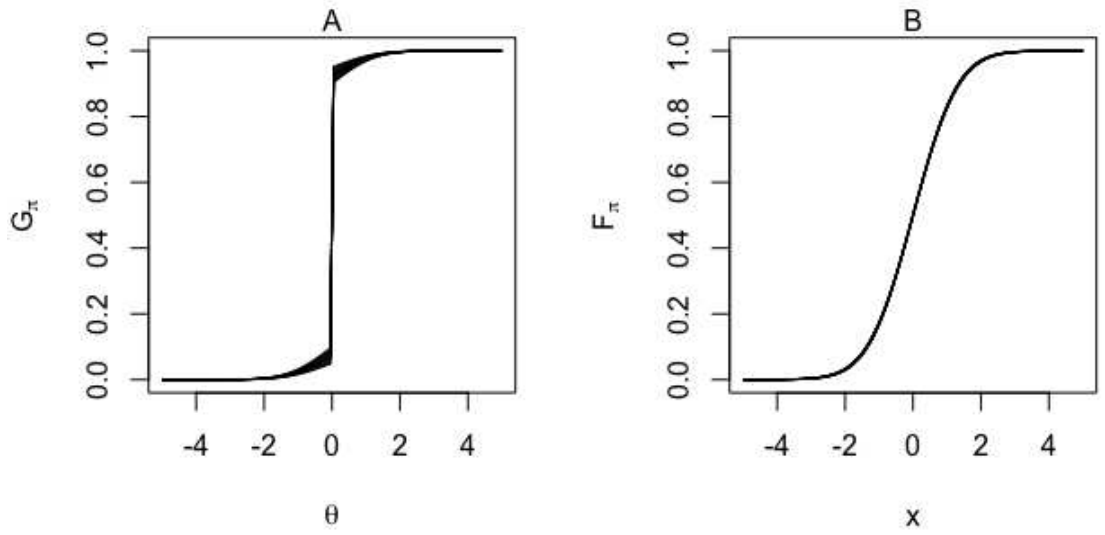

Figure 3.1: `Panel A`: Distribution functions for G_π, where $\pi = G_\pi(\{0\})$ ranges from 0.80 to 0.90. `Panel B`: The distribution functions for the induced distributions F_π. Note that they are overlapped.

In particular, we quantify the instability in this example by calculating the scaled L_2 diameters of $\{G_\pi\}$ and $\{F_\pi\}$ respectively, which are the maximum L_2 distances (scaled by 10000) between two distributions in a set. The scaled diameters are 32.1 and 0.5 for the mixing distribution and

compound distribution sets, which highlights the instability of the solutions to the mixing distribution recovery problem.

3.3 Methodology

3.3.1 Equivalence Classes: a New Perspective

As discussed in Section 3.2, the solutions to the mixing distribution recovery problem in (3.1) may not be unique. Here, we define $\mathcal{E}_{F,\mathcal{P}}$ as the set of solutions for a given compound distribution F and a component distribution family $\mathcal{P}$; that is,

$$\mathcal{E}_{F,\mathcal{P}} = \left\{ G : F(A) = \int P(A|\theta) \mathrm{d}G(\theta) \text{ for all } A \in \mathcal{B}_{\mathcal{X}} \right\}.$$

Note that $\mathcal{E}_{F,\mathcal{P}}$ can be viewed as an equivalence class with respect to the following equivalence relation. Distributions G and $\widetilde{G}$ are considered equivalent if they induce the same compound distribution for a given component distribution family.

If $\mathcal{E}_{F,\mathcal{P}}$ is a singleton set, then the mixing distribution G is identifiable. However, in the non-identifiable case, $\mathcal{E}_{F,\mathcal{P}}$ has more than one element. In fact, it is easy to verify that $\mathcal{E}_{F,\mathcal{P}}$ is closed under convex combinations. That is, if $G, \widetilde{G} \in \mathcal{E}_{F,\mathcal{P}}$, $\alpha G + (1-\alpha)\widetilde{G} \in \mathcal{E}_{F,\mathcal{P}}$ for all $\alpha \in [0,1]$. Therefore, the cardinality of $\mathcal{E}_{F,\mathcal{P}}$ can be zero, one, or uncountably infinite. In this paper, it is assumed that $\mathcal{E}_{F,\mathcal{P}}$ is non-empty.

When the solution to the mixing distribution recovery problem is not unique, and without scientific or domain justification, there is no reason to favor one solution in $\mathcal{E}_{F,\mathcal{P}}$ over another. This can be problematic for inference on events in the space $(\mathcal{T}, \mathcal{B}_{\mathcal{T}})$. To elaborate this point, let us revisit Example 3.2.1 from Section 3.2. Recall that for the Bernoulli component distribution family, the compound distribution F, induced by G_1 and G_2, is a Bernoulli distribution with success probability 0.5. In fact, if $\mathcal{T} = [0,1]$, then $\mathcal{E}_{F,\mathcal{P}}$ consists of all distributions with mean 0.5. Then, clearly knowledge of $\mathcal{P}$ and F gives no information about the probability of the event $\{\Theta = 0.5\}$, ranging from 0 to 1. If we restrict our attention on distributions supported on $\{0.25, 0.5, 0.75\} \subset \mathcal{T}$,

the restriction of $\mathcal{E}_{F,\mathcal{P}}$ can be further characterized as distributions $\widetilde{G}$, satisfying $\widetilde{G}(\{0.25\}) = \widetilde{G}(\{0.75\}) \in [0, 0.5]$ and $\widetilde{G}(\{0.5\}) = 1 - 2\widetilde{G}(\{0.25\})$.

In general, when we restrict our attention on mixing distributions supported over a finite set, a subset of $\mathcal{E}_{F,\mathcal{P}}$, it is straightforward to characterize the distributions in the subset. This may be of interest to the reader, as there are precedent studies on finite mixtures (Teicher, 1963; Yakowitz, 1969; Heinrich and Kahn, 2018).

Suppose $\mathcal{X} = \{x_1, \ldots, x_L\}$ and $\mathcal{T} = \{\theta_1, \ldots, \theta_M\}$. Our goal is to identify the mixing distributions with support $\mathcal{T}$ satisfying (3.1). Let f, g, and $p(\cdot|\theta)$ be the probability mass functions for F, G, and $P(\cdot|\theta)$, respectively. Then (3.1) can be expressed as

$$\boldsymbol{f} = \boldsymbol{P}\boldsymbol{g}, \tag{3.3}$$

where $\boldsymbol{P}$ is an $L \times M$ matrix with elements $p_{lm} = p(x_l|\theta_m)$, $\boldsymbol{g}$ is an M-dimensional vector with elements $g_m = g(\theta_m)$, and $\boldsymbol{f}$ is an L-dimensional vector with elements $f_l = f(x_l)$. In addition, we have $\sum_m g_m = \sum_l f_l = 1$.

Assume that $\boldsymbol{g}_0$ is a solution of (3.3), and is element-wise strictly greater than 0. If $\boldsymbol{P}$ has rank M, $\boldsymbol{g}_0$ is the unique solution to (3.3). However, if $\boldsymbol{P}$ has rank less than M then (3.3) has more than one solution. That is, $\mathcal{E}_{F,\mathcal{P}}$ has infinitely many elements, and we have

$$\mathcal{E}_{F,\mathcal{P}} = \left\{\boldsymbol{g} \geq 0 : \boldsymbol{f} = \boldsymbol{P}\boldsymbol{g}, \mathbf{1}^\top \boldsymbol{g} = 1\right\}.$$

The elements in $\mathcal{E}_{F,\mathcal{P}}$ can be expressed as $\boldsymbol{g} = \boldsymbol{g}_0 + \alpha\widetilde{\boldsymbol{g}}$, where $\widetilde{\boldsymbol{g}}$ is in the null space of $\boldsymbol{P}$, and α is a constant to ensure the nonnegativity of $\boldsymbol{g}$ entrywise.

Efron (2014, 2016) and Narasimhan and Efron (2020) considered a similar case as $L \geq M$, and further assumed that $\boldsymbol{g}$ can be approximated by a q-parameter exponential family. Those authors further proposed to maximize the regularized likelihood for parameter estimation.

However, the characterization of $\mathcal{E}_{F,\mathcal{P}}$ in terms of the null-space of $\mathcal{P}$ requires discrete component distributions, $\mathcal{T}$ to be a discrete lattice, and the knowledge of at least one distribution in

$\mathcal{E}_{F,\mathcal{P}}$. To better facilitate exploration of the solution set more generally, in Section 3.3.2 we present a method to recover different mixing distributions in $\mathcal{E}_{F,\mathcal{P}}$, according to a user-specified ansatz. In particular, one can change the user input to obtain different elements in $\mathcal{E}_{F,\mathcal{P}}$. In addition, if there is scientific or domain knowledge about the mixing distribution, it can be incorporated into the ansatz to help pick out a specific solution in $\mathcal{E}_{F,\mathcal{P}}$. In such a scenario, the ansatz can be used similar to a prior in Bayesian inference with nonidentifiable parameters; see Neath and Samaniego (1997) and Wechsler et al. (2013).

3.3.2 Integration-Disintegration Method

Here we present a nonparametric approach to recover a mixing distribution G in $\mathcal{E}_{F,\mathcal{P}}$. In developing the method, we first consider both the component distribution family $\mathcal{P}$ and the compound distribution F to be known. In practice, we only have a random sample of data from the compound distribution F. Consequently, we will later drop the assumption that F is known, and only consider $\mathcal{P}$ to be known. An estimate of F from the observed sample $\{X_i\}$ can be used instead.

Let us start with an initiating probability distribution G^0, which dominates some probability measure in $\mathcal{E}_{F,\mathcal{P}}$. We call the distribution G^0 an ansatz. Breidt et al. (2011), Butler et al. (2012, 2014, 2018), and Chi (2021) used the term ansatz to refer to a priori knowledge which helps facilitate solving stochastic inverse problems. We use the term ansatz similarly, as it helps facilitate solving the mixing distribution recovery problem.

Recall that ν is the probability kernel that defines $\mathcal{P}$. Then, for any distribution G^0 on $(\mathcal{T}, \mathcal{B}_{\mathcal{T}})$, we can construct a distribution $F^0_{X,\Theta}$ on $(\mathcal{X} \times \mathcal{T}, \mathcal{B}_{\mathcal{X}} \otimes \mathcal{B}_{\mathcal{T}})$ with ν and G^0 through the following *integration step*

$$F^0_{X,\Theta}(A, B) = \int_B \nu(\theta, A) \mathrm{d}G^0(\theta).$$

Note that the X-marginal distribution of $F^0_{X,\Theta}$, denoted by F^0, is not necessarily equal to the compound distribution F. There exists a probability kernel $\xi^0 : \mathcal{X} \times \mathcal{B}_{\mathcal{T}} \to [0, 1]$ for the constructed joint distribution $F^0_{X,\Theta}$. Its existence is a direct consequence of the fact that $\mathcal{B}_{\mathcal{T}}$ is countably generated and $\mathcal{T}$ is a subset of Euclidean space (Faden, 1985). Furthermore, the kernel can be

obtained through a *disintegration step*

$$F^0_{X,\Theta}(A, B) = \int_A \xi^0(x, B)\mathrm{d}F^0(x). \tag{3.4}$$

The disintegration theorem has been widely used in measure theory to restrict a measure to a given set of measure zero. Chang and Pollard (1997) drew the connection between conditional probability and disintegration.

As such, in (3.4) the probability kernel ξ^0 can be viewed as a (F^0, Q)-disintegration of $F^0_{X,\Theta}$, where the map $Q : \mathcal{X} \times \mathcal{T} \to \mathcal{X}$ is the coordinate projection $Q(x, \theta) = x$. As the disintegrating map Q will always be the coordinate projection in this paper, we will omit Q and simply refer to ξ^0 as the F^0-disintegration of $F^0_{X,\Theta}$.

From ξ^0 and F, we can obtain a new distribution G^1 on $(\mathcal{T}, \mathcal{B}_\mathcal{T})$ as

$$G^1(B) = \int_\mathcal{X} \xi^0(x, B)\mathrm{d}F(x).$$

It is important to note that G^1 is not necessarily a solution because $\int P(A|\theta)dG^1(\theta)$, in general, is not equal to $F(A)$. However, G^1 is an *update* of G^0 in the sense that, as will be shown later, the induced compound distribution from G^1 is closer to $F(A)$ than the induced compound distribution from G^0. With G^1 as our newly updated ansatz, we can repeat the three-step procedure, including integration, disintegration, and update, to approximate an element in $\mathcal{E}_{F,\mathcal{P}}$.

Our proposed iterative procedure can be summarized as follows. For each iteration $j = 0, 1, \ldots$, the iterative procedure can be categorized into the following three steps.

1. (*Integration*) Construct $F^j_{X,\Theta}$ from ν and G^j as

$$F^j_{X,\Theta}(A, B) = \int_B \nu(\theta, A)\mathrm{d}G^j(\theta).$$

2. (*Disintegration*) Disintegrate $F^j_{X,\Theta}$ with respect to F^j, the X-marginal of $F^j_{X,\Theta}$, to obtain ξ^j, such that

$$F^j_{X,\Theta}(A, B) = \int_A \xi^j(x, B)\mathrm{d}F^j(x).$$

3. (*Ansatz update*) Compute the updated ansatz from ξ^j and F as,

$$G^{j+1}(B) = \int_{\mathcal{X}} \xi^j(x, B)\mathrm{d}F(x).$$

We name the iterative procedure, the *integration-disintegration* (ID) method, as we consecutively integrate and disintegrate measures to step closer to a mixing distribution in $\mathcal{E}_{F,\mathcal{P}}$. We note that, when both $\mathcal{P}$ and F are known, the ID method is similar to that proposed by Kullback (1968), but instead of matching only marginals, we aim to match a marginal and a probability kernel. Additionally, we do not restrict ourselves to Lebesgue densities.

