NAVAL POSTGRADUATE SCHOOL

MONTEREY, CALIFORNIA

THESIS

DENSITY DECONVOLUTION WITH EPI-SPLINES

by

James C. Carbaugh

September 2015

Thesis Co-Advisors:	Johannes O. Royset
	Wei Kang
Second Reader:	Samuel E. Buttrey

Approved for public release; distribution is unlimited

THIS PAGE INTENTIONALLY LEFT BLANK

REPORT DOCUMENTATION PAGE

Form Approved OMB No. 0704–0188

1. AGENCY USE ONLY *(Leave Blank)*	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	09-2015	Master's Thesis 07-08-2013 to 09-25-2015

4. TITLE AND SUBTITLE

DENSITY DECONVOLUTION WITH EPI-SPLINES

5. FUNDING NUMBERS

6. AUTHOR(S)

Carbaugh, James C.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)

Naval Postgraduate School
Monterey, CA 93943-5000

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)

N/A

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

The views expressed in this document are those of the author and do not reflect the official policy or position of the Department of Defense or the U.S. Government. IRB Protocol Number: N/A.

12a. DISTRIBUTION / AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(maximum 200 words)*

Analysts tasked with developing probability density estimates may obtain data in sets of varying quality and quantity. Often low-fidelity data contaminated with measurement error, or "noise," may be abundant, but the cost of obtaining data free of these errors will limit the amount of high-fidelity data available. In such a scenario, the problem is to identify an estimate of a probability density function given scarce high-fidelity observations, knowledge of measurement errors, abundant "noisy" data, and available user knowledge of the density apart from empirical data. We solve this rich class of deconvolution problems through constrained optimization with first-order epi-splines, which are used for the first time to approximate densities to an arbitrarily high level of precision. We limit our scope to univariate densities where measurement errors are homoscedastic. Demonstrations come from a benchmark from the literature, historical medical data, and a scenario in uncertainty quantification in fluid dynamics. Results show that deconvolution via epi-splines is viable, comparable with a widely available deconvolution method, and shows potential for savings in resource budgets for data collection.

14. SUBJECT TERMS	15. NUMBER OF PAGES
Deconvolution, Probability Density Estimation, Epi-splines, Optimization, Functional Approximation, Non-Parametric Statistics, Uncertainty Quantification	73
	16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UU

NSN 7540-01-280-5500

Standard Form 298 (Rev. 2–89)
Prescribed by ANSI Std. 239–18

THIS PAGE INTENTIONALLY LEFT BLANK

Approved for public release; distribution is unlimited

DENSITY DECONVOLUTION WITH EPI-SPLINES

James C. Carbaugh
Lieutenant, United States Navy
B.S., U.S. Naval Academy, 2007

Submitted in partial fulfillment of the
requirements for the degrees of

MASTER OF SCIENCE IN OPERATIONS RESEARCH
and
MASTER OF SCIENCE IN APPLIED MATHEMATICS
from the

NAVAL POSTGRADUATE SCHOOL
September 2015

Author: James C. Carbaugh

Approved by: Johannes O. Royset, Ph.D.
Thesis Co-Advisor

Wei Kang, Ph.D.
Thesis Co-Advisor

Samuel E. Buttrey, Ph.D.
Second Reader

Patricia Jacobs, Ph.D.
Chair, Department of Operations Research

Craig Rasmussen, Ph.D.
Chair, Department of Applied Mathematics

THIS PAGE INTENTIONALLY LEFT BLANK

ABSTRACT

Analysts tasked with developing probability density estimates may obtain data in sets of varying quality and quantity. Often low-fidelity data contaminated with measurement error, or "noise," may be abundant, but the cost of obtaining data free of these errors will limit the amount of high-fidelity data available. In such a scenario, the problem is to identify an estimate of a probability density function given scarce high-fidelity observations, knowledge of measurement errors, abundant "noisy" data, and available user knowledge of the density apart from empirical data. We solve this rich class of deconvolution problems through constrained optimization with first-order epi-splines, which are used for the first time to approximate densities to an arbitrarily high level of precision. We limit our scope to univariate densities where measurement errors are homoscedastic. Demonstrations come from a benchmark from the literature, historical medical data, and a scenario in uncertainty quantification in fluid dynamics. Results show that deconvolution via epi-splines is viable, comparable with a widely available deconvolution method, and shows potential for savings in resource budgets for data collection.

THIS PAGE INTENTIONALLY LEFT BLANK

Table of Contents

List of Figures

THIS PAGE INTENTIONALLY LEFT BLANK

List of Tables

THIS PAGE INTENTIONALLY LEFT BLANK

List of Acronyms and Abbreviations

D/L drag/lift coefficient

DOD Department of Defense

FT Fourier Transform

FFT Fast Fourier Transform

FHS Framingham Heart Study

GAMS General Algebraic Modeling System

iid independent and identically distributed

lsc lower semicontinuous

MEP maximum entropy problem

MEP-E first-order epi-spline maximum entropy problem

MLP maximum log-likelihood problem

MLP-E first-order epi-spline maximum log-likelihood problem

MSE mean squared error

SBP systolic blood pressure

UQ uncertainty quantification

THIS PAGE INTENTIONALLY LEFT BLANK

Executive Summary

Analysts faced with scarce budgets for experimentation and probability density estimation must also account for the measurement error, sometimes called "noise," inherent in the data they acquire. Additionally, the cost to acquire data and the fidelity of the data obtained are often positively correlated. Scientists in the Department of Defense (DOD) use simulations to great effect as a resource saving measure, but simulations can vary widely in accuracy and computation time.

Our goal is to generate accurate univariate probability density estimates that can blend high-fidelity and low-fidelity data, account for noise, and make use of all available information about a given system. With inputs such as prior knowledge of measurement error, noisy and accurate data, and knowledge of the density apart from empirical observations (i.e., "soft" information), we address this rich class of deconvolution problems using constrained optimization with first-order epi-splines.

Density deconvolution reduces the noise inherent in estimates generated from observations contaminated with errors. Epi-splines have been shown to approximate density functions to an arbitrarily high level of accuracy. This thesis blends deconvolution and epi-splines, and is the first to use first-order epi-splines, which allow for closed form solutions of the convolution integral when errors are homoscedastic. First-order epi-splines also provide a convenient framework for imposing shape constraints generated by soft information.

We explore three cases to demonstrate our method. First, we show evidence of the accuracy of epi-spline deconvolution starting from a benchmark in the deconvolution literature. Second, we show that we can obtain deconvolution results comparable with a widely available software package. In an uncertainty quantification example from fluid dynamics, we produce density estimates blending high-fidelity and low-fidelity data showing that we can supplement small accurate data sets with abundant noisy data sets to produce density estimates comparable with those obtained exclusively by abundant accurate data. The result is that we can achieve dramatic reduction in computational expense with potential for further savings in other applications where the observations in a noisy data set contain identical and independently distributed errors.

THIS PAGE INTENTIONALLY LEFT BLANK

Acknowledgments

I am deeply grateful to my co-advisor, Dr. Johannes O. Royset, for his enormous patience, generous support, and innovative vision. I also thank my co-advisor, Dr. Wei Kang, and my second reader, Dr. Samuel E. Buttrey, for their encouragement, good humor, and insightful contributions.

My thanks goes also to the Operations Research and Applied Mathematics Departments, especially department chairs Dr. Patricia Jacobs, Dr. Robert Dell, Dr. Carlos Borges, and Dr. Craig Rasmussen, program officers Captain Jeffrey Hyink, Commander Thor Martinsen, Lieutenant Commander Connor McLemore, and academic associates Dr. Ron Fricker, Dr. Matt Carlyle, and Bard Mansager.

I express my warm gratitude to all those who enhanced my academic foundations, including Dr. Lyn Whitaker, Dr. Mathias Kolsch, Dr. Tom Lucas, Dr. Kyle Lin, Dr. Javier Salmeron, Dr. Ned Dimitrov, Dr. Jeremy Kozdon, Dr. Lucas Wilcox, Lieutenant Colonel Mark Raffetto, and Lieutenant Colonel Jon Roginski.

For encouraging me to pursue a dual degree program, I would like to thank Dr. Ralucca Gera and Lieutenant Phil Baxa.

My officemates, Captain Erik Rye, and Lieutenant Sam Chen, provided listening ears, and priceless advice.

I am forever grateful to my lovely wife, Hanna Carbaugh, for her countless sacrifices and tremendous endurance throughout this process. You understood the beauty of mathematics and patiently waited as I wrestled with my studies. Thank you for everything.

To my daughter, Mary, I have been so richly blessed by your arrival and I am overwhelmed by all the joy you have brought us.

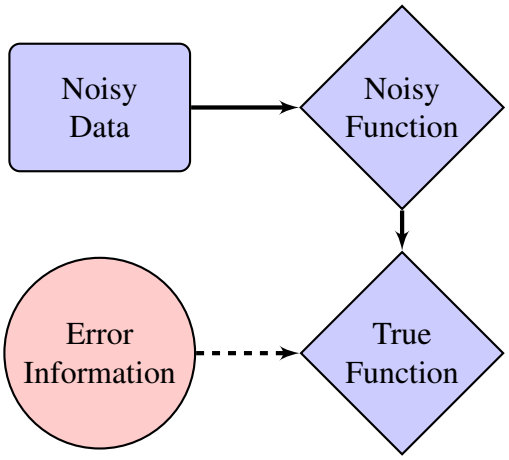
Finally, I must acknowledge above all my Lord and Savior Jesus Christ, whom, "For now we see through a glass, darkly; but then face to face: now I know in part; but then shall I know even as also I am known." –1 Corinthians 13:12

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 1:
Background

Measurement errors are inherent in all empirical data. When measurement errors are significant, generating estimates based on contaminated data may lead to poor results. This measurement error, often referred to as "noise," may be addressed by the application of a technique called deconvolution. Figure 1.1 shows a visualization of deconvolution in the broadest sense.

Figure 1.1: Deconvolution



The process known as deconvolution produces an estimate of a true function of interest. Methods and applications vary widely. Knowledge of errors is not always assumed.

We demonstrate a new method of deconvolution, that fuses hard data, "soft" information, and knowledge of the distribution of noise, to generate a probability density function for a random variable of interest. Hard data include noisy (and possibly some noise-free) empirical data. "Soft" information is defined as available user-provided system knowledge, such as continuity, monotonicity, tail convexity, or bounds on the density. We limit our scope to the estimation of univariate probability density functions where the noise is homoscedastic. This process is formulated as a constrained infinite-dimensional optimization problem, that is solved through an approximation by means of epi-splines.

Several distinct data sets provide opportunities for demonstration.

We continue this chapter with a brief literature review of the topics of deconvolution, the additive measurement model, and an introduction to the nascent application of epi-spline technology. The modern applications in the field of deconvolution may be subdivided, but not limited to, three main categories: signal processing, image processing, and probability density estimation. Epi-spline technology has also been used in probability density estimation, but this thesis represents the first use of first-order epi-splines. A qualitative overview of these topics is given to the reader in order to understand the significance of the various techniques.

1.1 Deconvolution

Deconvolution is the process by which we obtain a solution for f in the convolution of f and g defined as:

$$(f * g)(t) = \int_{-\infty}^{\infty} f(t-\tau)g(\tau)d\tau \tag{1.1}$$

In Equation (1.1), we say that f and g are convolved. Deconvolution is well-known as an ill-posed problem and falls into a larger class of inverse problems. For a tutorial review of the deconvolution problem, see [1].

1.1.1 Signal Processing

The application of deconvolution to gain inference on an input signal from one distorted by noise or some impulse response has one of its earliest applications in exploratory geophysics. In this field, a shot is fired into the ground and echoes as it passes through sedimentary layers of the earth. The echoes of the shot are recorded by a receiver and analyzed to make predictions about the various underground layers of the earth.

Enders Robinson's 1954 dissertation, and related paper, explain the use of time series decomposition of wavelets to extract a filtered time signal [2], [3] and expands on earlier work in statistics to use Fourier Transform (FT) methods for the analysis of stationary time series. These papers predate the use of the term deconvolution to explain such a process of extracting a time series signal filtered of errors, but they serve as an excellent

introduction to the application of deconvolution in geophysics. Further work on deconvolution in geophyics is given in [4]–[8].

Deconvolution in signal processing is not limited to geophysics. Since the intent is to extract a more accurate time series signal from a signal convolved with noise, the applications are vast. A seminal paper on the use of Fourier techniques in deconvolution for spectroscopy is given in [9]. Further treatment on deconvolution in spectroscopy may be found in [10]. Deconvolution has also been used widely in pharmacology to understand the effects of various substances on test subjects [11], [12]. Whereas in geophysics, a shot may be fired into the ground, in pharmacokinetics, a signal is generated by the intake of a certain drug.

1.1.2 Blind Deconvolution

Deconvolution refers to the process of recovering an input function or impulse function given a known output. Deconvolution may be subdivided into two different kinds of problems: one where the input function is accessible in some form, and one where it is not [13]. The scenario in which the input is unknown is known as *blind deconvolution*. Blind deconvolution presents a much more challenging problem than deconvolution in its traditional form since the input is unknown.

The adjective "blind" in the term *blind deconvolution* refers to the unknown nature of the input signal. One major application of blind deconvolution algorithms is in image processing. In this field, an image may be blurred noise taken to be from a point spread function (i.e., distortion with a mathematical description). To compute the true image, it is necessary to estimate the point spread function. Once an estimate is achieved, an improved picture follows. The widespread proliferation of digital photography has significantly raised the visibility of this field [14]. A thorough review of blind deconvolution techniques for imagery is given in [14]. Helpful visualizations of the performance of blind deconvolution techniques are given in [15].

1.1.3 Density Deconvolution

Density deconvolution refers to the application of deconvolution techniques in order to perform inference on a density of an input random variable given observations of an output variable. Density deconvolution is more restrictive than other deconvolution methods; the estimate of the density function of interest produced by the deconvolution method must adhere to axioms of probability such as nonnegativity and integration to one. A deconvolution method that uses FT may be highly accurate over large portions of the domain, but violate nonnegativity in the tails. The density estimate must then be amended in order to be an actual probability density function. In Chapter 2 we demonstrate how an epi-spline framework can use constraints to ensure whatever estimate is found via deconvolution meets the requirements of a density function.

1.2 Additive Measurement Model

Nonparametric deconvolution problems in statistics often involve density estimation of unobserved variables in measurement models of the form

$$Y = X + \varepsilon \tag{1.2}$$

where the problem is to identify the density of X given n independent and identically distributed (iid) observations of Y. In this case, each observation of X and ε are unknown, but the density of the errors ε may be assumed based on prior knowledge. Such a model is often described as the additive measurement model [16]. Within such a framework the literature is often focused on the nature of the error density [17]–[19], as well as rates of convergence [20]–[22]. For a detailed tutorial on density deconvolution; see [16]. For a comprehensive treatment of measurement error; see [23].

The R [24] package **decon** contains an example FT and kernel methods used in deconvolution [25]. The supporting literature provides a brief survey of fields in which measurement error may be significant, including medicine, bioinformatics, chemistry, astronomy, and econometrics, as well as an extensive review of kernel based methods and bandwidth selection methods for several cases. The authors compute density estimates using the Fast Fourier Transform (FFT) instead of the definitions provided and vastly re-

duce their computation time. Functions included in the R package **decon** can handle both Gaussian and Laplacian errors, as well as homoscedastic and heteroscedastic errors [25].

1.3 Epi-Splines

Royset and Wets have addressed the problem of "how to identify a function that best represents data and also satisfies information-driven constraints" [26] with the use of epi-splines. These epi-splines are piecewise polynomial functions described by a finite number of parameters [27]. By partitioning the domain of a function into a finite number of segments, epi-splines allow for the estimation of a function by selecting certain epi-parameters that define the estimate of the function in every segment. A solution of a functional identification problem generated by epi-splines will contain parameters such as slopes and intercepts in a first-order epi-spline setup or coefficients of polynomials more generally. Epi-splines have been used for a growing variety of applications, including financial tools, electricity demand forecasts, and probability density estimation [26]–[30].

This thesis shows the first computational use of first-order epi-splines. It also expands upon existing methodology with the addition of a convolution constraint in order to estimate the density of the input variable, X. First-order epi-splines allow for closed-form expressions of the convolution integral. These expressions preclude the necessity of numerical integration and the approximations and increases in computational cost that are unavoidable with numerical methods. The use of convolution as a constraint is explained fully in Section 2.3.3.

By combining an objective function such as maximum likelihood with constraints for convolution and soft information, we can deconvolve a probability density function generated from noisy data. At a minimum, the estimate of the probability density function must be entirely nonnegative and integrate to one. Additional user information may include support, continuity, differentiability, convex tails, among other characteristics. We explain this process completely in Chapter 2.

1.4 Overview

This thesis shows that constrained optimization using epi-spline technology may be a viable alternative for estimating the density of X in the additive measurement model of Equation (1.2) with Gaussian noise. One benefit of this approach is that we can use a sequence of constrained optimization models to combine information gained from both an abundant and noisy dataset and a scarce but accurate dataset.

We make our methodology explicit in Chapter 2, detailing steps for data processing, identifying objective functions, and listing all possible constraints of a given constrained optimization problem. In Chapter 3, we apply this method to an artificial data set of the convolution of a gamma density and Gaussian noise. Next, in Chapter 4, we compare our method with a widely available method on a medical data set of systolic blood pressure measurements. Finally, we show how this method may be used to fuse data from high-fidelity and low-fidelity simulations in an uncertainty quantification (UQ) scenario from fluid dynamics in Chapter 5. We conclude our findings in Chapter 6.

CHAPTER 2:
Methodology

2.1 Constrained Optimization Problems

We begin from the density estimation method of [27]. We consider an iid sample $x^1, \ldots, x^n$, of a random variable X, and a set F of lower semicontinuous (lsc) functions that are non-negative, integrate to one, and satisfy soft information about the density. The maximum log-likelihood problem (MLP) becomes

$$\textbf{MLP}: \max_f \sum_{i=1}^{n} \log(f(x^i)) \text{ s.t. } f \in F \tag{2.1}$$

where f satisfies additional criteria such as continuity, unimodality, tail convexity, etc., as well as a convolution constraint. The maximum entropy problem (MEP) becomes

$$\textbf{MEP}: \max_f -\int f(x)\log(f(x))dx \text{ s.t. } f \in F. \tag{2.2}$$

Given the additive measurement model shown in Equation (1.2), we denote the true density by f_X, a contaminated density by f_Y, and density of homoscedastic measurement error, by f_ε. The goal is to deconvolve the true density f_X from the contaminated density f_Y given knowledge of f_ε and soft information. We give a visual depiction of this process in Figure 2.1.

2.2 First-Order Epi-Splines

Following the setup in [26], we define first-order epi-splines. Given a closed interval $[l,u]$ in $\mathbb{R}$, we impose an evenly spaced mesh $m = \{m_k \mid k = 0,1,...N\}$ where $m_0 = l$ and $m_N = u$. An epi-spline density f is given with epi-parameters slope a^k and intercept a_0^k as

$$f(x) = a_0^k + a^k x \quad \text{for } x \in (m_{k-1}, m_k) \tag{2.3}$$

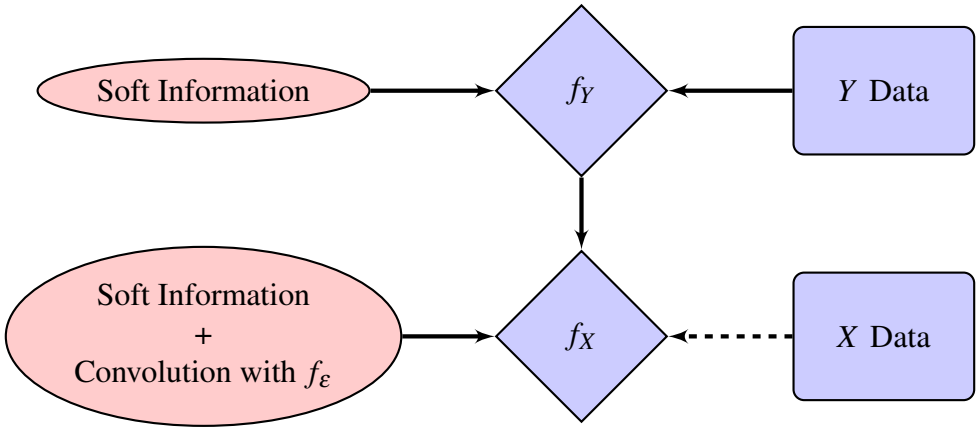
and for every $x \in [l,u]$, has

$$f(x) = \min\left\{ \lim_{t\downarrow 0} f(x+t),\ \lim_{t\downarrow 0} f(x-t) \right\}. \tag{2.4}$$

The second condition ensures that the value $f(x)$ is the smaller value of the left and right limits of the density at x. Figure 2.2 shows an epi-spline of order one on $\mathbb{R}$, that is piecewise affine and allows for jumps at mesh points [26].

Epi-splines are dense in the space of lsc functions under the epi-topology [27]. Thus, if the mesh is sufficiently fine, a first-order epi-spline can approximate to an arbitrary level of accuracy any lsc function [26]. Constrained optimization via epi-splines is therefore appropriate for estimation.

Figure 2.1: Deconvolution with Epi-Splines



Producing a deconvolved density f_X requires knowledge of a prior noisy density, f_Y, soft information and knowledge of measurement error density. Observations of X are not necessary for estimating f_X. They may be used if available.

2.2.1 First-Order Epi-Spline Optimization Problems

For computational purposes, we employ epi-splines to find approximate solutions of the given optimization problems. In the presence of n observations $x^1,\ldots,x^n$, we use a maximum log-likelihood objective function to obtain a density fit. Given a set $A \subset \mathbb{R}^{2N}$ of epi-parameters that ensure a resulting epi-spline density is nonnegative, integrates to

one, and satisfies soft information, the first-order epi-spline maximum log-likelihood problem (MLP-E) becomes

$$\textbf{MLP-E: } \max_{a_0^k, a^k} \sum_{i=1}^{n} \log(a_0^{k_i} + a^{k_i} x^i) \text{ s.t. } a \in A \subset \mathbb{R}^{2N}, \text{ where } k_i \text{ s.t. } x^i \in (m_{k-1}, m_k) \tag{2.5}$$

and the solution of epi-parameters $a = (a_0^1, a^1, a_0^2, a^2, \ldots, a_0^N, a^N)$ satisfies additional criteria such as constraints described in Section 2.3. We have an objective function that is concave in the epi-parameter. By the inclusion of the logarithm operator, the domain of the objective function is restricted to strictly positive values of the density at any interval.