The ID method can be written conveniently and concisely in terms of density functions. Let μ and γ be σ-finite measures on $(\mathcal{X}, \mathcal{B}_{\mathcal{X}})$ and $(\mathcal{T}, \mathcal{B}_{\mathcal{T}})$, respectively, where μ dominates F^0 and γ dominates G^0. Define a product measure on $(\mathcal{X} \times \mathcal{T}, \mathcal{B}_{\mathcal{X}} \otimes \mathcal{B}_{\mathcal{T}})$ as $\lambda = \mu \times \gamma$, which dominates $F^0_{X,\Theta}$. Then, the Radon-Nikodym derivative of $F^0_{X,\Theta}$ with respect to λ, i.e., the λ-density (or simply density) of $F^0_{X,\Theta}$ is

$$f^0(x, \theta) = p(x|\theta)g^0(\theta),$$

where $p(\cdot|\theta)$ and g^0 are the μ- and γ-densities of $P(\cdot|\theta)$ and G^0, respectively. Consequently, we can write the ID method in terms of density functions as, for $j = 0, 1, \ldots$

$$g^{j+1}(\theta) = g^j(\theta) \int \frac{f(x)}{f^j(x)} P(\mathrm{d}x|\theta) = g^j(\theta)\mathrm{E}_{P(\cdot|\theta)}\left[\frac{f}{f^j}\right]. \tag{3.5}$$

That is, evaluated at θ, the new ansatz density is the product of the previous ansatz density and a multiplicative factor, which is the expectation of a ratio between compound densities. Straightforward calculation shows that $g^j(\theta) \geq 0$ and that $\int g^j(\theta)\mathrm{d}\mu(\theta) = 1$, for all j.

Next, we will demonstrate that G^{j+1} from (3.5) is as close as, or closer than, G^j to an element of $\mathcal{E}_{F,\mathcal{P}}$. To quantify the closeness of G^j to G, we use the Kullback-Leibler (KL) divergence, denoted as $\mathcal{D}_{KL}(G||G^j)$. The following theorem describes the convergence of our procedure under mild conditions, the proof of which is included in the Appendix B.0.1.

Theorem 3. Let G be an element of the solution set $\mathcal{E}_{F,\mathcal{P}}$ determined by F and $\mathcal{P}$, and let G^0 be an ansatz and assume that

$$\mathcal{D}_{KL}(G||G^0) < \infty.$$

Further, let $\{G^j\}$ be the sequence of distributions defined by (3.5) and initiated by G^0, and let $\{F^j\}$ be the sequence of distributions induced by $\{G^j\}$ and $\mathcal{P}$ through (3.1). Then

(i) $\mathcal{D}_{KL}(G||G^j)$ forms a monotonically non-increasing sequence, hence has a limit;

(ii) $\mathcal{D}_{KL}(F||F^j) \to 0$ as $j \to \infty$.

Note that Theorem 3 establishes that the KL-divergence of F from F^j converges to zero, and thus, by Pinkser's inequality F^j converges to F in total variation. Suppose that if the KL-divergences of F^j from F also converge to zero, then, as established in the following theorem, the sequence of KL-divergences of G^j from G^{j+1} converges to zero.

Theorem 4. Under the assumptions of Theorem 3, if $\mathcal{D}_{KL}(F^j||F) \to 0$ as $j \to \infty$, then $\mathcal{D}_{KL}(G^j||G^{j+1}) \to 0$ as $j \to \infty$.

The proof of Theorem 4 is in Appendix B.0.2. A necessary condition for $\mathcal{D}_{KL}(F^j||F) \to 0$, is that for some J and all $j > J$, F also dominates F^j. Note that the convergence of the KL divergences of G^j from G^{j+1} to zero does not guarantee that the sequence $\{G^j\}$ has a limit. With additional assumptions, we establish the convergence of $\{G^j\}$ in the following theorem.

Theorem 5. Under the assumptions of Theorem 3 and 4, we define S_j, for $j = 0, 1, \ldots$, as

$$S_j = \lim_{k\to\infty} \sum_{\ell=1}^{k} \left| \mathcal{D}_{KL}(F^j||F) - \mathcal{D}_{KL}(F^j||F^{j+\ell}) \right|.$$

If $\{S_j\}$ forms a monotonically decreasing sequence whose limit is zero, then G^j converges to a probability distribution G^∞ in total variation. Furthermore, if for any $A \in \mathcal{B}_\mathcal{X}$, $P(A|\theta)$ is a continuous function on $(\mathcal{T}, \mathcal{B}_\mathcal{T})$, in terms of θ, then G^∞ belongs to $\mathcal{E}_{F,\mathcal{P}}$.

The proof is included in Appendix B.0.3. Note that, if $\mathcal{E}_{F,\mathcal{P}}$ is a singleton with element G, under the assumptions of Theorem 5, G^j converges in total variation to G, the unique mixing distribution that induces F through $\mathcal{P}$.

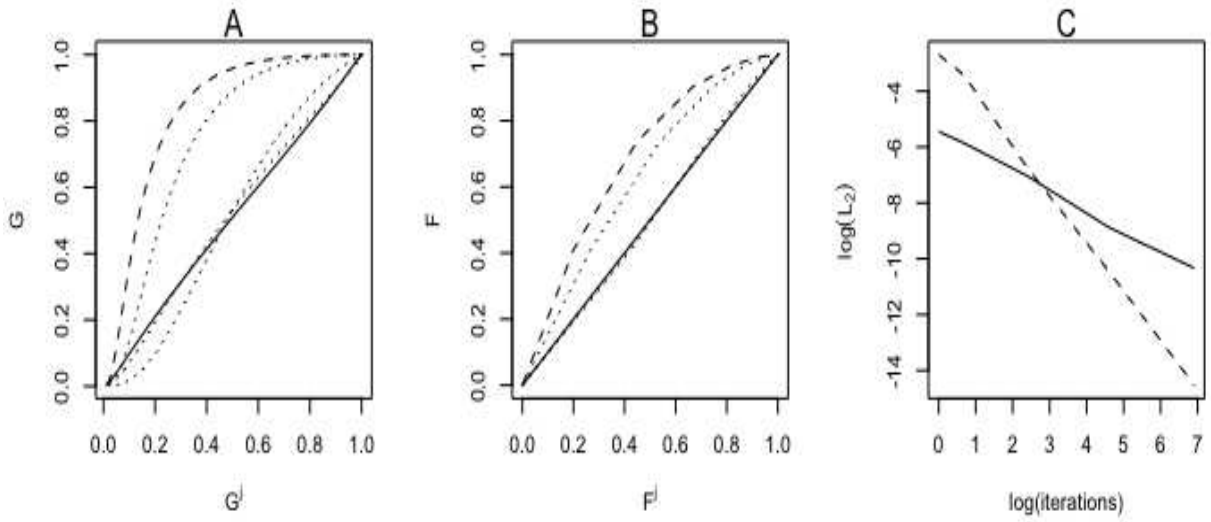

Figure 3.2: `Panel A:` The distribution functions for G^j, $j = 0, 1, 10, 100, 1000$ against the distribution function G. The solid line depicts G^{1000}, the dashed line is the ansatz, and the dotted lines are the remaining G^j. `Panel B:` The induced distributions functions of the induced F^j measures against the distribution function of F. `Panel C:` The log L_2 distances between G and G^j (solid line) and between F and F^j (dashed line) against the $\log(j + 1)$.

We demonstrate the ID method in Figure 3.2. Here, the mixing distribution G is a gamma distribution with shape 4 and rate 4, and the component distribution family consists of Poisson distributions with mean $\theta > 0$. The resulting compound distribution is a negative binomial with shape 4 and rate 0.8. We take the ansatz to be a folded normal with location 2 and scale 1. In Panel A, the cumulative distribution functions of G^j, defined by the ID method, are plotted against the distribution function of G. The closer the curves are to a straight line with slope one, the

closer the distribution functions match. The ansatz, depicted by the dashed curve, can be seen to be fairly different from G. The dotted curves, representing G^1, G^{10}, and G^{100}, indicate that G^j is approaching G. The solid curve for G^{1000} is nearly a straight line with slope 1, indicating that G^{1000} is a close approximation of G. Similarly, Panel B plots the cumulative distribution functions of F^j, induced by G^j, against the distribution function of the compound distribution F. Again, the ansatz G^0 does not induce a compound distribution F^0 that approximates F well. However, it can be seen that the induced compound distributions $\{F^j\}$ approach F, as depicted by the dotted curves approaching the solid curve, and the solid curve is nearly a straight line with slope 1. Panel C plots the log L_2 distance between G and G^j (solid curve) and between F and F^j (dashed curve) against $\log(j+1)$. Both curves are nearly straight lines with negative slopes, indicating the L_2 distances approach zero.

The recursive formula in (3.5) requires knowledge of F, the compound distribution. In practice, however, we observe a random sample, $X_1, X_2, \ldots, X_n$, from F. Standard estimation techniques lead to an estimate of F, denoted as $\hat{F}_n$. The empirical form of (3.5) is

$$\hat{g}_n^{j+1}(\theta) = \hat{g}_n^j(\theta) \int \frac{\hat{f}_n(x)}{\hat{f}_n^j(x)} P(\mathrm{d}x|\theta) \tag{3.6}$$

where $\hat{g}_n^0 = g^0$, $\hat{f}_n$ is μ-density of $\hat{F}_n$, and $\hat{f}_n^j(x) = \int p(x|\theta)\hat{g}_n^j(\theta)\mathrm{d}\theta$. The multiplicative factor is calculated through numerical integration since the component distributions are known. An alternative is to use sampling techniques, such as importance sampling (Hammersley and Morton, 1954; Tokdar and Kass, 2010). Additionally, for computational purposes, $\mathcal{T}$ can be approximated by a dense set of discrete points $\{\theta_1, \ldots, \theta_M\}$; see Efron (2016) and Narasimhan and Efron (2020). In the next section, computer simulations show that the ID method compares favorably to Efron's g-modeling.

3.4 Simulation

We assess the proposed ID method through simulation studies under four different scenarios. In the first three scenarios, we consider component distribution families that are single-parameter and additively closed. In such cases, the mixing distribution is identifiable from the compound distribution (Teicher, 1961). In the fourth scenario, we consider a nonidentifiable case with a bivariate mixing distribution. The details of the simulation scenarios and the specifics of the ID method are provided in subsequent subsections. To account for sampling variability, 100 datasets are generated under each simulation scenario.

We use the mean integrated squared error (MISE; Fryer, 1976) as a metric to evaluate the performance of the ID method. The MISE in recovering the mixing distribution is referred to as the mixing MISE (mMISE), while the MISE in recovering the induced compound distribution is referred to as the compound MISE (cMISE). On each simulated dataset, an estimate $\hat{g}_n^\ell$ of the mixing density is obtained by iteratively updating the ansatz via (3.6). For scenarios 1 and 2, we set $\ell = 100$ iterations. For scenarios 3 and 4, and for computational purposes, we lower the number of iterations to $\ell = 25$ and $\ell = 15$, respectively. The estimated mixing density induces an estimated compound density through $\hat{f}_n^\ell(x) = \int p(x \mid \theta)\hat{g}_n^\ell(\theta)\mathrm{d}\theta$. The following integrated squared errors are calculated,

$$\mathrm{E}\int \big(g(\theta) - \hat{g}_n^\ell(\theta)\big)^2 \,\mathrm{d}\gamma(\theta), \quad \text{and} \quad \mathrm{E}\int \Big(f(x) - \hat{f}_n^\ell(x)\Big)^2 \,\mathrm{d}\mu(x). \tag{3.7}$$

The mMISE and cMISE are estimated by averaging the above quantities over the 100 simulated datasets. In situations where the mixing distribution lacks identifiability, the mMISE is not a meaningful performance metric, and only the cMISE is considered.

For comparison, we also implement the g-modeling approach in Efron (2016) using the R package `deconvolveR` (Narasimhan and Efron, 2020). For numerical implementation of the ID and g-modeling approaches, following the strategy in Efron (2016) and Narasimhan and Efron (2020), the support of Θ_i is approximated by a discrete set.

3.4.1 Scenario 1

In the first scenario, we consider a gamma mixing distribution with shape parameter 4 and rate parameter 4, which belongs to the exponential family. The data generating process is

$$\Theta_i \overset{iid}{\sim} \text{Gamma}(4, 4), \qquad X_i \mid \Theta_i \overset{ind.}{\sim} \text{Poisson}(\Theta_i).$$

The component distribution family is the set of Poisson distributions indexed by the mean parameter. Hence, the mixing distribution is identifiable. The sample size of each simulated dataset is $n = 1000$.

The compound distribution is estimated by the empirical distribution, $\hat{F}_n(\{x\}) = \#\{X_i = x\}/n$. The discretized support of Θ_i is taken to be $\hat{\mathcal{T}} = (0.1, 0.2, \ldots, 5.0)$ in the numerical calculations. In this scenario, we find that the choice of the ansatz matters little. Therefore, we only present the results with the ansatz being a discrete uniform distribution over $\hat{\mathcal{T}}$.

Table 3.1 shows the estimated mMISE and cMISE for the ID and g-modeling approaches. The estimated cMISE for the empirical distribution as an estimator of the compound distribution is included as a benchmark. Despite that the true mixing distribution meets the exponential family assumption in the g-modeling approach, the estimated mMISE and cMISE are smaller for the ID approach than for the g-modeling approach, suggesting that the ID approach provides better recovery of the mixing and induced compound distributions.

3.4.2 Scenario 2

In the second scenario, we consider the mixing distribution to be a mixture of two normal distributions. The data generating process is

$$\Theta_i \overset{iid}{\sim} 0.4 \cdot \text{N}(-1, 0.5^2) + 0.6 \cdot \text{N}(1, 0.5^2), \qquad X_i \mid \Theta_i \overset{ind.}{\sim} \text{N}(\Theta_i, 1),$$

for $i = 1, \ldots, 1000$. The component distribution family is the set of normal distributions indexed by the mean parameter with a known variance of 1. The mixing distribution is identifiable.

Table 3.1: Estimated MISE in recovering the mixing distribution (mMISE) and the induced compound distribution (cMISE), calculated by averaging the quantities in (3.7) over 100 simulated datasets. Values have been multiplied by 10,000.