When observations of X are not available, we can use a maximum entropy formulation. The first-order epi-spline maximum entropy problem (MEP-E) takes the form

$$\textbf{MEP-E: } \max_{a_0^k, a^k} -\sum_{k=1}^{N} \int_{m_{k-1}}^{m_k} (a_0^k + a^k x) \log(a_0^k + a^k x) dx \text{ s.t. } a \in A \subset \mathbb{R}^{2N}. \tag{2.6}$$

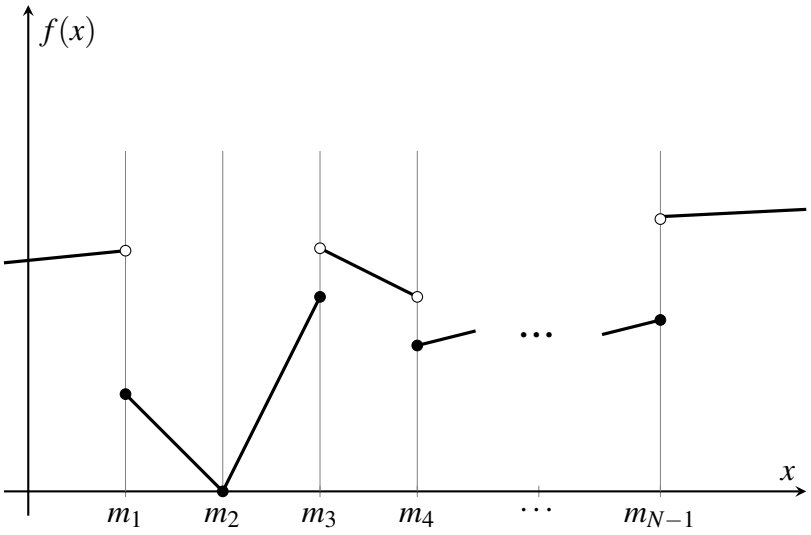
Since the entropy is composed with an affine function, and the density is entirely nonnegative, we have a concave objective function in the epi-parameter.

The density estimate solution of MLP-E or MEP-E, that we denote by $\hat{f}_X$, is then found by optimizing over $2N$ parameters. Though we focus on the f_X estimation problem here, we estimate f_Y in a similar manner, but without a convolution constraint or knowledge of measurement error.

2.3 Constraints

Detailed formula for our constraints are given in this section. Integration to one and nonnegativity are required for all densities and are described in 2.3.1 and 2.3.2. Convolution, described in 2.3.3, is required for the estimate of f_X only.

Figure 2.2: First-Order Epi-Splines



We show a mesh and epi-spline on $\mathbb{R}$. The epi-spline is defined by the epi-parameters slope and intercept on each interval. The value $f(x)$ is the smaller of the left and right limits at each mesh point.

2.3.1 Integrate to One

Densities must integrate to one. With a first-order epi-spline solution this integration takes the following form:

$$\sum_{k=1}^{N} \int_{m_{k-1}}^{m_k} a_0^k + a^k x dx = 1 \tag{2.7}$$

It follows that:

$$\sum_{k=1}^{N} a_0^k (m_k - m_{k-1}) + \sum_{k=1}^{N} \frac{a^k}{2} (m_k^2 - m_{k-1}^2) = 1 \tag{2.8}$$

is affine in the epi-parameter.

2.3.2 Nonnegativity

By constraining the function to be nonnegative at the endpoints of each interval, we ensure nonnegativity throughout the domain of the function. With the following pair of inequalities

$$a_0^k + a^k m_{k-1} \geq 0 \quad \forall k \tag{2.9}$$

$$a_0^k + a^k m_k \geq 0 \quad \forall k \tag{2.10}$$

we obtain halfspaces with respect to the epi-parameters.

2.3.3 Convolution

Recall that in the additive measurement model, the density of the noisy variable, f_Y, is the convolution of the density of the true variable, f_X, and the measurement error density, f_ε. Given a noisy density, f_Y, and knowledge of f_ε, we write the convolution equation as

$$f_Y(y) = \int f_\varepsilon(y-x) f_X(x) dx \quad \forall y. \tag{2.11}$$

Having already computed an estimate $\hat{f}_Y$, we can work to identify an estimate $\hat{f}_X$ such that $\hat{f}_Y = (\hat{f}_X * f_\varepsilon)$. We consider cases where the noise is Gaussian with known mean and variance, μ and σ^2, respectively, and f_X is approximated via first-order epi-splines. For computational purposes, we ensure that the right hand side of Equation (2.11) is close to $\hat{f}_Y$ with a tolerance δ and the inequality

$$\left| \hat{f}_Y(y) - \sum_{k=1}^{N} \int_{m_{k-1}}^{m_k} \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-x-\mu)^2 \big/ 2\sigma^2} (a_0^k + a^k x_k) dx \right| \leq \delta \tag{2.12}$$

for a finite number of y values evenly spaced within the interval $[m_0, m_N]$. With Gaussian measurement errors, we can obtain the closed form solution of these integrals. A nonzero δ is meaningful since we are working with estimates $\hat{f}_X$, $\hat{f}_Y$, and assumed knowledge of error f_ε, and there is uncertainty and possible errors in all three of these densities.

2.3.4 Density Continuity

Continuity may be required by the user. To ensure continuity, we set

$$a_0^k + a^k m_k = a_0^{k+1} + a^{k+1} m_{k-1} \quad \forall k = 1, \ldots, N-1 \tag{2.13}$$

which is affine with respect to the epi-parameters.

2.3.5 Non-increasing or Non-decreasing

Soft information may determine that the density of interest is nonincreasing. Since the density is not necessarily continuous, we use two constraints to establish a non-increasing function. First, we must have nonpositive derivatives, and secondly, we ensure that the function evaluated at the upper end of each interval is greater than or equal to the function evaluated at the lower end of the subsequent interval. We can use the following two inequalities.

$$a^k \leq 0 \quad \forall k \tag{2.14}$$

$$a_0^k + a^k m_k \geq a_0^{k+1} + a^{k+1} m_k \quad \forall k \tag{2.15}$$

These result in halfspaces with respect to the epi-parameters. We reverse the inequalities for non-decreasing estimates.

2.3.6 Unimodal

If soft information indicates the density is unimodal, the user can specify an interval of inflection points that contains the mode. We can achieve a unimodal estimate by requiring the function to be nondecreasing to the left of the left inflection point, I_L nonincreasing to the right of the right inflection point I_U, and requiring nonincreasing derivatives between the inflection points. With first-order epi-splines, five constraints are required for unimodality.

$$a^k \geq 0 \quad \text{for all } k \text{ with } m_k \leq I_L \tag{2.16}$$

$$a_0^k + a^k m_k \leq a_0^{k+1} + a^{k+1} m_k \quad \text{for all } k \text{ with } m_k \leq I_L \tag{2.17}$$

$$a^k \geq a^{k+1} \quad \text{for all } k \text{ with } m_k \geq I_L \text{ and } m_{k-1} \leq I_U \tag{2.18}$$

$$a^k \leq 0 \quad \text{for all } k \text{ with } m_{k-1} \geq I_U \tag{2.19}$$

$$a_0^k + a^k m_k \geq a_0^{k+1} + a^{k+1} m_k \quad \text{for all } k \text{ with } m_{k-1} \geq I_U \tag{2.20}$$

We find that the unimodal constraints are affine.

2.3.7 Tail Convexity

We can impose tail convexity for a function by ensuring that consecutive slopes are increasing outside of the inflection points. We introduce the following constraints enforced in addition to continuity:

$$a^k \leq a^{k+1} \quad \forall k \text{ with } m_k \leq I_L \tag{2.21}$$

$$a^k \leq a^{k+1} \quad \forall k \text{ with } m_{k-1} \geq I_U \tag{2.22}$$

to ensure convexity in the tails.

2.3.8 Maximum Change in Gradient

Since we have a first-order epi-spline framework, this epi-spline estimate is not differentiable. We can incorporate a certain "smoothness" by introducing a maximum change in slope, Δ, where

$$\left| a^k - a^{k+1} \right| \leq \Delta \quad \forall k = 1, \ldots, N-1. \tag{2.23}$$

2.4 Comparisons

We describe the method by which we measure the accuracy of our estimates numerically. When the true density is known a mean squared error (MSE) is computable as

$$\text{MSE} = \int_a^b (\hat{f}_X(x) - f_X(x))^2 f_X(x) dx. \tag{2.24}$$

which we integrate numerically via the midpoint rule.

For visual comparisons, we calculate kernel smoothing density estimates on certain data. Kernel estimates are generated using the standard kernel density estimator in MATLAB [31], a Gaussian kernel, and the default bandwidth calculation.

2.5 Procedures

We describe procedures for obtaining density estimates. Given an iid sample of Y, we obtain a solution to MLP-E, $\hat{f}_Y$, imposing constraints according to soft information. From

this solution, we record values of $\hat{f}_Y$ at evenly spaced intervals for use in the convolution constraint.

If an iid sample of X is available, we can obtain a solution of MLP-E given soft information and knowledge of measurement error. If no samples of X are available, we can obtain $\hat{f}_X$ as the solution to MEP-E. In addition to adhering to constraints imposed by soft information, MLP-E and MEP-E is constrained in estimates of f_X by convolution as described in Section 2.3.3. Chapter 3 and subsequent chapters will demonstrate how increasing soft information can improve estimates.

CHAPTER 3:
Gamma-Benchmark

3.1 Setup

We demonstrate the effectiveness of our model in simulations motivated by test instances in [32]. We consider an additive measurement model scenario where measurement errors have a normal density, f_ε, and f_X is a Gamma density. The convolution of these densities is f_Y.