Method	mMISE	cMISE
Scenario 1: G : Gamma$(4, 4)$, $P(\cdot \mid \theta)$: Poisson(θ)		
Empirical benchmark	–	7.8
ID (uniform ansatz)	3.2	4.7
g-modeling	6.2	5.6
Scenario 2: $G:\ 0.6 \cdot \mathrm{N}(-1, 0.5^2) + 0.4 \cdot \mathrm{N}(1, 0.5^2)$, $P(\cdot \mid \theta) : \mathrm{N}(\theta, 1)$		
KDE benchmark	–	7.3
ID (uniform ansatz)	2.8	4.3
g-modeling	6.4	8.7
Scenario 3: $G:\ 0.85 \cdot \delta_0 + 0.15 \cdot \mathrm{N}(0, 1)$, $P(\cdot \mid \theta) : \mathrm{N}(\theta, 1)$		
KDE benchmark	–	2.5
ID ($0.5 \cdot \delta_0 + 0.5 \cdot$ Logistic ansatz)	316.6	1.7
ID ($0.7 \cdot \delta_0 + 0.3 \cdot$ Logistic ansatz)	38.3	0.7
ID ($0.9 \cdot \delta_0 + 0.1 \cdot$ Logistic ansatz)	20.8	0.2
ID ($0.85 \cdot \delta_0 + 0.15 \cdot \mathrm{N}(0, 1)$ ansatz)	4.0	0.4
ID-EM ($0.5 \cdot \delta_0 + 0.5 \cdot$ Logistic ansatz)	9.4	0.3
ID-EM ($0.7 \cdot \delta_0 + 0.3 \cdot$ Logistic ansatz)	7.7	0.3
ID-EM ($0.85 \cdot \delta_0 + 0.15 \cdot \mathrm{N}(0, 1)$ ansatz)	5.5	0.2
g-modeling	62.1	0.5
Scenario 4: $G:\ \mathrm{N}(0, 1) \times$ Inv-Gamma$(6, 6)$, $P(\cdot \mid \mu, \sigma^2):\ \mathrm{N}(\mu, \sigma^2)$		
KDE benchmark	–	7.5
ID (bivariate uniform ansatz)	–	4.7
ID (scaled shifted beta $\times$ uniform ansatz)	–	6.3

The compound density is estimated using kernel density estimation (KDE) with a normal kernel. The choice of the kernel is found to have minimal impact. The discretized support of Θ_i is taken to be $\hat{\mathcal{T}} = \{-3.0, -2.9, \ldots, 3.0\}$. Again, the choice of the ansatz has little effect on the results, and we present the results based on a discrete uniform ansatz over $\hat{\mathcal{T}}$.

The simulation results are summarized in Table 3.1. Again, the ID approach results in smaller mMISE and cMISE estimates than the g-modeling approach. Notably, the ID approach better captures the bimodality of the mixing distribution. As an illustration, Figure 3.3 depicts the estimated

mixing densities on one of the 100 simulated datasets. This pattern is consistent across all 100 datasets.

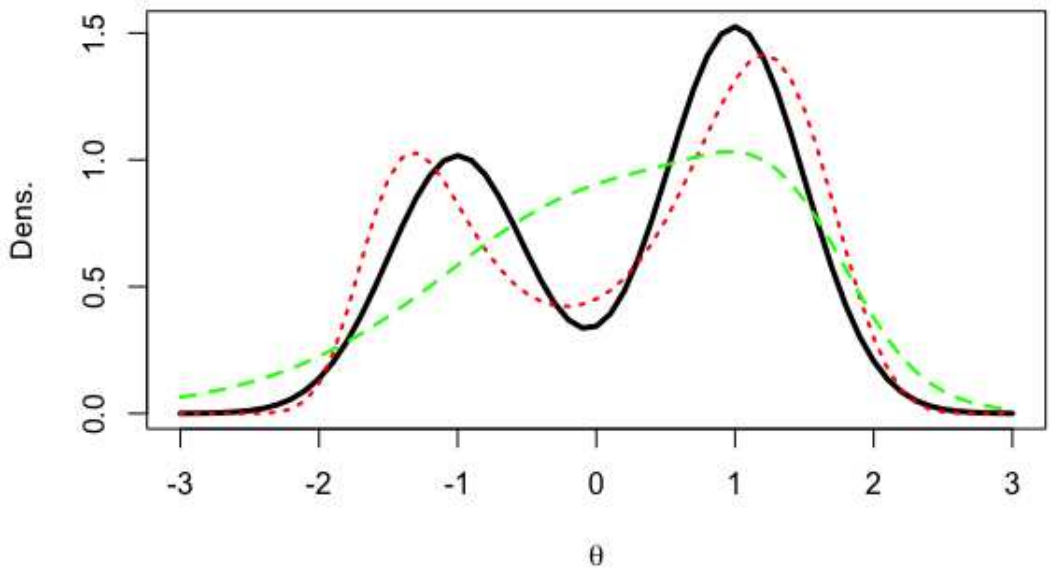

Figure 3.3: Simulation Scenario 2. The solid curve is the true mixing density. The dotted and dashed curves represent the estimated mixing densities using the ID and g-modeling approaches, respectively, on one of the 100 simulated datasets. The pattern is consistent across all 100 datasets.

3.4.3 Scenario 3 and the ID-EM Method

In the third scenario, we consider the mixing distribution to be a mixture of a Dirac measure at 0 and a normal distribution. The data generating process is

$$\Theta_i \overset{iid}{\sim} 0.85 \cdot \delta_0 + 0.15 \cdot \mathrm{N}(0,1), \qquad X_i \mid \Theta_i \overset{ind.}{\sim} \mathrm{N}(\Theta_i, 1),$$

where δ_0 is the Dirac measure at 0. Such a mixing distribution is motivated by the real application in Section 3.5.1 in which Θ_i characterizes the difference in expressions of gene i between the cancer and non-cancer groups. Similar to the real data, the sample size is taken to be $n = 6000$.

The component distribution family comprises normal distributions indexed by the mean parameter, with a known variance of 1, rendering the mixing distribution identifiable. However, as discussed in Section 3.2.3, the recovery of the mixing distribution often encounters instability in this scenario.

The standard ID approach. The compound density is estimated using KDE with a normal kernel. Under the modeling assumption, suppose that the form of the mixing distribution is known to be $G = p_0\delta_0 + (1 - p_0)G_*$, where δ_0 is the Dirac measure at 0, $p_0 \in (0, 1)$ is the point mass at 0, and G_* is dominated by the Lebesgue measure. We consider three ansatzes of the form $G^0 = p_0^0\delta_0 + (1 - p_0^0)G_*^0$. For the first two ansatzes, G_*^0 is a logistic distribution with mean 0 and scale 1. The first and second ansatzes have point masses of $p_0^0 = 0.5$ and 0.7 at 0, respectively. For the third ansatz, we set $G^0 = G$, the true mixing distribution. For computational purposes, we consider a discrete approximation of the mixing distribution on the grid $\hat{\mathcal{T}} = \{-3.6, -3.4, \ldots, 3.6\}$.

The simulation results are summarized in Table 3.1. The estimated cMISE for the kernel density estimator of the compound distribution is included as a benchmark. With any of the three ansatzes, the cMISE estimate for the ID approach is lower than the KDE benchmark. This suggests that the ID approach provides reasonable estimates of the mixing distribution, as demonstrated by the close proximity of their induced compound distributions to the true compound distribution. However, the mMISE estimates tend to be larger, indicating greater deviations of the estimated mixing distributions from the true mixing distribution. In addition, the mMISE values vary widely across different ansatz choices. These phenomena are due to the instability issue as discussed in Section 3.2.3.

As expected, the closer the initial ansatz to the true mixing distribution, the better the recovery of the mixing distribution. In the extreme case where the ansatz is the true mixing distribution itself (the third ansatz), the mMISE value is small. This suggests that the instability issue may be mitigated by employing a judiciously chosen ansatz that closely resembles the true mixing distribution. When there is little knowledge about the true mixing distribution, however, we propose the following modification of the ID method, namely the ID-EM method, to reduce its mMISE when the mixing distribution is a mixture of a Dirac measure and a continuous distribution.

The ID-EM approach. The ID-EM method leverages the form of the mixing distribution and proceeds by separately updating the continuous component and point mass of the ansatz at each iteration. With a normal component distribution family, the compound density can be expressed as $f(x) = p_0\phi(x) + (1-p_0)f_*(x)$. Here, $f_*(x) = \int \phi(x-\theta)g_*(\theta)\mathrm{d}\theta$, where g_* denotes the Lebesgue density of G_*.

Let G^0 denote the initial ansatz and $G^j = p_0^j\delta_0 + (1-p_0^j)G_*^j$ the updated ansatz at the start of iteration $j = 1, 2, \ldots$. Each iteration of the ID-EM method consists of two steps. In the first step, G^j is used to estimate f_* based on weighted KDE (Silverman, 1986). The weight of observation i is proportional to $G^j(\{\theta \neq 0\} \mid X_i)$, where

$$G^j(\{\theta \neq 0\} \mid x) = \frac{(1-p_0^j)f_*^j(x)}{p_0^j\phi(x) + (1-p_0^j)f_*^j(x)},$$

and $f_*^j(x) = \int \phi(x-\theta)g_*^j(\theta)\mathrm{d}\gamma(\theta)$. In essence, $G^j(\{\theta \neq 0\} \mid X_i)$ is the posterior probability that Θ_i arises from the continuous component of G^j, where G^j acts as the prior distribution. An update of g_*^j, denoted as g_*^{j+1}, is produced by plugging in the estimate of f_* into (3.6) and iterating 25 times.

In the second step, p_0^j is updated by

$$p_0^{j+1} = \underset{p\in(0,1)}{\arg\max} \prod_i^n \left\{p\phi(X_i) + (1-p)f_*^{j+1}(X_i)\right\},$$

where $f_*^{j+1}(x) = \int \phi(x-\theta)g_*^{j+1}(\theta)\mathrm{d}\theta$. Combining both steps, G^j is updated to $G^{j+1} = p_0^{j+1}\delta_0 + (1-p_0^{j+1})G_*^{j+1}$. While the first step is not formally an expectation, the two-step procedure is analogous to that in an expectation-maximization (EM) algorithm. Therefore, we refer to this iterative method as ID-EM. We iterate 25 times over both steps.

Table 3.1 presents the results of the ID-EM approach using the same three ansatzes as before. The cMISE estimates remain low. The mMISE estimates and their sensitivity to the choice of the ansatz are substantially reduced as compared to the standard ID approach. Notably, the results for the ID-EM approach compare favorably with those for the g-modeling approach.

The estimation of the point mass $p_0 = G(\{0\})$ may be of particular interest. For example, in the prostate cancer application (Section 3.5.1), p_0 characterizes the proportion of genes whose expressions show no difference between the cancer and non-cancer groups. Table 3.2 reports the estimated mean squared error (MSE) for the estimator of p_0 given by the ID-EM and g-modeling approaches.

Following Efron (2014) and Narasimhan and Efron (2020), we employ a parametric bootstrap to create a confidence interval (CI) for p_0. Based on a sample $\{X_i : i = 1, \ldots, n\}$, an estimate $\hat{G}$ of G is obtained. A bootstrap replication of the sample is obtained by first drawing $\Theta_i^{(b)} \sim \hat{G}$ and then drawing $X_i^{(b)} \mid \Theta_i^{(b)} \sim \mathrm{N}(\Theta_i^{(b)}, 1)$ for $i = 1, \ldots, n$. Finally, $\hat{G}^{(b)}$ is derived from the bootstrap replication $\{X_i^{(b)} : i = 1, \ldots, n\}$. Repeating this process for $b = 1, \ldots, B$, a 95% bootstrap CI of p_0 is constructed by taking the 0.025 and 0.975 sample quantiles of $\{\hat{G}^{(b)}(\{0\}) : b = 1, \ldots, B\}$. We use $B = 100$ bootstrap replications.

Table 3.2 summarizes the proportion of CIs covering the true value of 0.85 and the average CI length across the 100 simulated datasets. The coverage rate for the ID-EM approach is similar to that for the g-modeling approach and does not give rise to concerns. The ID-EM approach results in a shorter average CI length than the g-modeling approach. A nonparametric bootstrap procedure produced similar results (not shown).

Table 3.2: Simulation Scenario 3. Estimated MSE (multiplied by 10,000) for point estimator of p_0. Coverage rate and average length for bootstrap confidence interval of p_0.

Method	MSE	CI coverage	CI length
ID-EM ($0.5 \cdot \delta_0 + 0.5 \cdot$ Logistic ansatz)	8.6	96%	0.211
ID-EM ($0.7 \cdot \delta_0 + 0.3 \cdot$ Logistic ansatz)	6.9	89%	0.195
ID-EM ($0.85 \cdot \delta_0 + 0.15 \cdot \mathrm{N}(0, 1)$ ansatz)	5.0	98%	0.273
g-modeling	53.8	90%	0.264

3.4.4 Scenario 4

In the fourth scenario, we consider a bivariate mixing distribution with the latent variable $\boldsymbol{\Theta}_i = (\mu_i, \sigma_i^2)$. The data generating process is

$$\mu_i \overset{iid}{\sim} \text{N}(0,1), \quad \sigma_i^2 \overset{iid}{\sim} \text{Inv-Gamma}(6,6), \quad \text{and} \quad X_i \mid \mu_i, \sigma_i^2 \overset{ind.}{\sim} \text{N}(\mu_i, \sigma_i^2),$$

for $i = 1, \ldots, 1000$, where σ_i^2 is independent of μ_i. The component distribution family is the set of normal distributions indexed by both the mean and variance parameters. In this scenario, the mixing distribution lacks identifiability; see Section 3.2.2 and Teicher (1961).

The compound density is estimated using KDE with a normal kernel. The ID method is implemented with two ansatzes. The first ansatz is a discrete uniform distribution over a two-dimensional grid, $\hat{\mathcal{T}} = \{-4.0, -3.6, \ldots, 4.0\} \times \{0.5, 1.0, \ldots, 12.0\}$, consisting of 504 tuples. The second ansatz consists of a discrete approximation to a scaled and shifted Beta$(3,3)$ distribution over a grid of 126 equally spaced values ranging from -4 to 4 for the mean parameter, and an independent discrete uniform distribution over $\{1,2,3,4\}$ for the variance parameter. Note that the second ansatz is supported on a very coarse grid of σ^2. Since the `deconvolveR` package does not support a normal component distribution family indexed by both the mean and variance parameters, a comparison with the g-modeling approach is not implemented in this scenario.

Due to nonidentifiability, the estimated mixing distribution produced by the ID method depends on the choice of the ansatz and may not approximate the mixing distribution used in the data generating process. As a result, the mMISE is not used as a performance metric in this scenario. Instead, we compare the induced estimated compound distribution with the truth through the cMISE. Table 3.1 summarizes the cMISE estimates based on the two ansatzes, which are both lower than the KDE benchmark, suggesting that the ID method produces reasonable estimates of the mixing distribution as their induced compound distributions closely resemble the true compound distribution.

Summary. As summarized in Table 3.1, the ID method yields reasonable estimates of the mixing distribution in all four scenarios, as evidenced by its lower cMISE values compared to the empirical

and KDE benchmarks. It often outperforms the g-modeling approach, again as indicated by its lower mMISE and cMISE values. Being fully nonparametric, the ID method is widely applicable, irrespective of the parametric form of the true mixing distribution. Its flexibility allows for tailored approaches to tackle unstable scenarios by leveraging the specific form of the mixing distribution. Finally, in nonidentifiable scenarios, the ID method can produce different plausible estimates of the mixing distribution by using different ansatzes.