3.1.1 Data and Noise

We generate 5000 samples of a random variable from a Gamma density with a certain shape, α, and scale, β, so that

$$X \sim Gamma(\alpha = 5, \beta = 1) \tag{3.1}$$

and 5000 measurement errors

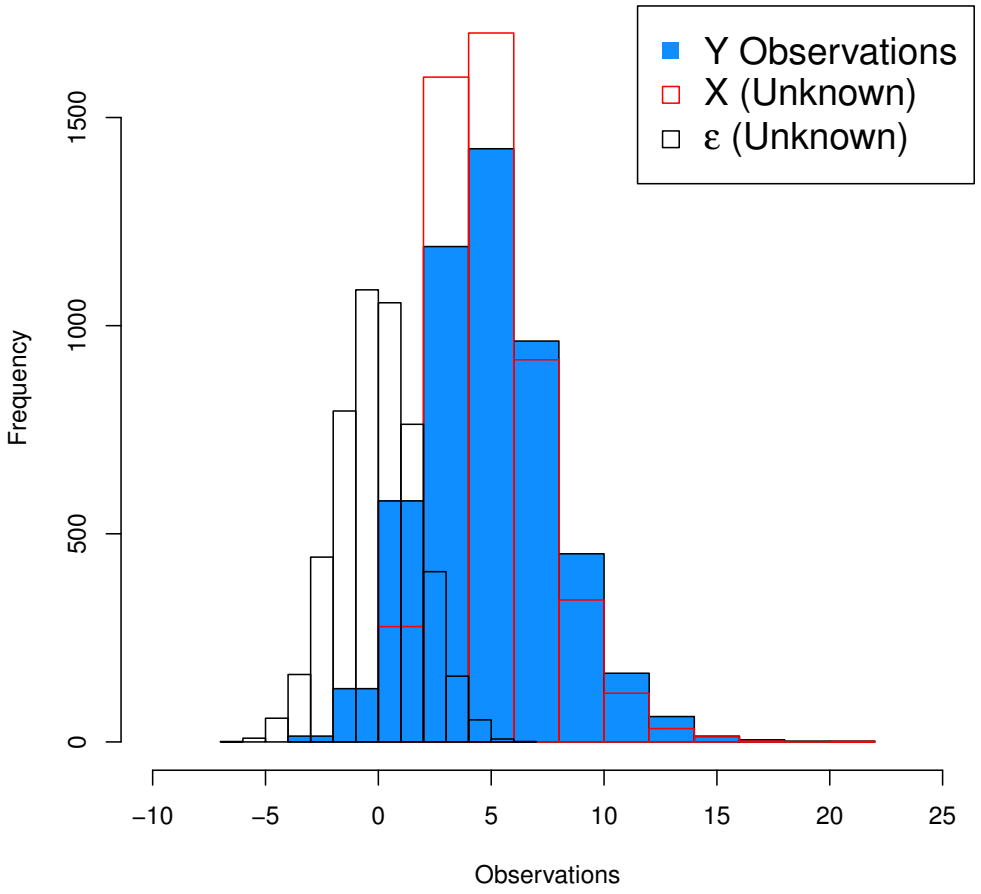
$$\varepsilon \sim Normal(\mu = 0, \sigma = \sqrt{3.2}). \tag{3.2}$$

Thus we record 5000 observations of Y with each observation being the sum of a single observation of X and a single observation of ε. A histogram of the observations of Y is given in Figure 3.1 and summary statistics for these observations is given in Table 3.1. We retain no knowledge of X or ε observations for density estimation but provide histograms of X and ε for illustrative purposes only in Figure 3.1. Separately, three observations of X are available for density estimation.

Table 3.1: Summary Statistics of Y Observations

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
−2.832	2.98	4.784	4.972	6.731	21.640

Figure 3.1: Sums of Gamma and Gaussian Observations



We generate 5000 observations of Y from the sums of 5000 samples of X and 5000 samples of ε. The blue histogram shows the frequency of Y and is filled to indicate the user's knowledge of these observations. The red and black outlined histograms show the frequency of the X and ε observations, respectively, that are unknown to the user.

3.1.2 Soft Information

Several items of soft information are available. Some settings we use for density estimation are the same for all the examples in this chapter and are shown in Table 3.2. The soft information we introduce incrementally is given in Table 3.3.

Table 3.2: Settings for Estimates of Artificial Data

Settings
Mesh Cardinality, $N = 1000$
Support, $[m_0, m_N]$: $[0, 23]$
Convolution Tolerance, $\delta = 5 \times 10^{-3}$
Convolution Values Checked: 101

Table 3.3: Soft Information on Density for Epi-Spline Estimates

Soft Information
Continuous
Unimodal
Inflection Points, $[I_L, I_U]$: $[\bar{Y} - \hat{\sigma}_{SE}, \bar{Y} + \hat{\sigma}_{SE}]$
Convex Tails
Maximum Change in Gradient, $\Delta = .00125$

$\bar{Y}$ and $\hat{\sigma}_{SE}$ represent the sample mean and estimated standard error.

3.2 Objective Criteria

The following sections document the performance of objective functions described in Section 2.3 in estimating f_X. We use MSE as a metric for accuracy as described in Section 2.4. To verify our results, we have run a simulation experiment detailed in Section 3.3.

3.2.1 Minimum Convolution Tolerance

Convolution is a constraint in MLP-E and MEP-E, we can also minimize the tolerance δ. This problem takes the form

$$\min_{a_0^k, a^k} \delta \text{ s.t. } a \in A \subset \mathbb{R}^{2N} \tag{3.3}$$

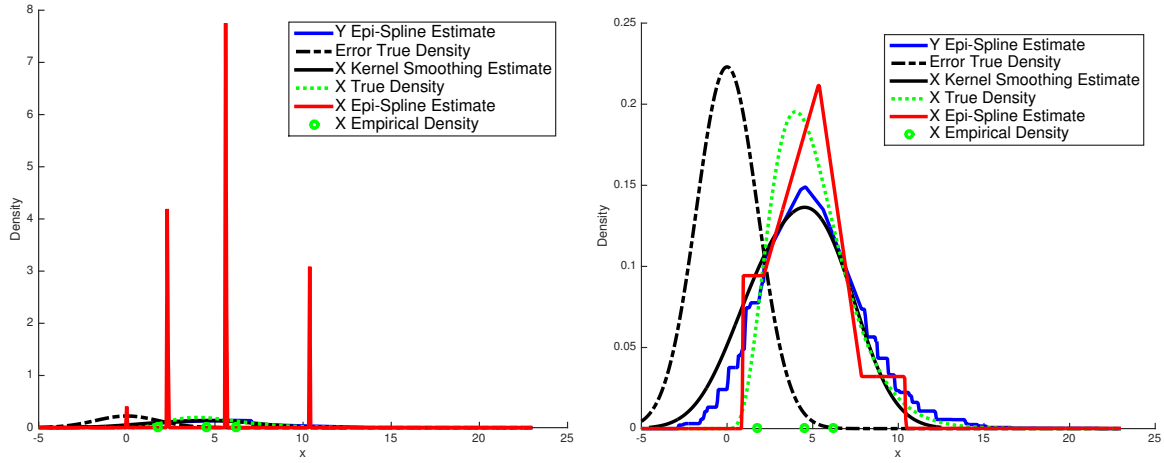
and the terms of the problem follow from Chapter 2. With extremely small values of δ or very large numbers of y values checked, MLP-E and MEP-E become infeasible. By relaxing the convolution tolerance slightly, we can incorporate the solution of this problem as a parameter in MLP-E or MEP-E.

3.2.2 Maximum Likelihood Estimation

We describe the results of MLP-E for the estimate of the density f_X and denote an optimal solution to this problem as $\hat{f}_X$.

We begin with the settings given in Table 3.2 and impose the continuity constraint. The result, $\hat{f}_X$, appears as a red line in Figure 3.2 and contains several distinct "spikes." Given the blue histogram shown in Figure 3.1, the user may reasonably believe that the underlying density is unimodal, with estimates for inflection points of approximately one sample standard deviation of Y from the sample mean, $\bar{Y}$. When we impose this constraint, the result improves dramatically as shown in Figure 3.2.

Figure 3.2: Continuous and Unimodal Maximum Likelihood Estimates



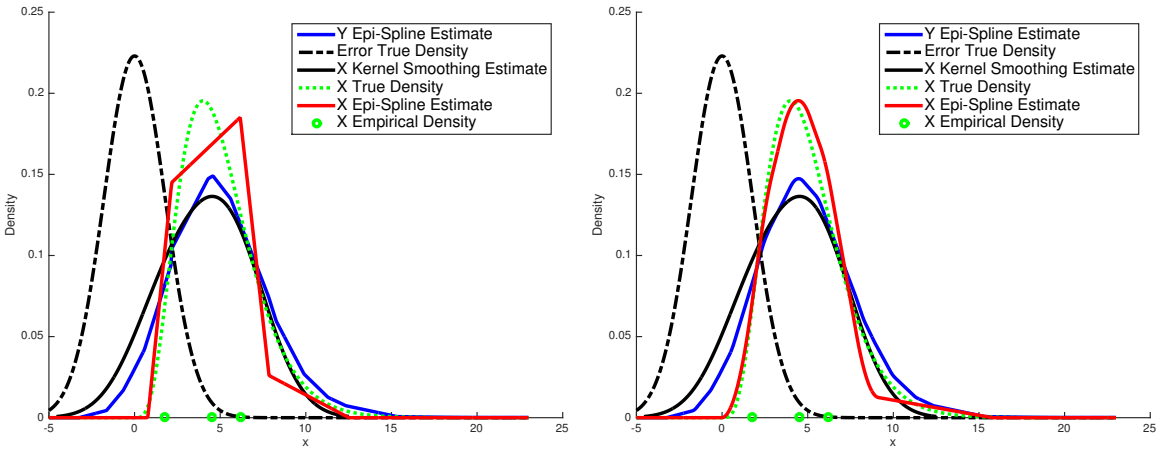
The blue line shows a solution of MLP-E for 5000 observations of Y. The dashed line shows the density of the Gaussian errors. The black line shows the density calculated from kernel methods using three observations of X. The dashed green line shows the true Gamma density. The red line shows the solution of MLP-E given three observations of X and certain soft information. We begin with continuity on the left and impose the unimodal constraint on the right. With continuity, MSE = 6.765×10^{-1}. The addition of the unimodal constraint improves the estimate dramatically; MSE = 9.3779×10^{-4}.

The visual "staircase" effect in Figure 3.2 may incline the user to impose convexity on the tails. We obtain the result of imposing this additional constraint in Figure 3.3, which still contains a sharp peak caused by a lack of bounds on gradient changes. By imposing such a bound, or a certain "smoothness," in addition to previously given constraints, $\hat{f}_X$ improves again. We document the numeric results of this evolving soft information in terms of MSE in Table 3.4.

To demonstrate the impact of the convolution constraint, we compare the same $\hat{f}_X$ of Figure 3.3 with an alternative produced without the convolution constraint. We show this

comparison in Figure 3.4. With convolution imposed as a constraint, the red line representing $\hat{f}_X$ has an extended right tail. Without the convolution constraint, the density appears nearly symmetric, producing a higher density estimate near the mode and eliminating the extended right tail. We see that the convolution constraint is key for identifying extended tail behavior.

Figure 3.3: Tail Convex and "Smooth" Maximum Likelihood Estimates



The blue line shows a solution of MLP-E for 5000 observations of Y. The dashed line shows the density of the Gaussian errors. The black line shows the density calculated from kernel methods using three samples of X, shown as green circles. The dashed green line shows the true Gamma density. The red line shows the solution of MLP-E given three samples of X and certain soft information. On the left, we have a density estimate that is continuous, and unimodal with convex tails. MSE = 7.3031×10^{-4}. On the right, we add "smoothness." MSE = 1.7953×10^{-4}.

3.2.3 Maximum Entropy Estimation

Without samples of X, we can use MEP-E, ensuring "maximum ignorance" in density estimation. An optimal solution of MEP-E is the most dispersed density possible given $\hat{f}_Y$ (i.e., an optimal solution of MLP-E for 5000 observations of Y), f_ε, and soft information. The orange line in Figure 3.5 shows the result of MEP-E. The dispersion of the density is evident in the right tail, which becomes apparently uniform and remains nonzero.

The advantage of such a method is clear in non-artificial examples when no high-fidelity data is available. With artificial data we have the ability to generate samples of X

and therefore can compare the results of MLP-E, MEP-E, and minimum δ for the same data and soft information. Table 3.5 shows a comparison of MSE and Figure 3.5 shows a graphical comparison. In a single trial, MEP-E produces the lowest MSE, but we offer repeated trials in Section 3.3 for further analysis.