3.5 Applications

3.5.1 Prostate Cancer

We apply the ID method to data from a prostate cancer study. Singh et al. (2002) conducted a microarray expression analysis to identify genes potentially associated with prostate cancer. The data were summarized by Efron and are available at `https://efron.ckirby.su.domains/LSI/datasets-and-programs/`. The dataset consists of gene expression levels from 102 men, 52 with cancer and 50 without cancer, across $n = 6033$ genes. For each gene i, Efron (2010, 2016) calculated the two-sample t-test statistic T_i, comparing the average gene expression between the cancer and cancer-free groups. A transformation of T_i to $X_i = \Phi^{-1}\left(F_{100}(T_i)\right)$ was applied, where F_{100} is the cdf of the t distribution with 100 degrees of freedom. The following model was justified by Efron (2010),

$$X_i \mid \Theta_i \overset{ind}{\sim} \mathrm{N}(\Theta_i, 1), \qquad \Theta_i \overset{iid}{\sim} G,$$

where Θ_i represents the effect size for gene i, with $\Theta_i = 0$ for null genes. The unknown mixing distribution G is of interest. Efron (2016) used the g-modeling approach to analyze the data under this model, assuming that G is a mixture of a Dirac measure at 0 and a continuous distribution G_*. We reanalyze the data using the ID method.

The compound density is estimated using KDE with a normal kernel. We take the ansatz to be $G^0 = 0.85 \cdot \delta_0 + 0.15 \cdot \mathrm{N}(0, 1)$. Here, the choice of the point mass of 0.85 in the ansatz corresponds

to its estimate provided by the g-modeling approach. The discretized support of G is taken to be $\hat{\mathcal{T}} = \{-3.6, -3.4, \ldots, 3.6\}$. We implement the ID-EM approach described in Section 3.4.3 with 25 iterations over the two steps. The estimated mixing distribution $\hat{G}$ puts probability 0.848 on $\{\Theta = 0\}$, indicating that 84.8% of the genes are estimated to have no association with prostate cancer. The estimated Lebesgue density of G_*, denoted by $\hat{g}_*(\theta)$, is shown as the solid curve in Figure 3.4 (Panel B). As a goodness-of-fit check, we perform a Kolmogorov–Smirnov (KS) test to determine if the sample $X_1, \ldots, X_n$ can plausibly be generated from the compound distribution induced by $\hat{G}$. The KS statistic is $D = 0.009$ with a p-value of 0.671, indicating no evidence of lack of fit.

To quantify uncertainty, we perform a parametric bootstrap with 100 bootstrap replications. See Section 3.4.3 for more details about the implementation of the parametric bootstrap. Panel A of Figure 3.4 shows a boxplot of $\hat{G}(\{0\})$ from the bootstrap replications, with a 95% bootstrap CI of $(0.793, 0.929)$ based on the 0.025 and 0.975 quantiles. In Panel B of Figure 3.4, the dashed vertical lines represent the bootstrap standard errors of $\hat{g}_*(\theta)$, extending from plus to minus one standard error.

From an empirical Bayes perspective (Efron, 2010), the estimate $\hat{G}$ can be used as a prior to then obtain a posterior distribution $\hat{G}(\cdot \mid X_i)$ for the effect size of each gene. We find 16 genes that have a posterior probability of having a null effect, $\hat{G}(\{\Theta_i = 0\} \mid X_i)$, less than 0.05. The numbers of genes with this posterior probability less than 0.2 and 0.5 are 60 and 168, respectively.

3.5.2 Word Count

Our second application concerns the analysis of word counts from Shakespeare's canon of works, which has a total word count of C = 884,647. The data were derived from a concordance (Spevack, 1968) of Shakespeare's work and are available in the R package `deconvolveR` (Narasimhan and Efron, 2020). The dataset consists of counts n_x, representing the number of words in the canon that appear x times for $x = 1, \ldots, 100$. In total, there are $n = \sum_{x=1}^{100} n_x =$

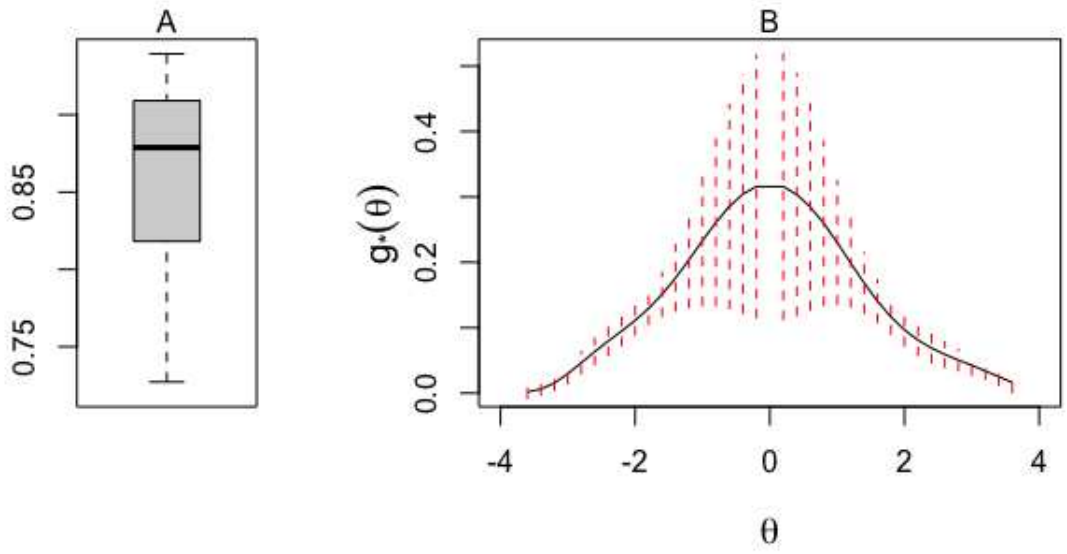

Figure 3.4: Panel A is a box plot of $\hat{G}(\{0\})$ from 100 parametric bootstrap replications. In Panel B, the solid curve shows $\hat{g}_*(\theta)$, the estimated Lebesgue density of the continuous component of the mixing distribution; the dashed vertical lines are plus and minus one standard error, estimated by bootstrap.

30,688 different words in the dataset. Words appearing more than 100 times, such as "and" and "the", are deemed less interesting and are not included in the dataset. Note that $\sum_{x=101}^{\infty} n_x = 846$.

Efron and Thisted (1976) posed the question: how many words did Shakespeare know? Alternatively, one might ask: if we hypothetically discovered another canon of works by Shakespeare, how many new words could we expect to find? This question is analogous to estimating the number of unseen species, a common goal in ecological studies.

To answer this question, we follow the model used by Narasimhan and Efron (2020). For the ith distinct word appearing 1–100 times in Shakespeare's canon, let X_i denote the number of times it appears. The count X_i is modeled by a Poisson distribution with rate Θ_i, truncated to the set $\{1, \ldots, 100\}$, where Θ_i is assumed to come from a mixing distribution G. In summary,

$$X_i \mid \Theta_i \overset{ind.}{\sim} \text{Trunc-Poisson}_{\{1,\ldots,100\}}(\Theta_i), \qquad \Theta_i \overset{iid}{\sim} G.$$

As a word of caution, with a truncated Poisson component distribution family, G is not identifiable; recall Section 3.3.1. Therefore, inference on G should ideally involve discussion of the class of equivalent mixing distributions.

The compound distribution is estimated by the empirical distribution. The discretized support of G is taken to be $\hat{\mathcal{T}} = \exp\{L\}$, where L is a set of 100 evenly spaced points ranging from -4 to 4.5. We use a discrete uniform ansatz over $\hat{\mathcal{T}}$. Let $\hat{g}_j$ denote the approximation for a probability mass function $g(\theta_j)$, $\theta_j \in \hat{\mathcal{T}}$, obtained according to (3.6). Then, approximately, the expected number of distinct new words, divided by the length of the actual canon C, in a newly discovered canon of word length Ct can be represented by $R(t)$ defined as

$$R(t) = \sum_{j=1}^{100} \hat{g}_j r_j(t), \text{ where}$$
$$r_j = \frac{\exp\{-\theta_j\}}{1-\exp\{-\theta_j\}} \left(1 - \exp\{-\theta_j t\}\right);$$

see Narasimhan and Efron (2020) for details and derivation.

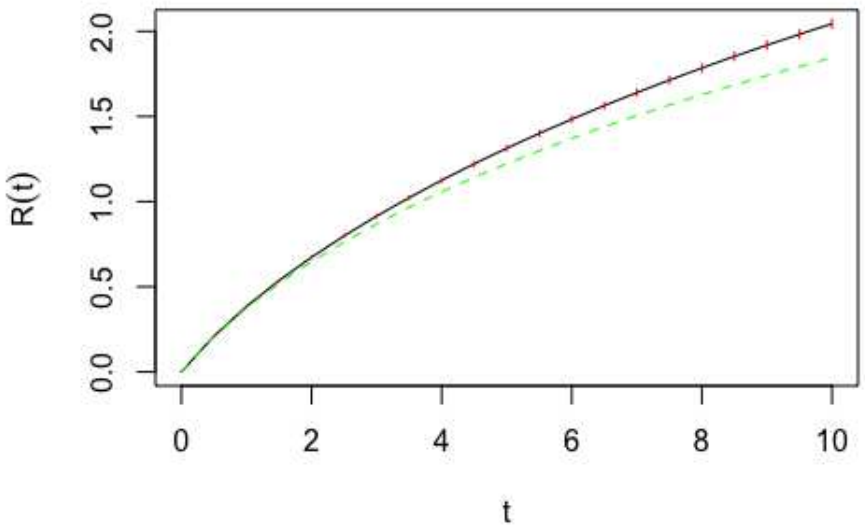

Figure 3.5: The solid black line is $R(t)$ resulting from the ID approach. The red vertical lines are plus and minus one standard deviation, derived from a parametric bootstrap. The green dashed line is $R(t)$ resulting from the g-modeling approach.

Figure 3.5 plots $R(t)$ over t resulting from the ID-method with a uniform ansatz on $\hat{\mathcal{T}}$, depicted by the solid black line. For example, because $R(3.39) \approx 1$, for a newly discovered canon of about 3.39 times larger than the known canon, we would expect to double the number of observed words in Shakespeare's vocabulary. The red vertical lines are plus and minus one standard deviation from a parametric bootstrap. The green dashed line is $R(t)$ produced by the g-modeling approximation on $\hat{\mathcal{T}}$. As a sensitivity analysis, we run the ID method with different ansatzes. The estimated mixing distributions map to nearly identical estimates of the number of distinct new words discovered in the new corpus, corroborating our conclusion in the nonidentifiable case.

3.6 Discussion

We propose a new flexible, nonparametric method for recovering a mixing distribution given a known component distribution family and a sample from a compound distribution. Our method is applicable to a wide range of previously studied scenarios of the mixing distribution problem including deconvolution models and the Normal Means problem. The proposed method is relatively intuitive, easily implementable, and compares well with the commonly used g-modeling approach under a number of scenarios. In scenarios where the mixing distribution is nonidentifiable, our method can approximate many mixing distributions which, in junction with the component distribution family, could have plausibly generated the observed data.

There are extensions to be explored with the proposed method. For example, in this work, the observed X_i values are generated by the same family of component distributions. In contrast, in some scenarios the component distribution family may differ for each data point. That is, $X_i|\Theta_i \sim P_i(\cdot|\Theta_i)$. An example of such a scenario is $X_i|\Theta_i \sim \text{Bin}(m_i, \Theta_i)$ where Θ_i is the observed probability of the binomial distribution and m_i is the observed number of trials. The parameters m_i may be designated by an experimenter for each draw of X_i. An extension of the proposed method to such a scenario is left as future work.

Although we teased nonidentifiable scenarios in this chapter, more research should be done into the equivalent classes of mixing distributions which induce the same compound distribution.

For example, in a nonidentifiable scenario, one may be interested in the probabilities assigned to an event E by all (or perhaps a subset of, such as all continuous) mixing distributions $G \in \mathcal{E}_{F,\mathcal{P}}$. Following similar notion as Dempster (1967), one may define lower and upper probabilities of E as

$$G_*(E) = \inf_{G \in \mathcal{E}_{F,\mathcal{P}}} G(E) \qquad \text{and} \qquad G^*(E) = \sup_{G \in \mathcal{E}_{F,\mathcal{P}}} G(E).$$

Given the convexity of $\mathcal{E}_{F,\mathcal{P}}$, the interval $(G_*(E), G^*(E))$ delineates the plausible range of probabilities for event E. Of course, this range may be trivial, e.g., $[0, 1]$ as in the first example in Section 3.2.2. However, it could be of interest for which classes $\mathcal{E}_{F,\mathcal{P}}$ and which events E, the resulting interval is not trivial.

Chapter 4

Summary and Future Work

In this dissertation we proposed novel approaches to approximating distributions in equivalence classes with and without identifiability. In Chapter 2, we showed that the coefficient distribution is not identifiable when the design variable takes values on a discrete lattice. Finding coefficient distributions in the RCR model was framed as a collection of SIPs. We proposed a novel iterative method that at each iteration finds a solution to each SIP by disintegration and averages the local solutions to update the current global ansatz. The initial global ansatz can be changed so that the iterative method approximates different distributions within the desired equivalence class. We studied the convergence properties of the method and proposed an analogous empirical version. The iterative method compared well against the moment-matching estimator proposed by Beran and Hall (1992) and the Radon transform estimator proposed by Hoderlein et al. (2010) in simulation studies. In the same simulation studies, we also compared our method with a Bayesian nonparametric approach, a Dirichlet process mixture model. The DPM performed well under one simulation scenario, but not the other. In contrast, our proposed method appeared robust to the different data-generating processes. Finally, we demonstrated our proposed method on an acupuncture clinical trial, where we showed that despite a statistically significant treatment effect, the plausible percentage of patients who benefit from acupuncture could range from about 22.3% to 98.7%, according to our analysis.

In Chapter 3, we summarized the identifiability of mixing distributions and emphasized the possibility of unstable recovery scenarios. With the equivalence class perspective to nonidentifiability in mind, we proposed a similar iterative method as that in Chapter 2, but for the recovery of mixing distributions. The method can be used in identifiable and nonidentifiable scenarios. The user-specified ansatz can be changed to guide the iterative method to approximate different mixing distributions under nonidentifiability. Again, we studied the convergence properties of the method, and proposed an analogous empirical version. We compared our method with the well-known

G-modeling approach proposed by Efron (2014, 2016) through simulation studies, and found our method to perform well. In recovery scenarios that suffer from instability, we demonstrated that the flexible approach can be adapted to stabilize the recovery problem, by incorporating it within an EM-like approach. The method of Chapter 3 was demonstrated on two datasets: a prostate cancer dataset (Singh et al., 2002; Efron, 2016) and the Shakespeare's word count dataset (Spevack, 1968; Narasimhan and Efron, 2020). In the analysis of the prostate cancer dataset, under an empirical Bayesian framework, we estimated the mixing distribution, taken to be a prior, and consequently derived a posterior distribution to help identify which genes may be associated with prostate cancer in men. The Shakespeare's word count dataset was analyzed to answer the hypothetical question of how many new unique words we would expect to find in a newly discovered canon by Shakespeare.