Table 3.4: Error Reduction Due to Increasing Soft Information

Soft Information	MSE
Nonnegative Support	9.18×10^{-2}
Nonnegative Support, Continuous	6.765×10^{-1}
Nonnegative Support, Continuous, Unimodal	9.3779×10^{-4}
Nonnegative Support, Continuous, Unimodal, Convex Tails	7.3031×10^{-4}
Nonnegative Support, Continuous, Unimodal, Convex Tails, "Smoothness"	1.7953×10^{-4}

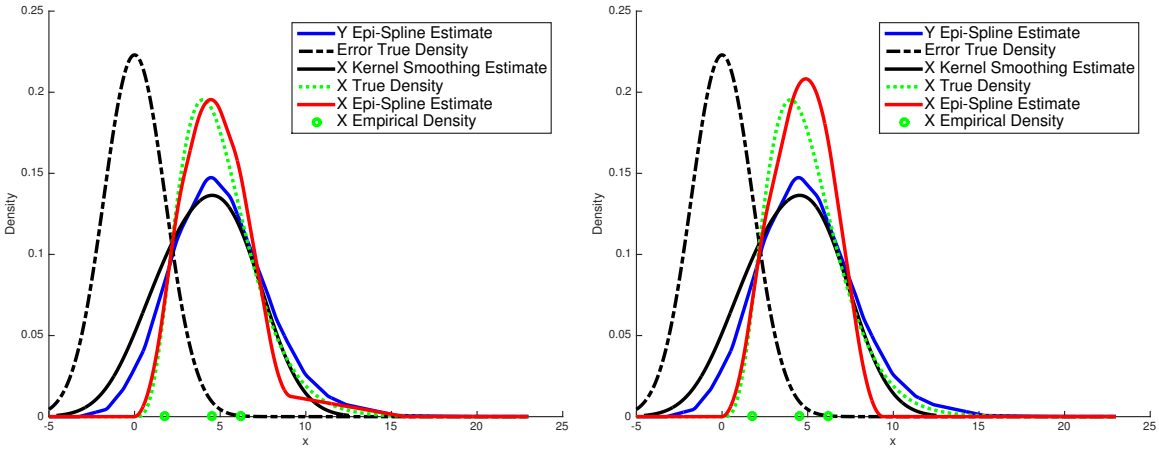
The results of MLP-E show how increasing soft information reduces MSE, with the largest reduction achieved through the addition of the unimodal constraint.

3.3 Numerical Experiments

We show through repeated trials that our results do not appear by chance. We generate 30 pairs of Y and X data sets (i.e., 5000 Y observations and three X observations), and obtain a solution of MLP-E for each pair. Separately we generate 30 sets of 5000 Y observations and obtain a solution $\hat{f}_X$ of MEP-E for each trial. The data produced for each replication of each experiment is unique. We record the MSE of each trial and compute averages of MSE for both sets of replications. More detailed pseudocode describing the execution of these replications is available within Algorithm 1 of the Appendix. For all trials in these experiments we include all the soft information given in Table 3.3.

We present our results in Figure 3.6. Consistent differences between the two methods become apparent in the right tails. Solutions of MEP-E often remain nonzero and become apparently uniform in the right tail. Solutions obtained via MLP-E exhibit more accurate right tail behavior. Soft information ensures that mean MSE for MLP-E and MEP-E do not vary widely on average.

Figure 3.4: The Effect of the Convolution Constraint



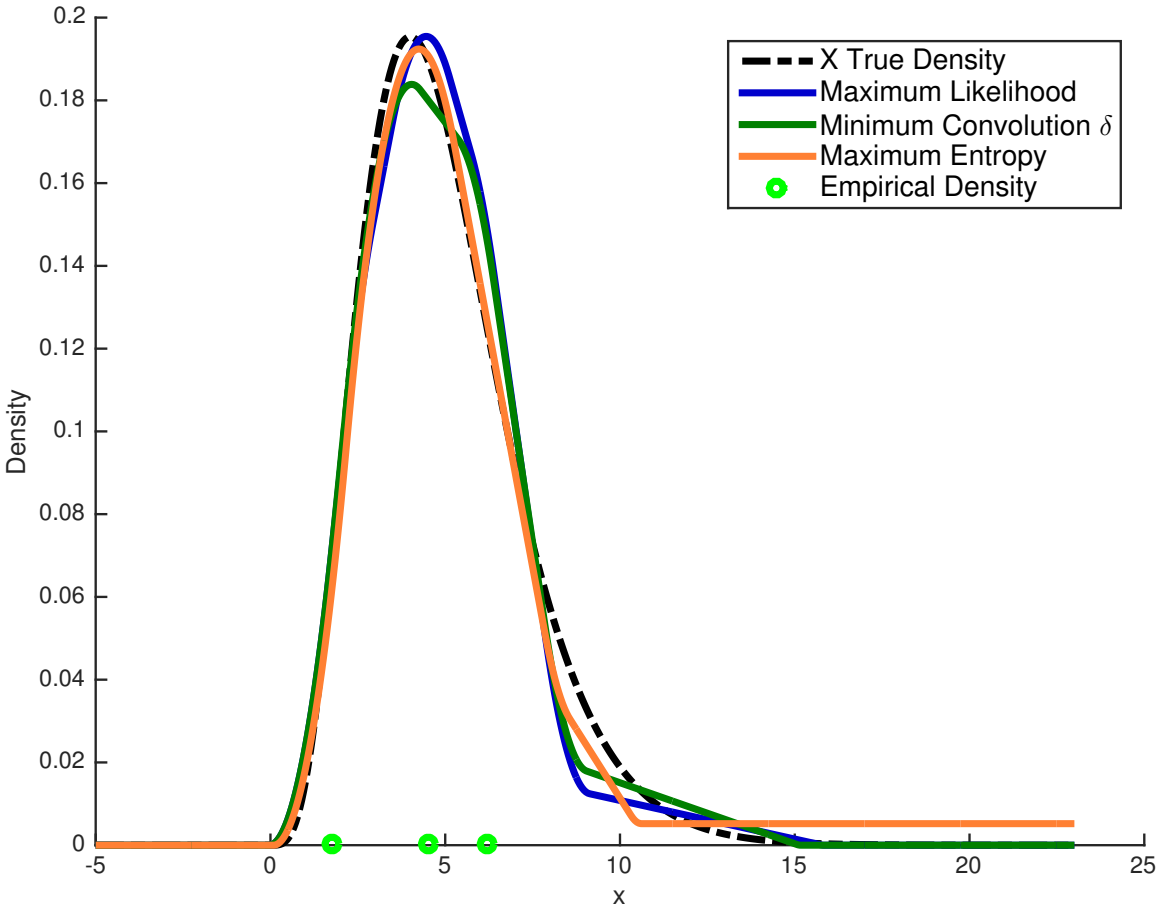
The blue line shows a solution of MLP-E for 5000 observations of Y. The dashed line shows the density of the Gaussian errors. The black line shows the density calculated from kernel methods using three samples of X, shown as green circles. The dashed green line shows the true Gamma density. The red line shows the solution of MLP-E given three samples of X and certain soft information. On the left, we include the convolution constraint, but on the right, we do not. By removing this constraint, we see that convolution is the key to right identifying tail behavior.

Table 3.5: Error Results for Differing Criteria

Problem	MSE
MLP-E	1.7953×10^{-4}
MEP-E	4.1694×10^{-5}
Minimum Convolution Tolerance	1.2398×10^{-4}

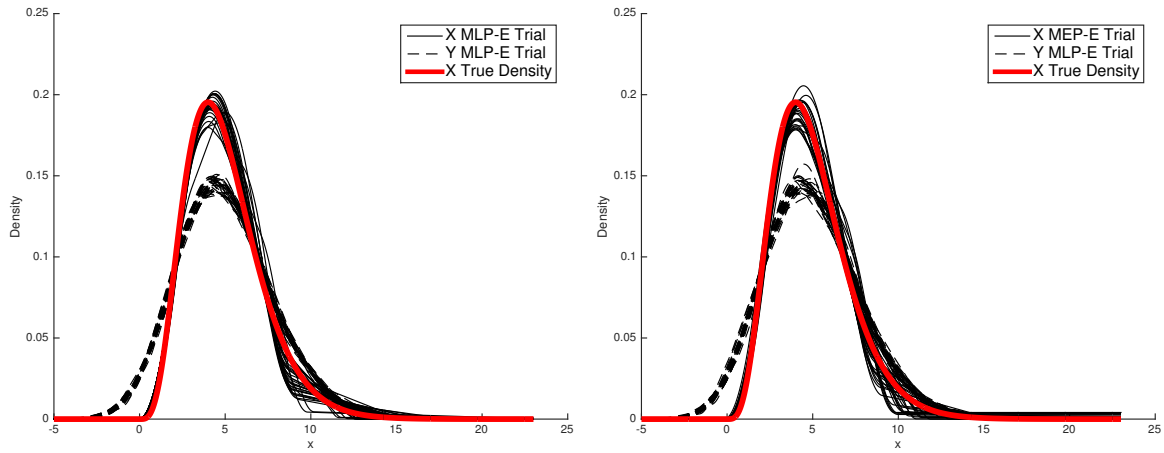
MSE is shown for the same data and soft information.

Figure 3.5: Comparison of Objective Functions



The black dashed line shows the true density. The blue line shows a solution of MLP-E given three samples of X, shown as green circles. The green line shows a solution found by minimizing the convolution tolerance, δ. The orange line is a solution of MEP-E. Since the observations of X near or below the mode, MLP-E gives the lowest density in the right tails. The density is most dispersed throughout its support in MEP-E, which creates a near uniform nonzero right tail. Minimizing convolution tolerance underestimates the density at the mode.

Figure 3.6: Repeated Trials of Epi-Spline Estimates



(a) Maximum Likelihood (b) Maximum Entropy

The solid black lines show solutions of MLP-E for three random samples of X in (a), and solutions of MEP-E in (b). The dashed lines are solutions of MLP-E for 5000 observations of Y. The red lines show the true density. Each trial is performed on distinct data sets. Repeated trials of MLP-E and MEP-E show slight differences in right tail behavior. With MLP-E, right tails approach zero, while with MEP-E right tails often become nearly uniform and nonzero. In (a), average MSE = 9.4×10^{-5}. In (b), average MSE = 1.14×10^{-4}.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 4:
Biostatistics Data

We now turn to the application of MEP-E on a medical data set and a comparison with another deconvolution method. Density deconvolution literature is abundant within the field of biostatistics. Through the R package **decon**, we have obtained real data on systolic blood pressure (SBP) from the Framingham Heart Study (FHS) described in [23]. The data contain homoscedastic measurement errors.

4.1 Deconvolution with R Package decon

The authors of [25] provide explicit instructions to use their deconvolution method for density estimation. They use FHS data from 1,615 male subjects whose SBP is measured and recorded twice at a first visit and then again measured and recorded twice at a second visit eight years later.

They compare the difference of the average SBP measurement at each visit for each subject and each difference is displayed in a histogram containing 1,615 differences within Figure 4.1. Measurement errors are assumed to be normally distributed based on the appearance of this histogram; therefore the differences are used to compute the standard deviation of measurement errors. The assumption of normality is justified further within [23].

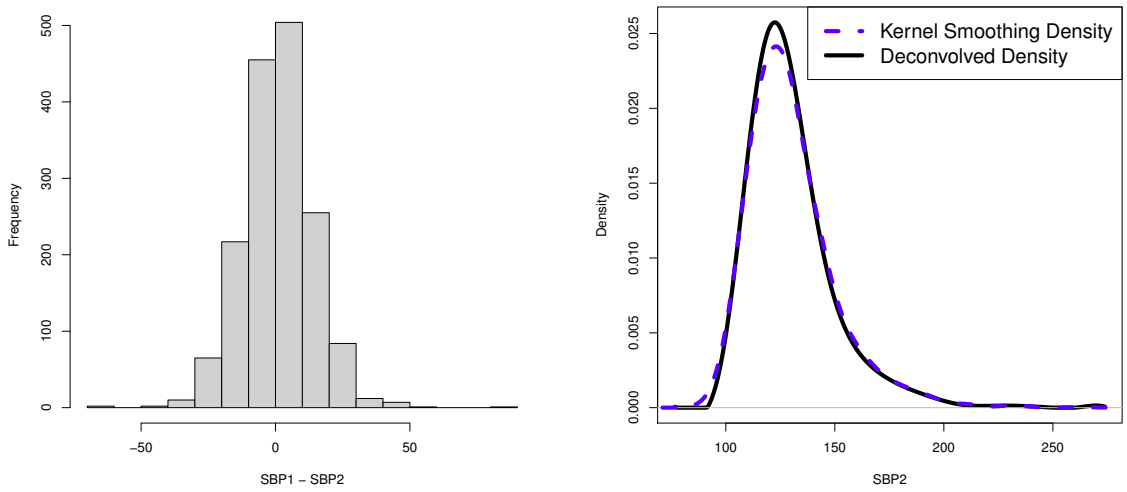
Their goal is to produce an accurate density of average SBP at the second visit by accounting for measurement error. The authors perform the deconvolution via a FFT algorithm, and their results on the SBP data are shown in Figure 4.1.

4.2 Deconvolution of Systolic Blood Pressure Data Using Epi-Splines

Our goal is to compare the results of MEP-E with the results achieved with **decon**. By incorporating estimated measurement error, the empirical data, and soft information, we

can achieve competitive results. The intent is to show that epi-spline estimates are capable of deconvolution in a non-artificial example.

Figure 4.1: The R package **decon** Results



On the left, we see the histogram of the differences of the average SBP reading at each visit. On the right, we see the results of deconvolution via **decon**. The blue dashed line shows a kernel smoothing density and the black solid line shows the deconvolved estimated density.

4.2.1 Setup

To begin, we relate the terms described in [25] to the additive measurement model of Equation (1.2). In **decon**, the numeric vectors containing each subject's average SBP reading at the first and second visit are defined as $SBP1$ and $SBP2$, respectively. Since the goal is to deconvolve $SBP2$, we describe each subject's second visit average SBP reading as an observation of Y. Therefore, we have 1,615 observations of Y, and we show summary statistics in Table 4.1.

Table 4.1: Summary Statistics of SBP2 Observations

Minimum	1st Quartile	Median	Mean	3rd Quartile	Maximum
87.5	117.0	126.5	130.0	139.5	263.0

No observations of X are available. Following the assumptions in [25] the esti-

mate of measurement error is given as

$$\varepsilon \sim Normal(\mu = 0, \sigma = \sqrt{.5\text{var}(SBP1 - SBP2)}. \tag{4.1}$$

We first obtain a solution $\hat{f}_Y$ of MLP-E given 1,615 observations of Y. Second, we use $\hat{f}_Y$ and knowledge of measurement error to obtain an optimal solution $\hat{f}_X$ of MEP-E. Table 4.2 provides the settings that remain unchanged for all estimates and Table 4.3 includes the soft information we impose incrementally.

Table 4.2: Settings for Estimates of SBP

Settings
Mesh Cardinality, $N = 1000$
Support, $[m_0, m_N]$: $[80, 265]$
Convolution Tolerance, $\delta = 2 \times 10^{-3}$
Convolution Values Checked: 101

Table 4.3: Soft Information for Systolic Blood Pressure Density

Soft Information
Continuous
Unimodal
Inflection Points, $[I_L, I_U]$: $[110, 140]$
Convex Tails
Maximum Change in Gradient, $\Delta = 3.75 \times 10^{-5}$

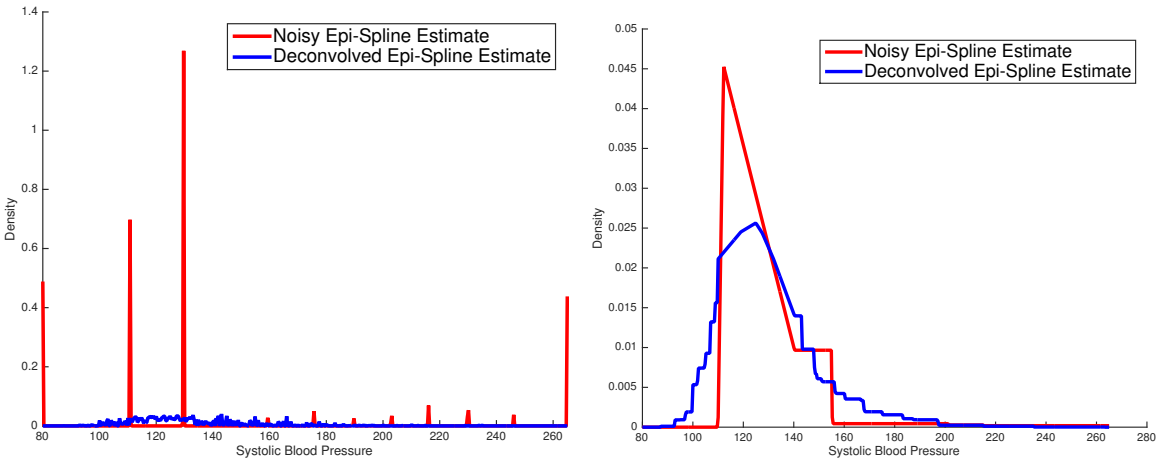
4.2.2 Evolving Soft Information

We impose soft information in a sequence similar to Chapter 3. In all figures within this chapter, the blue lines show optimal solutions of MLP-E of the given noisy SBP data and the red lines show the deconvolved estimate from optimal solutions of MEP-E.

We show the change in density estimates as we increase the soft information included in our estimate from continuity to unimodality in Figure 4.2. In a continuous estimate, we see how MEP-E disperses the density with peaks at the lower and upper bounds of the support. Again, the effect of imposing the unimodal constraint is dramatic. The unimodal estimate reduces the density in the tails and increases the density at the mode,

but produces a "staircase" effect and a sharp peak. As shown in Chapter 3, we can eliminate these effects with the constraints of tail convexity and maximum changes in gradient, which we denote by Δ. We include all soft information in Section 4.3.

Figure 4.2: Evolving Soft Information for Systolic Blood Pressure Density



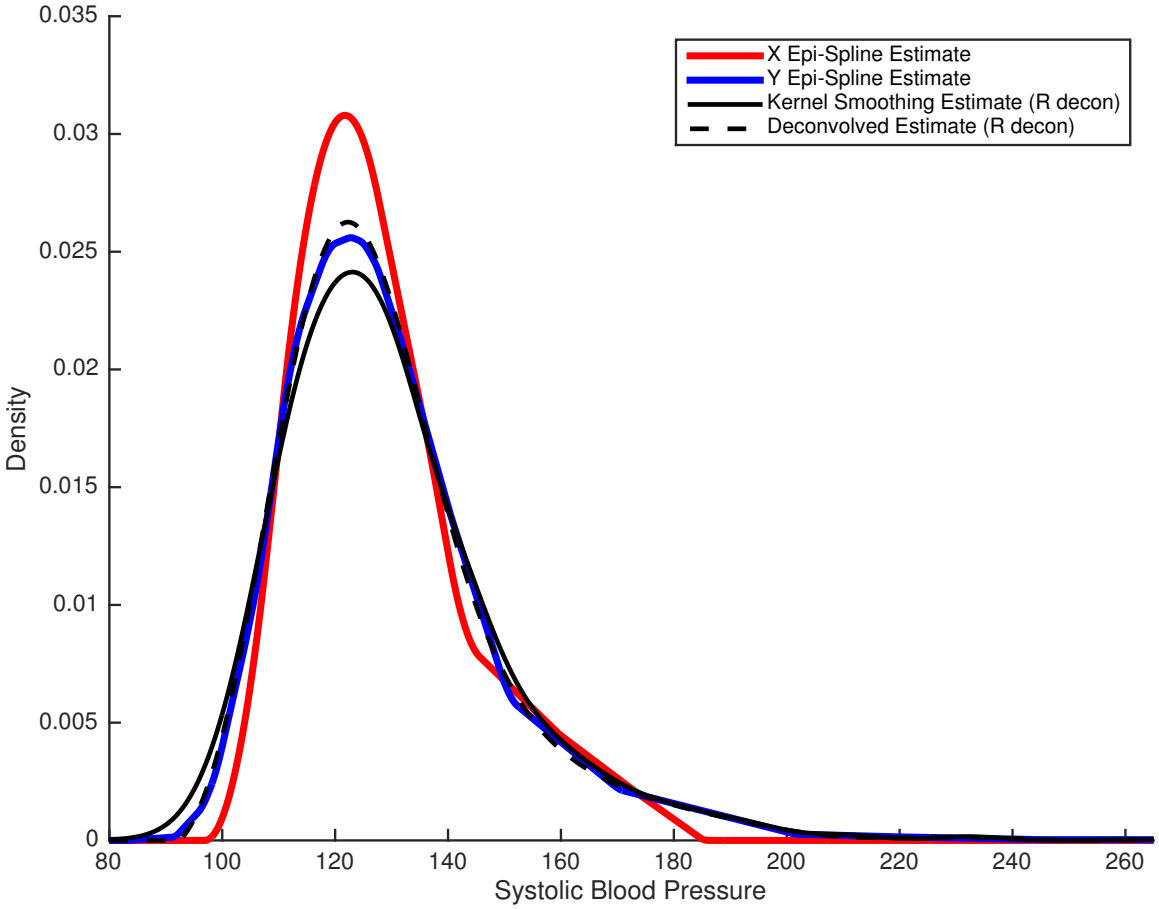
The blue line is the estimate of the density of SBP readings by from the solution of MLP-E. The red line is the deconvolved estimate given by the solution of MEP-E. We begin with continuity on the left and add the unimodal constraint on the right. The sharp peaks of both estimates show the necessity of imposing a maximum change in gradient.

4.3 Comparison of Deconvolution Methods

With all soft information, the deconvolved epi-spline density is shown in Figure 4.3. Here we also show the kernel density with a black solid line and the deconvolved estimate from **decon** with a black dashed line. We note that **decon** computes a certain kernel bandwidth adjustment [25], so that we focus on the differences in the effects of deconvolution by each method, rather than the differences between corresponding densities.