It is of interest that similar iterative methods, with a similar equivalence class perspective, can be used in seemingly unrelated modeling scenarios. In particular, both methods made use of disintegration and integration of measures to step closer to the target solution set. In addition, both methods used an initiating user-specified distribution, called a global ansatz in Chapter 2 and simply an ansatz in Chapter 3. The convergence properties were similar between the two. In the RCR model's approach, each output distribution induced by the sequence of global ansatzes converges to the corresponding unknown output distribution with respect to the KL-divergence. The analogous result was shown in Chapter 3, where the sequence of induced compound distributions converges to the unknown target compound distribution with respect to the KL-divergence. In practice, both methods make use of the data to estimate distributions, output distributions and the compound distribution, respectively.

Although the methods in each chapter are similar, the modeling scenarios presented in each chapter differ in a few aspects. In Chapter 2, we only considered coefficient distributions with Lebesgue densities. Whereas in Chapter 3, we took a more general approach, where we discussed densities of the mixing distribution and compound distribution with respect to general dominating measures. In future work, it may be of interest to relax the Lebesgue restriction of the solution class

of Chapter 2. As discussed in Section 2.2.2, and shown in Appendix A.0.2, not all distributions that induce the target set of output distributions have Lebesgue densities in $\mathbb{R}^2$, despite each output distribution being absolutely continuous with respect to the Lebesgue measure in $\mathbb{R}$.

Moreover, in Chapter 3, the equivalence class is characterized by a probability kernel and a single distribution. Whereas the equivalence class of Chapter 2 is characterized by a finite set of maps and their associated image probability measures. Finally, for the RCR models in this dissertation, the coefficient distributions are always nonidentifiable. Whereas in the mixing distribution recovery problem we consider both identifiable and nonidentifiable scenarios.

As the approaches to both problems share similarities, it may be of interest for future work to generalize the approaches and adapt them to different modeling scenarios and problems that also suffer from nonidentifiability. Additionally, to facilitate inference, plausible probability intervals on events of interest can be investigated, as described in the conclusion section of each chapter. A related problem, finding which sets in the σ-algebras corresponding to the solution set have nontrivial probability ranges, can further be researched.

The finite-sample and asymptotic behaviors of the empirical versions of each method can be explored in more detail. Such an example of a finite-sample behavior that may be investigated further is a similar pattern which emerged in both methods in the simulation studies. In Chapter 2, the original nonparametric estimates of each output distribution had greater KS-distances from the corresponding true output distribution on average than did each induced output distribution from the resulting estimate of the coefficient distribution. Analogously, in Chapter 3, the original nonparametric estimate of the compound distribution had a higher mean integrated square error from the true compound distribution than did the compound distribution induced by the estimate of the mixing distribution. We leave it as future work to investigate why this pattern emerges in both modeling scenarios.

For the methods in both Chapters 2 and 3, extensions of each can be investigated. For the iterative method of the RCR models, the model can be extended from the simple RCR model to include vectors for B_i and x_i. Additionally, the current iterative method can be investigated

when the design variable is no longer conditioned upon, and instead is considered random. For the iterative method to recover mixing distributions, a possible extension is to substitute the single compound distribution for a set of compound distributions $F_i(\cdot) = F(\cdot|\eta_i)$, where η_i is an observed quantity. That is,

$$F(A|\eta_i) = \int_{\mathcal{T}} P(A|\theta, \eta_i) \mathrm{d}G(\theta),$$

for each $A \in \mathcal{B}_{\mathcal{X}}$ and each i. The recovery problem becomes finding the mixing distribution G from a given set of compound distributions $F(\cdot|\eta_i)$ and a set of component distributions $P(\cdot|\theta, \eta_i)$. An example of such a case is the family of component distributions is a family of Binomial distributions where the number of trials η_i, is observed and is mixed over the success probability parameter θ.

Bibliography

Atienza, N., Garcia-Heras, J., and Munoz-Pichardo, J. M. (2006). A new condition for identifiability of finite mixture distributions. Metrika, 63(2):215–221.

Barnard, G. A. (1954). Sampling inspection and statistical decisions. Journal of the Royal Statistical Society. Series B (Methodological), 16(2):151–174.

Beran, R., Feuerverger, A., and Hall, P. (1996). On nonparametric estimation of intercept and slope distributions in random coefficient regression. The Annals of Statistics, 24(6):2569–2592.

Beran, R. and Hall, P. (1992). Estimating coefficient distributions in random coefficient regressions. The annals of Statistics, pages 1970–1984.

Beran, R. and Millar, P. W. (1994). Minimum distance estimation in random coefficient regression models. The Annals of Statistics, 22(4):1976–1992.

Bingham, D., Butler, T., and Estep, D. (2024). Inverse problems for physics-based process models. Annual Review of Statistics and Its Application, 11.

Bowman, A. W. (1984). An alternative method of cross-validation for the smoothing of density estimates. Biometrika, 71(2):353–360.

Breidt, J., Butler, T., and Estep, D. (2011). A measure-theoretic computational method for inverse sensitivity problems I: Method and analysis. SIAM Journal on Numerical Analysis, 49(5):1836–1859.

Butler, T., Estep, D., and Sandelin, J. (2012). A computational measure theoretic approach to inverse sensitivity problems II: A posteriori error analysis. SIAM Journal on Numerical Analysis, 50(1):22–45.

Butler, T., Estep, D., Tavener, S., Dawson, C., and Westerink, J. J. (2014). A measure-theoretic computational method for inverse sensitivity problems III: multiple quantities of interest. SIAM/ASA Journal on Uncertainty Quantification, 2(1):174–202.

Butler, T., Jakeman, J., and Wildey, T. (2018). Combining push-forward measures and Bayes' rule to construct consistent solutions to stochastic inverse problems. SIAM Journal on Scientific Computing, 40(2):984–1011.

Butucea, C. and Comte, F. (2009). Adaptive estimation of linear functionals in the convolution model and applications. Bernoulli, 15(1):69–98.

Carroll, R. J. and Hall, P. (1988). Optimal rates of convergence for deconvolving a density. Journal of the American Statistical Association, 83(404):1184–1186.

Casella, G. and Berger, R. L. (2002). Statistical inference. Duxbury advanced series. Duxbury/Thomson Learning, Pacific Grove, CA, 2nd ed. edition.

Chang, J. T. and Pollard, D. (1997). Conditioning as disintegration. Statistica Neerlandica, 51(3):287–317.

Chi, J. (2021). Sliced inverse approach and domain recovery for stochastic inverse problems. PhD thesis, Colorado State University.

Crowder, M. J. (1978). Beta-binomial anova for proportions. Journal of the Royal Statistical Society. Series C (Applied Statistics), 27(1):34–37.

Cui, Y. and Hannig, J. (2022). A fiducial approach to nonparametric deconvolution problem: discrete case.

Deans, S. R. (1983). The Radon Transform and Some of Its Applications. Wiley, New York.

Dempster, A. (1967). Upper and lower probabilities induced by a multivalued mapping. The Annals of Mathematical Statistics, 38(2):325–339.

Dunker, F., Eckle, K., Proksch, K., and Schmidt-Hieber, J. (2019). Tests for qualitative features in the random coefficients model. Electronic Journal of Statistics, 13:2257–2306.

Dunker, F., Mendoza, E., and Reale, M. (2021). Regularized maximum likelihood estimation for the random coefficients model. arXiv preprint arXiv:2104.08402.

Durlauf, S. N., Hansen, L. P., Heckman, J. J. J. J., and Matzkin, R. L. (2020). Handbook of econometrics Volume 7A. Handbooks in economics. Elsevier, North Holland, Amsterdam.

Efron, B. (2010). Large-Scale Inference: Empirical Bayes Methods for Estimation, Testing, and Prediction. Institute of Mathematical Statistics Monographs. Cambridge University Press.

Efron, B. (2014). Two Modeling Strategies for Empirical Bayes Estimation. Statistical Science, 29(2):285 – 301.

Efron, B. (2016). Empirical Bayes deconvolution estimates. Biometrika, 103(1):1–20.

Efron, B. and Thisted, R. (1976). Estimating the number of unseen species: How many words did Shakespeare know? Biometrika, 63(3):435–447.

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

Faden, A. M. (1985). The existence of regular conditional probabilities: Necessary and sufficient conditions. The Annals of probability, 13(1):288–298.

Fan, J. (1991). On the optimal rates of convergence for nonparametric deconvolution problems. The Annals of Statistics, 19(3):1257–1272.

Feller, W. (1943). On a General Class of "Contagious" Distributions. The Annals of Mathematical Statistics, 14(4):389 – 400.

Fisk, P. (1967). Models of the second kind in regression analysis. Journal of the Royal Statistical Society: Series B (Methodological), 29(2):266–281.

Froehlich, B. (1973). Some estimators for a random coefficient regression model. Journal of the American Statistical Association, 68(342):329–335.

Fryer, M. J. (1976). Some Errors Associated with the Non-parametric Estimation of Density Functions. IMA Journal of Applied Mathematics, 18(3):371–380.

Gaillac, C. and Gautier, E. (2022). Adaptive estimation in the linear random coefficients model when regressors have limited variation. Bernoulli, 28(1):504–524.

Goutis, C. (1997). Nonparametric estimation of a mixing density via the kernel method. Journal of the American Statistical Association, 92(440):1445–1450.

Hadamard, J. (1902). Sur les problèmes aux dérivées partielles et leur signification physique. Princeton university bulletin, pages 49–52.

Hall, P. and Meister, A. (2007). A ridge-parameter approach to deconvolution. The Annals of Statistics, 35(4):1535–1558.

Hamedani, G. G. and Tata, M. N. (1975). On the determination of the bivariate normal distribution from distributions of linear combinations of the variables. The American Mathematical Monthly, 82(9):913–915.

Hammersley, J. M. and Morton, K. W. (1954). Poor man's monte carlo. Journal of the Royal Statistical Society. Series B (Methodological), 16(1):23–38.

Heinrich, P. and Kahn, J. (2018). Strong identifiability and optimal minimax rates for finite mixture estimation. The Annals of Statistics, 46(6A):2844–2870.

Hermann, P. and Holzmann, H. (2021). Bounded support in linear random coefficient models: Identification and variable selection. arXiv preprint arXiv:2107.03245.

Hildreth, C. and Houck, J. P. (1968). Some estimators for a linear model with random coefficients. Journal of the American Statistical Association, 63(322):584–595.

Hoderlein, S., Klemelä, J., and Mammen, E. (2010). Analyzing the random coefficient model nonparametrically. Econometric Theory, 26(3):804–837.

Holzmann, H. and Meister, A. (2020). Rate-optimal nonparametric estimation for random coefficient regression models. Bernoulli, 26(4):2790–2814.

Jiang, W. and Zhang, C.-H. (2009). General maximum likelihood empirical Bayes estimation of normal means. The Annals of Statistics, 37(4):1647 – 1684.

Johnson, L. W. (1977). Stochastic parameter regression: An annotated bibliography. International Statistical Review, 45(3):257–272.

Karlis, D. and Xekalaki, E. (2005). Mixed poisson distributions. International Statistical Review / Revue Internationale de Statistique, 73(1):35–58.

Kendall, M. G., Stuart, A., Ord, J. K., Arnold, S. F., and O'Hagan, A. (1994). Kendall's advanced theory of statistics. Edward Arnold, London, 6th ed. edition.

Koopmans, T. C. and Reiersol, O. (1950). The identification of structural characteristics. The Annals of Mathematical Statistics, 21(2):165–181.

Kullback, S. (1959). Information theory and statistics. Wiley publication in mathematical statistics. Wiley, New York.

Kullback, S. (1968). Probability densities with given marginals. The Annals of Mathematical Statistics, 39(4):1236–1243.

Laird, N. (1978). Nonparametric maximum likelihood estimation of a mixing distribution. Journal of the American Statistical Association, 73(364):805–811.

Lehmann, E. L. and Casella, G. (1998). Theory of Point Estimation. Springer Texts in Statistics. Springer Nature, New York, NY, second edition. edition.

Liu, C. (2023). Another look at the problem of many-normal-means.

Liu, M. C. and Taylor, R. L. (1989). A consistent nonparametric density estimator for the deconvolution problem. The Canadian Journal of Statistics / La Revue Canadienne de Statistique, 17(4):427–438.

Lomax, K. S. (1954). Business failures: Another example of the analysis of failure data. Journal of the American Statistical Association, 49(268):847–852.

Masten, M. A. (2018). Random coefficients on endogenous variables in simultaneous equations models. Review of Economic Studies, pages 1193–1259.

Narasimhan, B. and Efron, B. (2020). deconvolveR: A g-modeling program for deconvolution and empirical bayes estimation. Journal of Statistical Software, 94(11):1–20.

Neath, A. A. and Samaniego, F. J. (1997). On the efficacy of Bayesian inference for nonidentifiable models. The American Statistician, 51(3):225–232.

Nelder, J. (1968). Regression, model-building and invariance. Journal of the Royal Statistical Society: Series A (General), 131(3):303–315.

Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. The Annals of Mathematical Statistics, 33(3):1065 – 1076.

Patrick, E. and Hancock, J. (1966). Nonsupervised sequential classification and recognition of patterns. IEEE Transactions on Information Theory, 12(3):362–372.

Robbins, H. (1964). The empirical bayes approach to statistical decision problems. The Annals of Mathematical Statistics, 35(1):1–20.

Rosenblatt, M. (1956). Remarks on some nonparametric estimates of a density function. The Annals of Mathematical Statistics, 27(3):832 – 837.

Rothenberg, T. J. (1971). Identification in parametric models. Econometrica, 39(3):577–591.

Silverman, B. W. (1986). Density estimation for statistics and data analysis. Monographs on statistics and applied probability. Chapman and Hall, London ;.

Singh, D., Febbo, P. G., Ross, K., Jackson, D. G., Manola, J., Ladd, C., Tamayo, P., Renshaw, A. A., D'Amico, A. V., Richie, J. P., Lander, E. S., Loda, M., Kantoff, P. W., Golub, T. R., and Sellers, W. R. (2002). Gene expression correlates of clinical prostate cancer behavior. Cancer Cell, 1(2):203–209.

Skellam, J. G. (1948). A probability distribution derived from the binomial distribution by regarding the probability of success as variable between the sets of trials. Journal of the Royal Statistical Society. Series B (Methodological), 10(2):257–261.

Spevack, M. (1968). A Complete and Systematic Concordance to the Works of Shakespeare., volume 1-6. Hildesheim: George Olms.

Sprent, P. and Smeeton, N. C. (2007). Applied nonparametric statistical methods. Texts in statistical science. Chapman & Hall/CRC, Boca Raton, 4th ed. edition.

Stefanski, L. A. and Carroll, R. J. (1990). Deconvolving kernel density estimators. Statistics, 21(2):169–184.

Strauss, W. A. (2008). Partial differential equations : an introduction. Wiley, Hoboken, N.J, 2nd ed. edition.

Sun, L. (2020). Topics On Empirical Bayes Normal Means. PhD thesis, The University of Chicago.

Tallis, G. M. (1969). The identifiability of mixtures of distributions. Journal of Applied Probability, 6(2):389–398.

Teicher, H. (1960). On the mixture of distributions. The Annals of Mathematical Statistics, 31(1):55–73.

Teicher, H. (1961). Identifiability of mixtures. The Annals of Mathematical Statistics, 32(1):244–248.

Teicher, H. (1963). Identifiability of finite mixtures. The Annals of mathematical statistics, 34(4):1265–1269.

Tikhonov, A. N., Goncharsky, A., Stepanov, V. V., and Yagola, A. G. (1995). Numerical Methods for the Solution of Ill-Posed Problems, volume 328 of Mathematics and Its Applications. Springer Netherlands, Dordrecht, 1 edition.

Tokdar, S. T. and Kass, R. E. (2010). Importance sampling: a review. WIREs Computational Statistics, 2(1):54–60.

Vickers, A. J. (2006). Whose data set is it anyway? sharing raw data from randomized trials. Trials, 7(1):1–6.

Vickers, A. J., Rees, R. W., Zollman, C. E., McCarney, R., Smith, C. M., Ellis, N., Fisher, P., and Van Haselen, R. (2004). Acupuncture for chronic headache in primary care: large, pragmatic, randomised trial. BMJ, 328(7442):744.

Wechsler, S., Izbicki, R., and Esteves, L. G. (2013). A Bayesian look at nonidentifiability: A simple example. The American Statistician, 67(2):90–93.

Yakowitz, S. (1969). A consistent estimator for the identification of finite mixtures. The Annals of mathematical statistics, 40(5):1728–1735.

Yang, L. (2018). Infinite dimensional stochastic inverse problems. PhD thesis, Colorado State University.

Zhai, J. and Jiang, H. (2023). Two-sample test with g-modeling and its applications. Statistics in Medicine, 42(1):89–104.

Zhang, C.-H. (1990). Fourier methods for estimating mixing densities and distributions. The Annals of Statistics, 18.

Appendix A

Random Coefficient Regression Appendix

A.0.1 Proof of Theorem 1

Define $\mathcal{B}_r(a,b)$ as an open ball centered at (a,b) with radius r. Let $\phi_r(a,b)$ be a continuous nonnegative function, indexed by r, satisfying $0 < \phi_r(a,b) \leq 1$ if $(a,b) \in \mathcal{B}_r(0,0)$ and $\phi_r(a,b) = 0$ otherwise.

Chi (2021) showed that $\mathcal{E}$ is convex; hence, it is sufficient to find one distribution, different from $F_{A,B}$ on a set with a positive Lebesgue measure, that induces the same set of output distributions as $F_{A,B}$ on a finite set $\mathcal{X}$, i.e., $\mathcal{F}_{\mathcal{X}}$.

First, we construct a function $h_1(a,b) \neq 0$ such that $g_1(a,b) = \widetilde{f}_{A,B}(a,b) + h_1(a,b)$ satisfies the following properties:

(i) g_1 is a valid density;

(ii) $f_{A,B}(a,b)$ and g_1 have the same Radon transform for x_1; that is,

$$\int \widetilde{f}_{A,B}(y - bx_1, b)\mathrm{d}b = \int g_1(y - bx_1, b)\mathrm{d}b, \quad \text{for all } y.$$

Note that (i) ensures that g_1 and $\widetilde{f}_{A,B}$ induce the same output distribution at a single design point x_1, i.e., $F_{Y_{x_1}}$. Furthermore, the Radon transform of h_1 is constant 0. Construction of h_1 proceeds as follows.

Let $\mathcal{B}_r(a_0,b_0)$ be a ball centered at (a_0,b_0) with radius r such that

$$\operatorname{ess\,inf}_{(a,b)\in\mathcal{B}_r(a_0,b_0)} \widetilde{f}_{A,B}(a,b) > 0. \tag{A.1}$$

Its existence is due to the fact that $\widetilde{f}_{A,B}$ is continuous almost everywhere with respect to the Lebesgue measure. Because $\widetilde{f}_{A,B}$ was assumed to be Riemann integrable, the density $\widetilde{f}_{A,B}$ is

continuous almost everywhere by Lebesgue-Vitali's theorem. Let δ be any positive number less than the essential infimum.

Next, consider the open balls $\mathcal{B}_s(a_{1,1}, b_{1,1})$ with $0 < s < r2^{-J-1}$, $a_{1,1} = a_0$, and $b_{1,1} = b_0$. Denote $\kappa_{1,1}(a, b) = \phi_s(a - a_{1,1}, b - b_{1,1})$, and $\kappa_{1,2}(a, b) = \phi_s(a - a_{1,2}, b - b_{1,2})$, where $a_{1,2} = a_{1,1} + \tau_1 \sin \arctan(-x_1)$, $b_{1,2} = b_{1,1} + \tau_1 \cos \arctan(-x_1)$, and $\tau_1 = 2s$. Note that $\kappa_{1,1}$ and $\kappa_{1,2}$ are supported over $\mathcal{B}_s(a_{1,1}, b_{1,1})$ and $\mathcal{B}_s(a_{1,2}, b_{1,2})$, respectively. In addition, $\mathcal{B}_s(a_{1,2}, b_{1,2})$ can be obtained by translating $\mathcal{B}_s(a_{1,1}, b_{1,1})$ by $2s$. In fact, since the distance between the centers is $\tau_1 = 2s$, both $\mathcal{B}_s(a_{1,1}, b_{1,1})$ and $\mathcal{B}_s(a_{1,2}, b_{1,2})$ are contained within $\mathcal{B}_r(a_0, b_0)$ and $\mathcal{B}_s(a_{1,1}, b_{1,1}) \cap \mathcal{B}_s(a_{1,2}, b_{1,2}) = \emptyset$.

The function h_1 is defined as

$$h_1(a, b) = \delta\kappa_{1,1}(a, b) - \delta\kappa_{1,2}(a, b).$$

Note that

$$\int h_1(y - bx_1, b)\mathrm{d}b = \delta \int \phi_s(y - bx_1 - a_{1,1}, b - b_{1,1})\mathrm{d}b - \delta \int \phi_s(y - bx_1 - a_{1,2}, b - b_{1,2})\mathrm{d}b.$$

Letting $b' = b - b_{1,2} + b_{1,1}$, the second integral on the right hand side can be expressed as

$$\begin{aligned}\delta \int \phi_s(y - bx_1 - a_{1,2}, b - b_{1,2})\mathrm{d}b &= \delta \int \phi_s(y - b'x_1 - (b_{1,2} - b_{1,1})x_1 - a_{1,2}, b' - b_{1,1})\mathrm{d}b' \\ &= \delta \int \phi_s(y - b'x_1 - a_{1,1}, b' - b_{1,1})\mathrm{d}b',\end{aligned}$$

where the last equality holds because $x_1 \cos \arctan(-x_1) + \sin \arctan(-x_1) = 0$. That is, the Radon transform of h_1 is 0 for $x = x_1$. It is easy to see that for g_1, (i) and (ii) are satisfied.

In Panel A of Figure A.1, the construction of h_1 is depicted for $x_1 = 0$. The open ball $\mathcal{B}_r(a_0, b_0)$ is depicted by the large dotted circle, a region over which $f_{A,B}$ is strictly positive. The supports of $\kappa_{1,1}$ and $\kappa_{1,2}$, i.e., $\mathcal{B}_s(a_{1,1}, b_{1,1})$ and $\mathcal{B}_s(a_{1,2}, b_{1,2})$, are shown as the dashed circle and solid circle,

respectively. Since $\mathcal{B}_s(a_{1,2}, b_{1,2})$ is a translation of $\mathcal{B}_s(a_{1,1}, b_{1,1})$ along the direction, shown as a dash-dotted line, integrating $h_1(a, b)$ along the lines, parallel to the dash-dotted line, is zero.

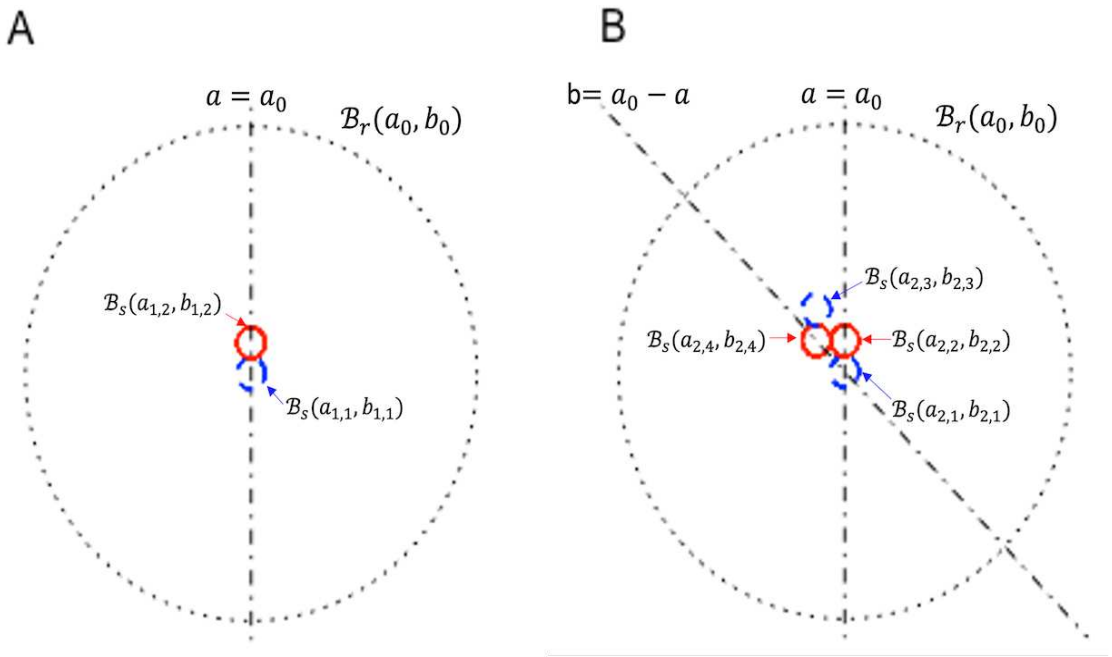

Figure A.1: A representation of the supports of h_1 and h_2 for $x_1 = 0$ (A) and $x_2 = 1$ (B). Here, the large dotted circle indicates $\mathcal{B}_r(a_0, b_0)$, over which the density is strictly positive. A: The supports of $\kappa_{1,1}$ and $\kappa_{1,2}$, $\mathcal{B}_s(a_{1,1}, b_{1,1})$ and $\mathcal{B}_s(a_{1,2}, b_{1,2})$, are shown as the dashed and solid circles, respectively. Note that $\mathcal{B}_s(a_{1,2}, b_{1,2})$ is a translation of $\mathcal{B}_s(a_{1,1}, b_{1,1})$ along the dashed vertical line. The function h_1 is constructed by adding $\delta\kappa_{1,1}$ and subtracting $\delta\kappa_{1,2}$, and it is zero outside of $\mathcal{B}_s(a_{1,1}, b_{1,1})$ and $\mathcal{B}_s(a_{1,2}, b_{1,2})$. B: $\mathcal{B}_s(a_{2,k}, b_{2,k})$ for $k = 1, 2, 3, 4$ are depicted with solid circles for even k and dashed circles for odd circles, and are the support sets for the associated $\kappa_{2,k}$ functions. $\mathcal{B}_s(a_{2,1}, b_{2,1}) = \mathcal{B}_s(a_{1,1}, b_{1,1})$ and $\mathcal{B}_s(a_{2,2}, b_{2,2}) = \mathcal{B}_s(a_{1,2}, b_{1,2})$. $\mathcal{B}_s(a_{2,3}, b_{2,3})$ is a translation of $\mathcal{B}_s(a_{2,2}, b_{2,2})$ and $\mathcal{B}_s(a_{2,4}, b_{2,4})$ a translation of $\mathcal{B}_s(a_{2,1}, b_{2,1})$ along lines with slope -1, such that all four circles are non-overlapping. h_2 is characterized by adding or subtracting $\delta\kappa_{2,k}$ according to the parity of k, and is equal to zero everywhere else. Integrating h_2 along a line with slope -1 results in zero, while maintaining the property that integrating over a vertical line also results in zero.

Next, for $j = 2, \ldots, J$, we define a function $h_j(a, b)$ such that $g_j(a, b) = \widetilde{f}_{A,B}(a, b) + h_j(a, b)$ satisfies similar properties as (i) and (ii), namely:

(iii) g_j is a valid density;

(iv) $\widetilde{f}_{A,B}(a, b)$ and g_j have the same Radon transforms for j design points $x_1, \ldots, x_j$.

For each $j = 2, \ldots, J$, we recursively define a sequence of tuples $(a_{j,k}, b_{j,k})$ as

$$a_{j,k} = \begin{cases} a_{j-1,k} & 1 \leq k \leq 2^{j-1} \\ a_{j-1,2^j-k+1} + \tau_j \sin \arctan(-x_j) & 2^{j-1} < k \leq 2^j \end{cases}$$

and

$$b_{j,k} = \begin{cases} b_{j-1,k} & 1 \leq k \leq 2^{j-1} \\ b_{j-1,2^j-k+1} + \tau_j \cos \arctan(-x_j) & 2^{j-1} < k \leq 2^j \end{cases},$$

where τ_j is the minimum distance such that

$$\bigcup_{k=1}^{2^{j-1}} \mathcal{B}_s(a_{j,k}, b_{j,k}) \bigcap \bigcup_{k=2^{j-1}+1}^{2^j} \mathcal{B}_s(a_{j,k}, b_{j,k}) = \emptyset.$$

Note that $\tau_j < s2^{j+1}$, thus the union of the 2^j balls is always contained within $\mathcal{B}_r(a_0, b_0)$.