Both **decon** and MEP-E reduce the density in the tails and increase the density in the mode for exactly the same data. These changes in density are precisely those noted in the explanation of the R package **decon** in [25]. We see that deconvolution via epi-splines produces remarkably similar effects.

Figure 4.3: Deconvolution Method Comparison



A comparison of our methodology on Systolic Blood Pressure data against a FFT method in the R package **decon** is shown. We compare the difference between the black lines from the **decon** package and the difference between the epi-spline estimates in blue and red. Both deconvolved estimates reduce density in tails and increase density at the mode.

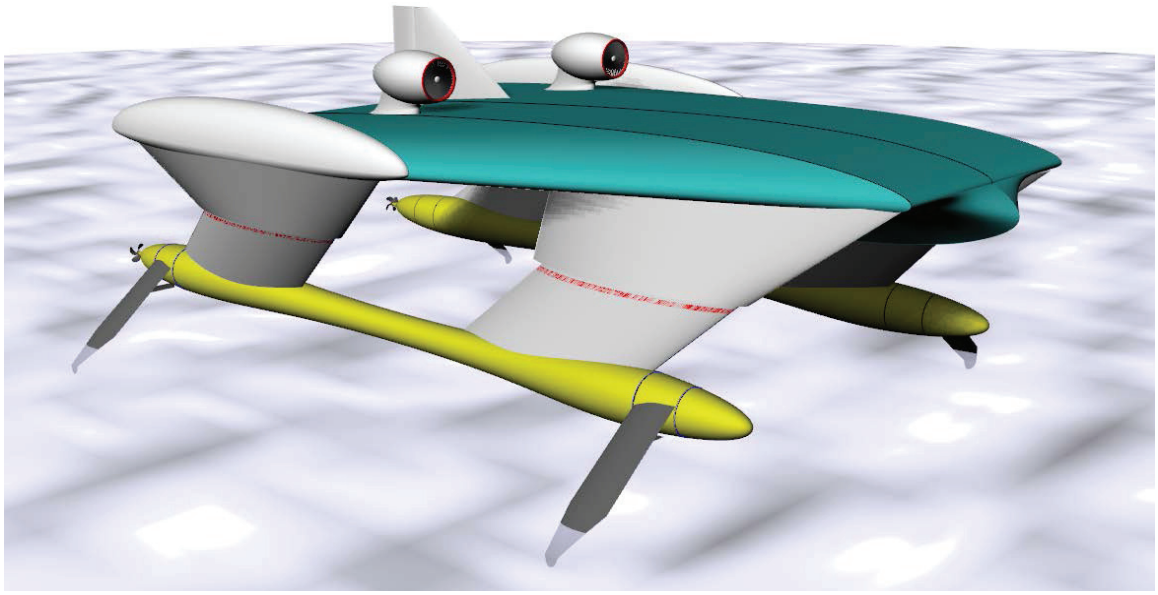
THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 5:
High-Fidelity and Low-Fidelity Simulation Output

5.1 Hydrofoil Concept

In this chapter we examine an example in UQ from fluid dynamics. We have received a data set from Dr. S. Brizzolara of the MIT SeaGrant program that contains the output of high-fidelity and low-fidelity fluid dynamics simulations for a hydrofoil concept displayed in Figure 5.1. The goal is to produce a method that can reduce the number of high-fidelity simulations required to produce an estimate of the density of high-fidelity performance by supplementing a few observations from a high-fidelity data set with many low-fidelity simulations.

Figure 5.1: Hydrofoil Design Concept

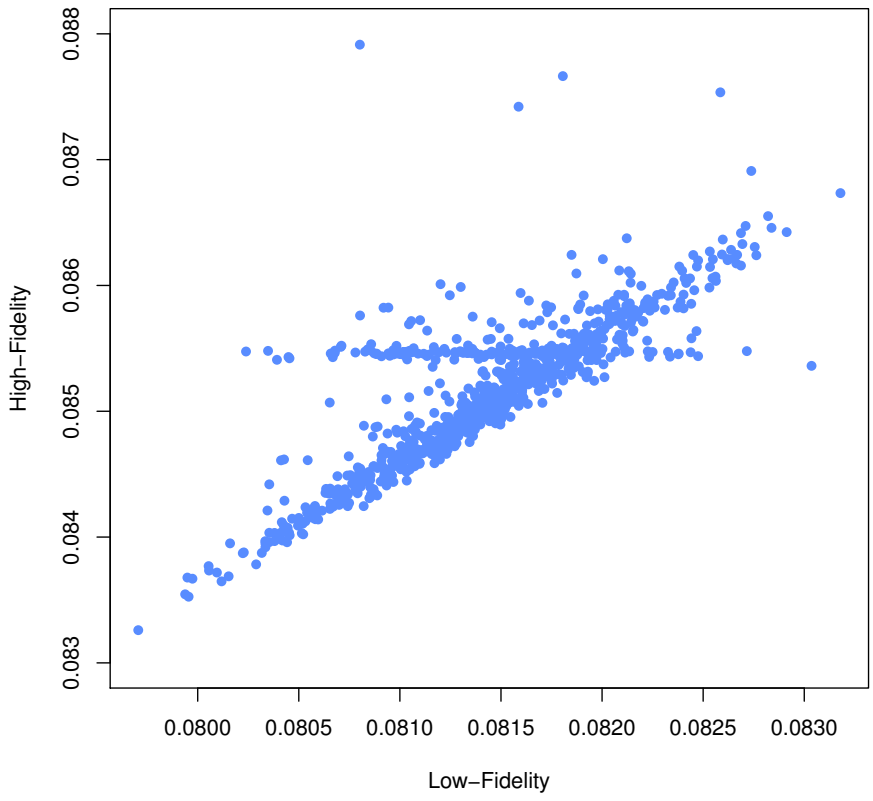


This image represents a design under development for an advanced high-speed hydrofoil. Fluid Dynamics simulations to compute characteristics such as Drag/Lift vary widely in fidelity and computational cost. The image and data provided for this experiment come courtesy of Dr. S. Brizzolara of the MIT SeaGrant program.

5.2 Computation Time

A discussion of the difference in computation time is helpful to understand the desire to reduce the number of high-fidelity observations. The simulation calculates the drag/lift coefficient (D/L) using two differing methods. The high-fidelity method solves using the Navier-Stokes equations requiring four hours of computation time on eight cores for a single solution. By contrast, the low-fidelity solve uses the panel method, which only takes five seconds on one core. The data contain 898 pairs of high-fidelity and low-fidelity observations of D/L. We compare the data by fidelity in Figure 5.2.

Figure 5.2: Comparison of High-Fidelity and Low-Fidelity Simulation Data



Each point represents a different hydrofoil shape, for which the drag/lift coefficient is computed via a high-fidelity simulation and a low-fidelity simulation. Horizontal and diagonal patterns appear in the scatterplot.

5.3 Comparison of Output Data

Figure 5.2 shows that there appear to be two underlying relationships between high-fidelity and low-fidelity data. For much of the data, high-fidelity and low-fidelity results are clearly positively correlated, but many other points show that high-fidelity D/L is largely unchanged as low-fidelity D/L is increased. We refer loosely to these two patterns as horizontal and diagonal portions.

We show kernel smoothing density estimates of high-fidelity data and low-fidelity data in Figure 5.3, computed in a manner described in Section 2.4. The low-fidelity kernel density is unimodal, but the high-fidelity kernel density is bimodal. Our goal is to identify the salient characteristics of the high-fidelity density by fusing together few high-fidelity samples and many low-fidelity samples in the following process, described in further detail in Algorithm 2 within the Appendix.

5.4 Regression on a Sample for Deconvolution

To relate this UQ scenario to Equation (1.2), we consider high-fidelity Drag/Lift as an observation of X and a low-fidelity Drag/Lift as an observation of Y. Using regression we may describe high-fidelity estimates in terms of low-fidelity data. We adopt the linear regression in the form

$$X = \beta_0 + \beta Y + \varepsilon \tag{5.1}$$

and since errors are symmetric, we can write

$$\beta_0 + \beta Y = X + \varepsilon. \tag{5.2}$$

We use an affine transformation with the coefficients of regression to obtain a new variable

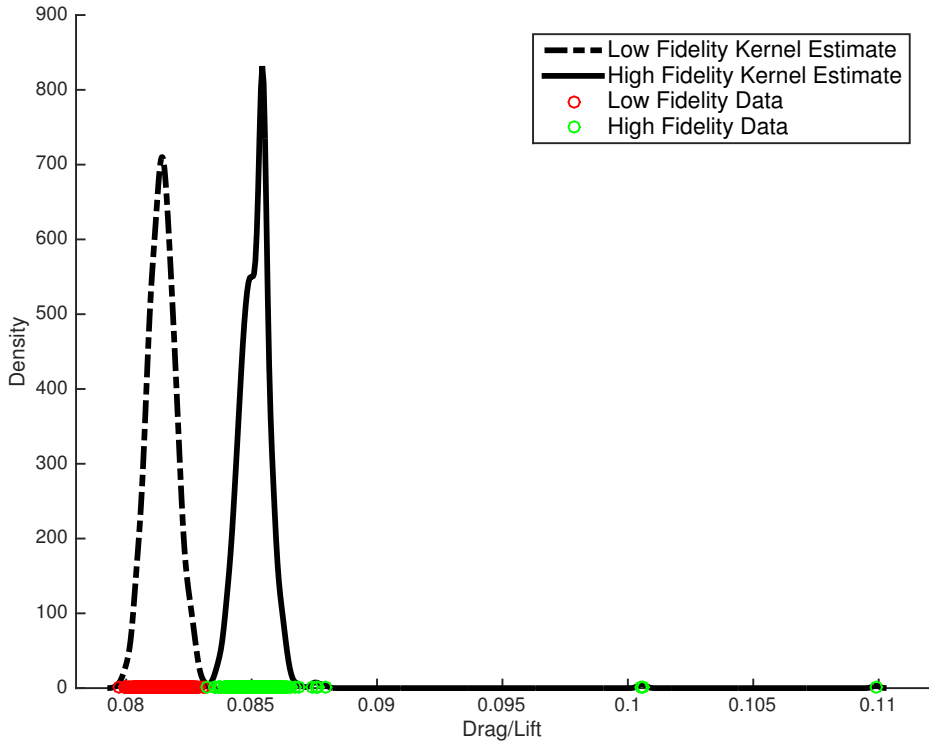
$$Y' = \beta_0 + \beta Y \tag{5.3}$$

so that

$$Y' = X + \varepsilon. \tag{5.4}$$

Thus we can use observations of the variable Y' to obtain a solution $\hat{f}_{Y'}$ of MLP-E to work towards a deconvolved solution.