Similar to the construction of h_1, for $j = 2, \ldots, J$ and $k \leq 2^j$, let $\kappa_{j,k}(a, b) = \phi_s(a - a_{j,k}, b - b_{j,k})$. We now define the functions h_j as

$$h_j(a, b) = \delta \sum_{k=1}^{2^{j-1}} \left(\kappa_{j,2k-1}(a, b) - \kappa_{j,2k}(a, b) \right). \tag{A.2}$$

Note that all functions, $\kappa_{j,k}$, $k = 1, 2, \ldots, 2^j$ have non-overlapping supports, i.e., $\mathcal{B}_s(a_{j,k}, b_{j,k})$. For example, in Panel B of Figure A.1, $\mathcal{B}_s(a_{2,1}, b_{2,1})$ and $\mathcal{B}_s(a_{2,2}, b_{2,2})$ are identical to $\mathcal{B}_s(a_{1,1}, b_{1,1})$ and $\mathcal{B}_s(a_{1,2}, b_{1,2})$ from Panel A, respectively. In addition, $\mathcal{B}_s(a_{2,3}, b_{2,3})$ and $\mathcal{B}_s(a_{2,4}, b_{2,4})$ are the resulting open balls from translating $\mathcal{B}_s(a_{2,2}, b_{2,2})$ and $\mathcal{B}_s(a_{2,1}, b_{2,1})$, respectively, along the direction of lines with slope $-1/x_2$, by τ_2. Outside the union of $\mathcal{B}_s(a_{2,i}, b_{2,i})$ for $i = 1, 2, 3, 4$, h_2 is zero. Here, $\mathcal{B}_s(a_{2,1}, b_{2,1})$ and $\mathcal{B}_s(a_{2,3}, b_{2,3})$ are depicted as dashed circles to indicate that the associated functions $\kappa_{2,1}$ and $\kappa_{2,3}$ are added, as shown in (A.2). Similarly, $\mathcal{B}_s(a_{2,2}, b_{2,2})$ and $\mathcal{B}_s(a_{2,4}, b_{2,4})$ are depicted with solid circles, and the associated functions $\kappa_{2,2}$ and $\kappa_{2,4}$ are subtracted. As will

be shown below, integrating h_2 over lines parallel to either dash-dotted lines in the right Panel of Figure A.1 is zero, i.e. the Radon transforms for both x_1 and x_2 are constant zero.

Now, we will show that g_j satisfies (iii) for each $j = 2, \ldots, J$. By (A.2) and by the definition of τ_j, the support of h_j is the union of the non-overlapping open balls $\mathcal{B}_s(a_{j,k}, b_{j,k})$ for $k = 1, \ldots, 2^j$. Thus the minimum of h_j is $-\delta$. Note that because the diameter of $\cup_k \mathcal{B}_s(a_{j,k}, b_{j,k})$ is less than r, it must be that $\cup_k \mathcal{B}_s(a_{j,k}, b_{j,k}) \subset \mathcal{B}_r(a_0, b_0)$. Then, because $f_{A,B} \geq \delta$ on $\mathcal{B}_r(a_0, b_0)$, it follows that $g_j(a, b) \geq f_{A,B}(a, b) - \delta \geq 0$. Next notice that by (A.2), the integral of h_j can be evaluated as

$$\iint h_j(a, b)\mathrm{d}a\mathrm{d}b = 2^{j-1}\delta \int_{\mathcal{B}_s(0,0)} \phi_s(a, b)\mathrm{d}a\mathrm{d}b - 2^{j-1}\delta \int_{\mathcal{B}_s(0,0)} \phi_s(a, b)\mathrm{d}a\mathrm{d}b = 0,$$

hence g_j integrates to one.

To show (iv), we can rearrange the $\kappa_{j,k}$ functions in (A.2) so that h_j can be equivalently defined as

$$h_j(a, b) = h_{j-1}(a, b) - h_{j-1}(a - \tau_j \sin \arctan(-x_j), b - \tau_j \cos \arctan(-x_j)). \tag{A.3}$$

Due to the recursive nature of (A.3), h_j can be expressed in terms of the h_k functions for any $k \in 1, \ldots, j-1$. Hence to show (iv), it is sufficient to demonstrate that h_j has a Radon transform that is constant zero for x_j. The Radon transform of h_j for x_j is

$$\begin{aligned}\int h_j(y - bx_j, b)\mathrm{d}b = &\int h_{j-1}(y - bx_j, b)\mathrm{d}b \\ &- \int h_{j-1}(y - bx_j - \tau_j \sin \arctan(-x_j), b - \tau_j \cos \arctan(-x_j))\mathrm{d}b.\end{aligned}$$

By a change of variables of $b' = b - \tau_j \cos \arctan(-x_j)$,

$$\int h_{j-1}(y - bx_j - \tau_j \sin \arctan(-x_j), b - \tau_j \cos \arctan(-x_j))\mathrm{d}b = \int h_{j-1}(y - b'x_j, b')\mathrm{d}b'.$$

Therefore, h_j has a Radon transform that is constant zero for x_j.

Thus, we have constructed a probability density function $g_J(a,b)$, different from $\widetilde{f}_{A,B}$ on a set with positive Lebesgue measure, where the Radon transforms of g_J and $\widetilde{f}_{A,B}$ are equal for each $x \in \mathcal{X}$.

A.0.2 Analytic Forms of $f^3_{A,B}$ and $f^4_{A,B}$

The following are the analytic forms of $f^3_{A,B}$ and $f^4_{A,B}$

$$f^3_{A,B}(a,b;d) = \begin{cases} \widetilde{f}_{A,B}(a,b) - \min\left(\widetilde{f}_{A,B}(a,b), \widetilde{f}_{A,B}(a+dm_3(b), b-2dm_3(b))\right) & b \notin (-d,0] \\ \widetilde{f}_{A,B}(a,b) + \sum_{k=1}^{\infty} \min\left(\widetilde{f}_{A,B}(a,b+kd), \widetilde{f}_{A,B}(a+kd,b-kd)\right) & \\ \qquad + \sum_{k=1}^{\infty} \min\left(\widetilde{f}_{A,B}(a-kd,b+kd), \widetilde{f}_{A,B}(a,b-kd)\right) & b \in (-d,0] \end{cases}$$

$$f^4_{A,B}(a,b;d) = \begin{cases} \widetilde{f}_{A,B}(a,b) - \min\left(\widetilde{f}_{A,B}(a,b), \widetilde{f}_{A,B}(a+dm_4(b), b-2dm_4(b))\right) & b \notin [0,d) \\ \widetilde{f}_{A,B}(a,b) + \sum_{k=1}^{\infty} \min\left(\widetilde{f}_{A,B}(a,b+kd), \widetilde{f}_{A,B}(a+kd,b-kd)\right) & \\ \qquad + \sum_{k=1}^{\infty} \min\left(\widetilde{f}_{A,B}(a-kd,b+kd), \widetilde{f}_{A,B}(a,b-kd)\right) & b \in [0,d) \end{cases}$$

where d is a positive constant. The functions $m_i : \mathbb{R} \to \mathbb{Z}$, for $i = 3, 4$, map $b \in \mathbb{R}$ to an integer such that $b - dm_3(b) \in (-d, 0]$ and $b - dm_4(b) \in [0, d)$. In Panels D and E of Figure 2.1, both functions with $d = 2/33$ are shown.

As an example of a distribution that is in $\mathcal{E}$ but is not absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^2$, consider the following:

$$F^5_{A,B}\left((-\infty,a],(-\infty,b]\right) = \Phi_2\left((a,b);M_1\right) - p\Phi_2\left((a,b);M_2\right) + p\Phi_1\left(a;1\right)\mathbb{I}\{b \geq 0\},$$

where $\Phi_2(\cdot, \Sigma)$ is a bivariate normal distribution function centered at $(0, 0)$ with covariance matrix Σ, and $\Phi_1(\cdot, \sigma^2)$ is a normal distribution with mean 0 and variance σ^2. It can be verified that for

$$M_1 = \begin{bmatrix} 1 & -1 \\ -1 & 3 \end{bmatrix},$$
$$M_2 = \begin{bmatrix} 1 & -1 \\ -1 & 2 \end{bmatrix},$$

and for $0 < p < 0.7071$, $F^5_{A,B}$ is a valid distribution in $\mathcal{E}$.

A.0.3 Proof of Lemma 1

Suppose $(a_0, b_0) \in \text{supp}\left(\widetilde{f}_{A,B}\right)$. To show $(a_0, b_0) \in \bigcap_{x\in\mathcal{X}} Q_x^{-1}(\text{supp}(f_{Y_x}))$ it is sufficient to show $y_{0,x} = a_0 + b_0 x \in \text{supp}\left(f_{Y_x}\right)$ for each $x \in \mathcal{X}$.

For $x \in \mathcal{X}$ and $\epsilon > 0$, we have

$$\begin{aligned}
\mathbb{P}\left(Y_x \in (y_{0,x} - \epsilon, y_{0,x} + \epsilon)\right) &= \int_{y_{0,x}-\epsilon}^{y_{0,x}+\epsilon} \int_{\mathbb{R}} \widetilde{f}_{A,B}\left(y_{0,x} - bx, b\right) \mathrm{d}b\mathrm{d}y \\
&= \int_{\mathbb{R}} \int_{-bx+y_{0,x}-\epsilon}^{-bx+y_{0,x}+\epsilon} \widetilde{f}_{A,B}(a, b)\mathrm{d}a\mathrm{d}b \\
&\geq \iint_{\mathcal{B}_\epsilon(a_0,b_0)} \widetilde{f}_{A,B}(a, b)\mathrm{d}a\mathrm{d}b > 0.
\end{aligned}$$

Here, the open ball, $\mathcal{B}_\epsilon(a_0, b_0)$, in $\mathbb{R}^2$ centered at (a_0, b_0) with radius ϵ is a subset of the region $Q_x^{-1}(y_{0,x} - \epsilon, y_{0,x} + \epsilon)$. Thus, $\mathbb{P}\left(Y_x \in (y_{0,x} - \epsilon, y_{0,x} + \epsilon)\right) > 0$, which implies that $(a_0, b_0) \in Q_x^{-1}\left(\text{supp}\left(f_{Y_x}\right)\right)$ for all $x \in \mathcal{X}$.

A.0.4 Local Solutions are Valid Densities

Let $f^{(0)}_{A,B}$ be an arbitrary density whose support contains Λ_0, and let $f^{(1)}_{A,B,x}$ be the update according to (2.3) for $x \in \mathcal{X}$. In addition, let μ_2 be the Lebesgue measure on $\mathbb{R}^2$. We will show

that $f^{(1)}_{A,B,x}$ is a valid density function with support supp $\left(f^{(0)}_{A,B}\right) \bigcap Q_x^{-1}(\text{supp}(f_{Y_x}))$, and the Radon transform of $f^{(1)}_{A,B,x}$ is f_{Y_x}.

To show that $f^{(1)}_{A,B,x} > 0$ μ_2-almost everywhere on supp $\left(f^{(0)}_{A,B}\right) \bigcap Q_x^{-1}(\text{supp}(f_{Y_x}))$ and is zero μ_2-almost everywhere else, it is enough to show that supp $\left(f^{(0)}_{Y_x}\right)$ contains supp (f_{Y_x}). Let $y_0 \in$ supp (f_{Y_x}). Then for $\epsilon > 0$,

$$0 < \int_{y_0-\epsilon}^{y_0+\epsilon} f_{Y_x}(t)\mathrm{d}t = \int_{y_0-\epsilon}^{y_0+\epsilon} \int_{\mathbb{R}} f_{A,B}(y_0 - bx, b)\mathrm{d}b.$$

Because supp $(f_{A,B}) \subseteq$ supp $\left(f^{(0)}_{A,B}\right)$,

$$0 < \int_{y_0-\epsilon}^{y_0+\epsilon} \int_{\mathbb{R}} f^{(0)}_{A,B}(y_0 - bx, b)\mathrm{d}b = \int_{y_0-\epsilon}^{y_0+\epsilon} f^{(0)}_{Y_x}(t)\mathrm{d}t.$$

Therefore, $y_0 \in$ supp $\left(f^{(0)}_{Y_x}\right)$.

We finish by simultaneously showing that $f^{(1)}_{A,B,x}$ integrates to 1 and its Radon transform for x is f_{Y_x}. By (2.3) and by a change of variables,

$$\iint f^{(1)}_{A,B,x}(a,b)\mathrm{d}a\mathrm{d}b = \int \left(\int f^{(0)}_{A,B}(y - sx, s) \frac{f_{Y_x}(y)}{f^{(0)}_{Y_x}(y)} \mathrm{d}s \right) \mathrm{d}y.$$

We recognize the inside integral on the right hand side as the Radon transform of $f^{(1)}_{A,B,x}$ for x, which is equal to

$$\int f^{(0)}_{A,B}(y - sx, s)\mathrm{d}s \frac{f_{Y_x}(y)}{f^{(0)}_{Y_x}(y)} = f^{(0)}_{Y_x}(y) \frac{f_{Y_x}(y)}{f^{(0)}_{Y_x}(y)} = f_{Y_x}(y).$$

Then

$$\iint f^{(1)}_{A,B,x}(a,b)\mathrm{d}a\mathrm{d}b = \int f_{Y_x}(y)\mathrm{d}y = 1.$$

A.0.5 Bregman Projection onto $\mathcal{E}_x$

It is desired to show that

$$F^{(1)}_{A,B,x} = \underset{\widetilde{F}_{A,B}\in\mathcal{E}_x}{\arg\min}\, \mathcal{D}_{KL}(\widetilde{F}_{A,B}||F^{(0)}_{A,B}).$$

By the principle of minimum discrimination information, $F^{(1)}_{A,B,x}$ achieves the minimum if $\mathcal{D}_{KL}(F^{(1)}_{A,B,x}||F^{(0)}_{A,B}) = \mathcal{D}_{KL}(F_{Y_x}||F^{(0)}_{Y_x})$. Consider the following

$$\begin{aligned}\mathcal{D}_{KL}(F^{(1)}_{A,B,x}||F^{(0)}_{A,B}) &= \iint f^{(1)}_{A,B,x}(a,b)\log\frac{f^{(1)}_{A,B,x}(a,b)}{f^{(0)}_{A,B}(a,b)}\mathrm{d}a\mathrm{d}b \\ &= \iint f^{(0)}_{A,B}(a,b)\frac{f_{Y_x}(a+bx)}{f^{(0)}_{Y_x}(a+bx)}\log\frac{f_{Y_x}(a+bx)}{f^{(0)}_{Y_x}(a+bx)}\mathrm{d}a\mathrm{d}b.\end{aligned}$$

By a change of variables, $y = a + bx$ and $s = b$, the last iterative integral above can be expressed as

$$\iint f^{(0)}_{A,B}(y-sx,s)\frac{f_{Y_x}(y)}{f^{(0)}_{Y_x}(y)}\log\frac{f_{Y_x}(y)}{f^{(0)}_{Y_x}(y)}\mathrm{d}s\mathrm{d}y = \int f_{Y_x}(y)\log\frac{f_{Y_x}(y)}{f^{(0)}_{Y_x}(y)}\mathrm{d}y.$$

A.0.6 Proof of Theorem 2

For ease of notation, where convenient we write f in place of $\widetilde{f}_{A,B}$, $f^{(\ell)}$ in place of $f^{(\ell)}_{A,B}$, and denote the ratio of the output density functions as $\xi^{(\ell)}_x$, i.e., for $\ell = 0, 1, \ldots$

$$\xi^{(\ell)}_x(a,b) = \frac{f_{Y_x}(a+bx)}{f^{(\ell)}_{Y_x}(a+bx)}.$$

Then, (2.5) can be written as

$$f^{(\ell+1)}(a,b) = f^{(\ell)}(a,b)\frac{1}{J}\sum_{x\in\mathcal{X}}\xi^{(\ell)}_x(a,b), \quad \text{for } \ell = 0, 1, \ldots. \tag{A.4}$$

We begin by showing (i). The KL-divergence of f with respect to $f^{(\ell+1)}$ is

$$\mathrm{E}_f \log \frac{f}{f^{(\ell+1)}} = \mathrm{E}_f \log \frac{f}{f^{(\ell)}} - \mathrm{E}_f \log \left(\sum_{\mathcal{X}} \xi_x^{(\ell)} \frac{1}{J} \right).$$

Rearranging the terms above, we obtain the following expression

$$\begin{aligned} \mathrm{E}_f \log \frac{f}{f^{(\ell)}} - \mathrm{E}_f \log \frac{f}{f^{(\ell+1)}} &= \mathrm{E}_f \log \left\{ \sum_{\mathcal{X}} \xi_x^{(\ell)} \frac{1}{J} \right\} \\ &\geq \sum_{\mathcal{X}} \frac{1}{J} \mathrm{E}_f \log \xi_x^{(\ell)}, \end{aligned} \tag{A.5}$$

where the inequality is a direct consequence of Jensen's inequality.

Next, we will show that $\mathrm{E}_f \log \xi_x^{(\ell)} \geq 0$ for $\ell = 0, 1, \ldots$ and $x \in \mathcal{X}$. Straightforward calculation yields

$$-\mathrm{E}_f \log \xi_x^{(\ell)} = \iint \log \frac{f_{Y_x}^{(\ell)}(a+bx)}{f_{Y_x}(a+bx)} f(a,b) \mathrm{d}a\mathrm{d}b.$$

Recall that $\log(x) \leq x - 1$ for all $x > 0$. Thus, we have

$$\begin{aligned} \iint \log \frac{f_{Y_x}^{(\ell)}(a+bx)}{f_{Y_x}(a+bx)} f(a,b) \mathrm{d}a\mathrm{d}b &\leq \iint \frac{f(a,b)}{f_{Y_x}(a+bx)} f_{Y_x}^{(\ell)}(a+bx) \, \mathrm{d}a\mathrm{d}b \;\; - \;\; 1 \\ &= \int \left(\int \frac{f(y-sx,s)}{f_{Y_x}(y)} \mathrm{d}s \right) f_{Y_x}^{(\ell)}(y) \mathrm{d}y - 1 \\ &= \int f_{Y_x}^{(\ell)}(y) \mathrm{d}y - 1 = 0. \end{aligned}$$

Hence,

$$\mathrm{E}_f \log \xi_x^{(\ell)} \geq 0.$$

Thus, by recalling (A.5), we conclude that

$$\mathrm{E}_f \log \frac{f}{f^{(\ell)}} - \mathrm{E}_f \frac{f}{f^{(\ell+1)}} \geq 0,$$

for any ℓ. Therefore, the KL-divergences of f with respect to $f^{(\ell)}$ is a non-increasing sequence with a lower bound of zero and by the monotone convergence theorem has a convergence.

Next, we show (iii). From (i) and (A.5), we have $1/J \sum_{\mathcal{X}} \mathrm{E}_f \log \xi_x^{(\ell)} \to 0$ as $\ell \to \infty$. Therefore, $\mathrm{E}_f \log \xi_x^{(\ell)} \to 0$ as $\ell \to \infty$, for each $x \in \mathcal{X}$. By a straightforward change of variables, it can be seen that $\mathrm{E}_f \log \xi_x^{(\ell)}$ equals the KL-divergence of f_{Y_x} from $f_{Y_x}^{(\ell)}$; hence, $\mathrm{E}_f \log \xi_x^{(\ell)} \geq 0$. Therefore, $\mathrm{E}_f \log \xi_x^{(\ell)} \to 0$, for each $x \in \mathcal{X}$.

Finally, we show (ii). By the convexity property of the KL-divergence,

$$\begin{aligned}\mathcal{D}_{KL}(f^{(\ell+1)}||f^{(\ell)}) &= \mathcal{D}_{KL}\left(\frac{1}{J}\sum_x f_{A,B,x}^{(\ell+1)}||\frac{1}{J}\sum_x f_{A,B}^{(\ell)}\right) \\ &\leq \frac{1}{J}\sum_x \mathcal{D}_{KL}(f_{A,B,x}^{(\ell+1)}||f_{A,B}^{(\ell)}).\end{aligned}$$

Within the proof in A.0.5, it was shown that $\mathcal{D}_{KL}(f_{A,B,x}^{(\ell+1)}||f_{A,B}^{(\ell)}) = \mathcal{D}_{KL}(f_{Y_x}||f_{Y_x}^{(\ell)})$. Hence, $\mathcal{D}_{KL}(f^{(\ell+1)}||f^{(\ell)}) \leq 1/J \sum_x \mathcal{D}_{KL}(f_{Y_x}||f_{Y_x}^{(\ell)})$, which goes to zero by (iii).

A.0.7 (Q_x, μ_1)-disintegration of μ_2

We show that the (Q_x, μ_1)-disintegration of μ_2, denoted by $\{\mu_{Q_x^{-1}(y)}\}$ is the set of disintegrating measures defined by,

$$\mu_{Q_x^{-1}(y)}(f) = \int f(y - sx, s)\mathrm{d}\mu_1(s),$$

where f is a measurable function. For $\{\mu_{Q_x^{-1}(y)} : y \in \mathbb{R}\}$ to be a (Q_x, μ_1)-disintegration of μ_2, it must satisfy the following three conditions given in Chang and Pollard (1997):

(i) $\mu_{Q_x^{-1}(y)}$ is a σ-finite measure on $\mathcal{B}_\Lambda$ concentrated on the set $\{Y = y\}$;

and, for each nonnegative measurable function g on $\mathbb{R}^2$:

(ii) $y \to \mu_{Q_x^{-1}(y)}(g)$ is $\mathcal{B}_\mathbb{R}$ measurable;

(iii) $\mu_2(g) = \mu_1\left(\mu_{Q_x^{-1}(y)}(g)\right)$.

For (i), let $f_y(a,b) = \mathbb{I}\{(a,b) \in \Lambda : Q_x(a,b) \neq y\}$. Then

$$\mu_{Q_x^{-1}(y)}(f_y) = \int \mathbb{I}\{(y - sx, s) \in \Lambda : y \neq y\} \mathrm{d}\mu_1(s) = 0.$$

Additionally, define disjoint sets $\{C_i\}$ such that for each $i \in \mathbb{Z}$, $C_i = \mathbb{I}\{(a,b) \in \Lambda : i \leq b < i+1\}$. The countable union $\bigcup_{i \in \mathbb{Z}} C_i = \Lambda$. It can be seen that $\mu_{Q_x^{-1}(y)}(C_i) = \sqrt{1+x^2} < \infty$ for each i. Condition (ii) holds since $\mu_{Q_x^{-1}(y)}(f)$ is a composition of measurable operations on g.

For (iii), let g be a nonnegative measurable function on Λ. Then

$$\mu_2(g) = \int g(a,b) \mathbb{I}\{(a,b) \in \Lambda\} \, \mathrm{d}\mu_2(a,b).$$

By a change of variables, $\mu_2(g)$ can be expressed as

$$\mu_2(g) = \int g(y - sx, x) \mathbb{I}\{(y - sx, s) \in \Lambda\} \, \mathrm{d}\mu_2(s,y) = \int \left(\int g \mathrm{d}\mu_{Q_x^{-1}(y)} \right) \mathrm{d}\mu_1$$

which is equal to $\mu_1\left(\mu_{Q_x^{-1}(y)}(g)\right)$ as desired.

Appendix B

Mixing Distribution Recovery Appendix

B.0.1 Proof of Theorem 3

Note that, since $\mathcal{D}_{KL}(G||G^0) < \infty$, G^0 dominates G (Kullback, 1959, 1968). Consider $\mathcal{D}_{KL}(G||G^j) - \mathcal{D}_{KL}(G||G^{j+1})$, which is

$$\begin{aligned}\mathcal{D}_{KL}(G||G^j) - \mathcal{D}_{KL}(G||G^{j+1}) &= \mathrm{E}_G \log \frac{g}{g^j} - \mathrm{E}_G \log \frac{g}{g^{j+1}} \\ &= \mathrm{E}_G \log \mathrm{E}_{P(\cdot|\theta)} \frac{f}{f^j} \\ &\geq \mathrm{E}_F \log \frac{f}{f^j} = \mathcal{D}_{KL}(F||F^j) \geq 0,\end{aligned}$$

where the first inequality is a consequence of Jensen's inequality and the law of total expectation. Since $\mathcal{D}_{KL}(G||G^0) < \infty$, it follows from the inequality above that, for each $j = 0, 1, \ldots$, $\mathcal{D}_{KL}(G||G^j) < \infty$ and hence, G^j dominates G. In addition, the above inequality shows that for each j

$$\mathcal{D}_{KL}(G||G^j) \geq \mathcal{D}_{KL}(G||G^{j+1}).$$

Therefore, by the monotone convergence theorem, the sequence of KL-divergences $\{\mathcal{D}_{KL}(G||G^j)\}$ has a convergence. Then, $\mathcal{D}_{KL}(G||G^j) - \mathcal{D}_{KL}(G||G^{j+1}) \to 0$, and hence, $\mathcal{D}_{KL}(F||F^j) \to 0$.

B.0.2 Proof of Theorem 4

Consider $\mathcal{D}_{KL}(G^j||G^{j+1})$, which is

$$\mathcal{D}_{KL}(G^j||G^{j+1}) = -\mathrm{E}_{G^j} \log \mathrm{E}_{P(\cdot|\theta)} \frac{f}{f^j} \leq \mathcal{D}_{KL}(F^j||F)$$

where the last inequality holds by Jensen's inequality, the law of total expectation, and basic logarithm properties. By assumption $\mathcal{D}_{KL}(F^j||F) \to 0$ as $j \to 0$, thus, $\mathcal{D}_{KL}(G^j||G^{j+1}) \to 0$ as $j \to 0$.

B.0.3 Proof of Theorem 5

By Pinsker's inequality, and the relation between total variation and the L_1 metric, it follows that $\int |g^j(\theta) - g^{j+k}(\theta)|\mathrm{d}\gamma(\theta) \leq \sqrt{2\mathcal{D}_{KL}(G^j||G^{j+k})}$. Thus, consider the KL-divergence of G^j from G^{j+k},

$$\begin{aligned}
\mathcal{D}_{KL}(G^j||G^{j+k}) &= -\sum_{\ell=0}^{k-1} \mathrm{E}_{G^j} \log \mathrm{E}_{P(\cdot|\theta)} \frac{f}{f^{j+\ell}} \\
&\leq \sum_{\ell=0}^{k-1} \mathrm{E}_{F^j} \log \frac{f^{j+\ell}}{f} \\
&= \mathcal{D}_{KL}(F^j||F) + \sum_{\ell=1}^{k-1} \left(\mathcal{D}_{KL}(F^j||F) - \mathcal{D}_{KL}(F^j||F^{j+\ell})\right),
\end{aligned}$$

where the inequality is a consequence of Jensen's inequality and the law of total expectation. Recalling the definition of S_j,

$$\mathcal{D}_{KL}(F^j||F) + \sum_{\ell=1}^{k-1} \left(\mathcal{D}_{KL}(F^j||F) - \mathcal{D}_{KL}(F^j||F^{j+\ell})\right) \leq \mathcal{D}_{KL}(F^j||F) + S_j,$$

for any k. By assumption, $\mathcal{D}_{KL}(F^j||F)$ and S_j converge to zero as j increases to infinity. Then for any $\epsilon > 0$, we can find a J such that for all $j > J$, $\mathcal{D}_{KL}(F^j||F) + S_j < \epsilon$. Therefore, for any $\epsilon > 0$ we can find a J such that for $j > J$ and any k, $\int |g^j(\theta) - g^{j+k}(\theta)|\mathrm{d}\gamma(\theta) \leq \sqrt{2\mathcal{D}_{KL}(G^j|G^{j+k})} < \epsilon$. Then, the sequence of L_1 distances between G^j and G^m is a Cauchy sequence in the L_1 metric space, which is complete, and hence has a convergence G^∞. Then because $L_1(G^j, G^\infty)$ is proportional to the total variation between G^j and G^∞, the convergence of $G^j \to G^\infty$ in the L_1 metric implies convergence in total variation.

Recall $\mathcal{D}_{KL}(F||F^j) \to 0$ as $j \to \infty$, so by Pinsker's inequality,

$$\sup_{A \in \mathcal{B}_{\mathcal{X}}} |F^j(A) - F(A)| \to 0$$

as $j \to \infty$. Hence, for any $A \in \mathcal{B}_{\mathcal{X}}$,

$$F(A) = \lim_{j \to \infty} F^j(A) = \lim_{j \to \infty} \int P(A|\theta) g^j(\theta) \mathrm{d}\gamma(\theta). \tag{B.1}$$

For any $A \in \mathcal{B}_{\mathcal{X}}$, $P(A|\theta)$ is bounded, and $P(A|\theta)$ is continuous, by the Helly-Bray theorem, the limit in (B.1) can be passed under the integral to obtain

$$F(A) = \lim_{j \to \infty} F^j(A) = \lim_{j \to \infty} \int P(A|\theta) g^j(\theta) \mathrm{d}\gamma(\theta) = \int P(A|\theta) g^\infty(\theta) \mathrm{d}\gamma(\theta) = F^\infty(A).$$