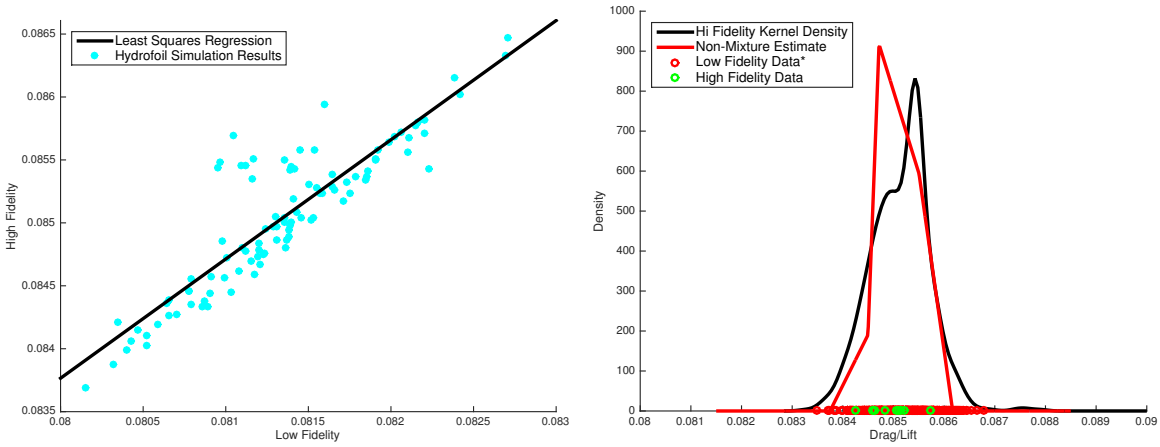
Figure 5.3: Estimates Obtained by Kernel Smoothing



Kernel smoothing estimates show the disparity between results of simulations of differing levels of fidelity. Red points indicate low-fidelity points, and the dashed-dotted line gives an estimated density given these low-fidelity points. Green points indicate high-fidelity points, and the solid line gives an estimated density given the high-fidelity points. Notice that the solid line is bimodal. High-fidelity outliers are also displayed.

We may perform a single regression, naively, on a sample of the data to generate an estimate of measurement error. That is, from the regression line shown in Figure 5.4, we obtain a residual standard error $\hat{\sigma}_{RSE}$ that we use as the estimate of the standard deviation of measurement errors. With the coefficients shown in Table 5.1 we transform 700 low-fidelity observations of Y into observations of Y', shown by the red points in Figure 5.4. Additionally, we obtain 10 samples of high-fidelity data, shown by the green points, and obtain a solution $\hat{f}_X$ of MLP-E, shown by the red line. We compare the result to a kernel smoothing estimate achieved using 898 samples of high-fidelity data. The estimate $\hat{f}_X$ fails to capture the bimodal behavior of the density. A more sophisticated approach is required to obtain a better estimate.

Figure 5.4: Regression on a Sample for Deconvolution



On the left, we show a single linear regression on a sample of 100 points. On the right, we show the kernel smoothing estimated density of 898 high-fidelity points with the black line. The red line is the deconvolved solution of MLP-E given 700 transformed low-fidelity points 10 high-fidelity points. Performance of the estimate is significantly altered without partitioning of the subsample for multiple univariate regressions. The deconvolution causes the mode of the epi-spline density to fall to the left of the peak kernel smoothing density.

Table 5.1: Single Regression Data

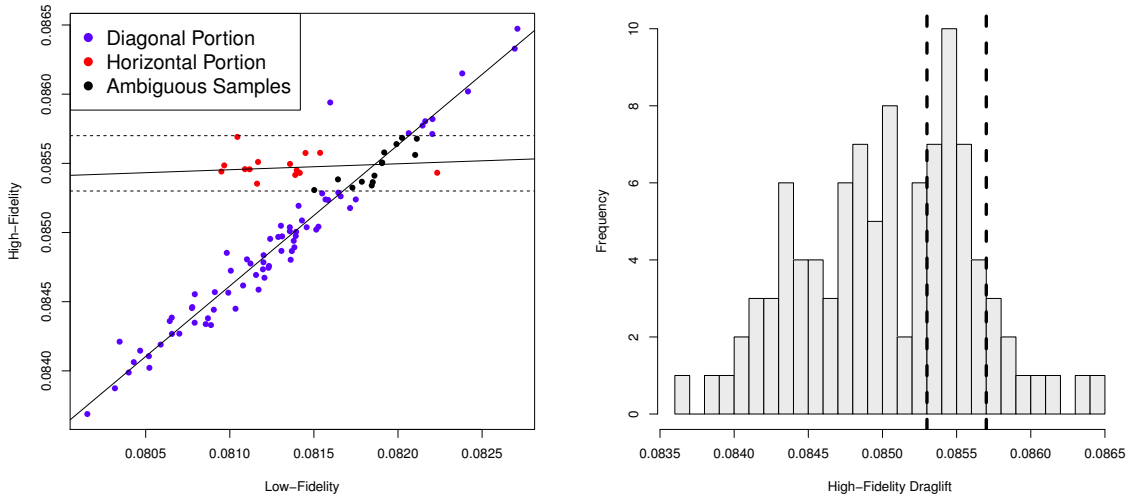
Entire Sample
Intercept, β_0: 7.982×10^{-3}
Slope, β: 9.473×10^{-1}
Residual Standard Error, $\hat{\sigma}_{RSE}$: 2.604×10^{-4}

5.5 Regressions on Partitions of a Sample for Deconvolution

Figure 5.5 shows a sample may preserve the horizontal and diagonal patterns discussed in Section 5.3. Though distinct patterns appear, the 14 points that fall in the intersection of these patterns present a certain ambiguity. We may assume that all points for which the high-fidelity D/L falls within a lower bound .0853 and upper bound .0857 belong to a horizontal portion of the data and the rest in the diagonal pattern. Alternatively, we may assume that the points that may be described as ambiguous (i.e., those that fall within the horizontal and diagonal patterns), should be distinct from those clearly in the horizontal pattern.

We record a weight p for the number of points in the horizontal portion divided by the sample size of 100. If ambiguous points are grouped with the other horizontal points, then 28 points fall in the horizontal pattern of 100 sampled points. Thus, $p = .28$, and $p = .14$ if the ambiguous points are grouped with the diagonal pattern, since 14 points fall in the ambiguous region. We weight density estimates of the horizontal and diagonal portions accordingly to produce a mixture.

Figure 5.5: Regressions on Partitions of a Sample for Deconvolution



We sample 100 points without replacement from the data set containing all high-fidelity and low-fidelity simulation output. The diagonal and horizontal patterns are clear, but there is an ambiguous region. The scatterplot on the left visualizes the partition into horizontal and diagonal sections, with a regression over red and black points, and a separate regression over the blue points. The bars within the dashed lines of the histogram represent those points assumed to be in the horizontal pattern.

We compute a linear regression for the horizontal pattern and the diagonal pattern for each method of partitioning the sample of 100. Two of the four regressions we compute appear as two intersecting lines in Figure 5.5 given the partition that groups all the ambiguous points together with the horizontal points as described in Figure 5.5. Coefficients β and β_0, and residual standard errors, $\hat{\sigma}_{RSE}$, for all regressions are given in Table 5.2 and identified by the groups of points included for the particular regression. Within Figure 5.5 and Table 5.2, the points that clearly fall along a diagonal pattern are described as diagonal

points, those clearly along the horizontal pattern are called horizontal points, and the points that do not clearly belong to a particular group are called ambiguous.

Table 5.2: Partition Regressions

Points Included	Intercept, β_0	Slope, β	$\hat{\sigma}_{RSE}$
Horizontal and Ambiguous	8.281×10^{-2}	3.280×10^{-2}	1.114×10^{-4}
Diagonal	3.588×10^{-2}	1.0006	1.819×10^{-4}
Horizontal	8.787×10^{-2}	-2.932×10^{-2}	8.704×10^{-5}
Diagonal and Ambiguous	2.982×10^{-3}	1.008	1.270×10^{-4}

For both partitions, the horizontal component density $\hat{f}_{Xh}$ we assume to be Normal($\mu = \bar{Y}'$,$\sigma = \hat{\sigma}_{RSE}$), where $\bar{Y}'$ and $\hat{\sigma}_{RSE}$ are found through the corresponding regression on the horizontal portion. We perform no deconvolution on the horizontal component, but we perform deconvolution on the diagonal component in the following manner.

We obtain 700 observations of Y' using the coefficients from the diagonal regression and a solution $f_{Y'}$ of MLP-E. From the remaining 98 high-fidelity and low-fidelity pairs of observations, we sample 10 additional high-fidelity observations without replacement from outside the region containing the horizontal pattern. The $\hat{\sigma}_{RSE}$ of the diagonal regression we use as an estimate of measurement error. For this diagonal portion, with the given data and soft information for this portion shown in Table 5.3 we obtain a solution $\hat{f}_{Xd}$ of MLP-E. The result is shown in Figure 5.6.

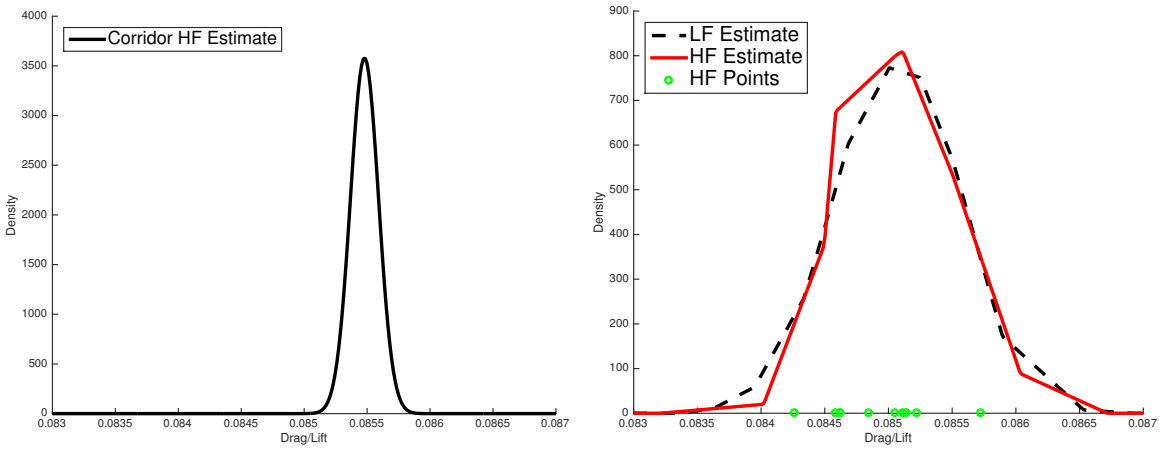
Table 5.3: Settings and Soft Information for Hydrofoil Data

Estimation Settings
Mesh Cardinality, $N = 1000$
Support, $[m_0, m_N]$: $[.0815, .0885]$
Convolution Tolerance, $\delta = 1 \times 10^{-2}$
Convolution Values Checked: 11
Continuous
Unimodal
Inflection Points, $[I_L, I_U]$: $[.0845, .0855]$
Convex Tails

Noticeable differences appear between the two component densities in Figure 5.6. The density at the mode for the horizontal portion is far greater than the peak density in the

diagonal component. The deconvolution of the diagonal portion, shown by the red line in Figure 5.6 reduces the density in the tails and increases the density in the mode, due to the observations of high-fidelity data points.

Figure 5.6: Mixture Density Components



The left plot shows the assumed normal density for the horizontal portion. On the right, we have deconvolved a density estimate of the diagonal portion using residual standard error from regression of the diagonal portion as an estimate of the standard deviation of the noise. The dashed line shows the result of MLP-E on 700 low-fidelity observations. The red line shows the deconvolved solution of MLP-E on a sample of 10 high-fidelity observations displayed as green circles.

5.6 Mixture Density Estimates

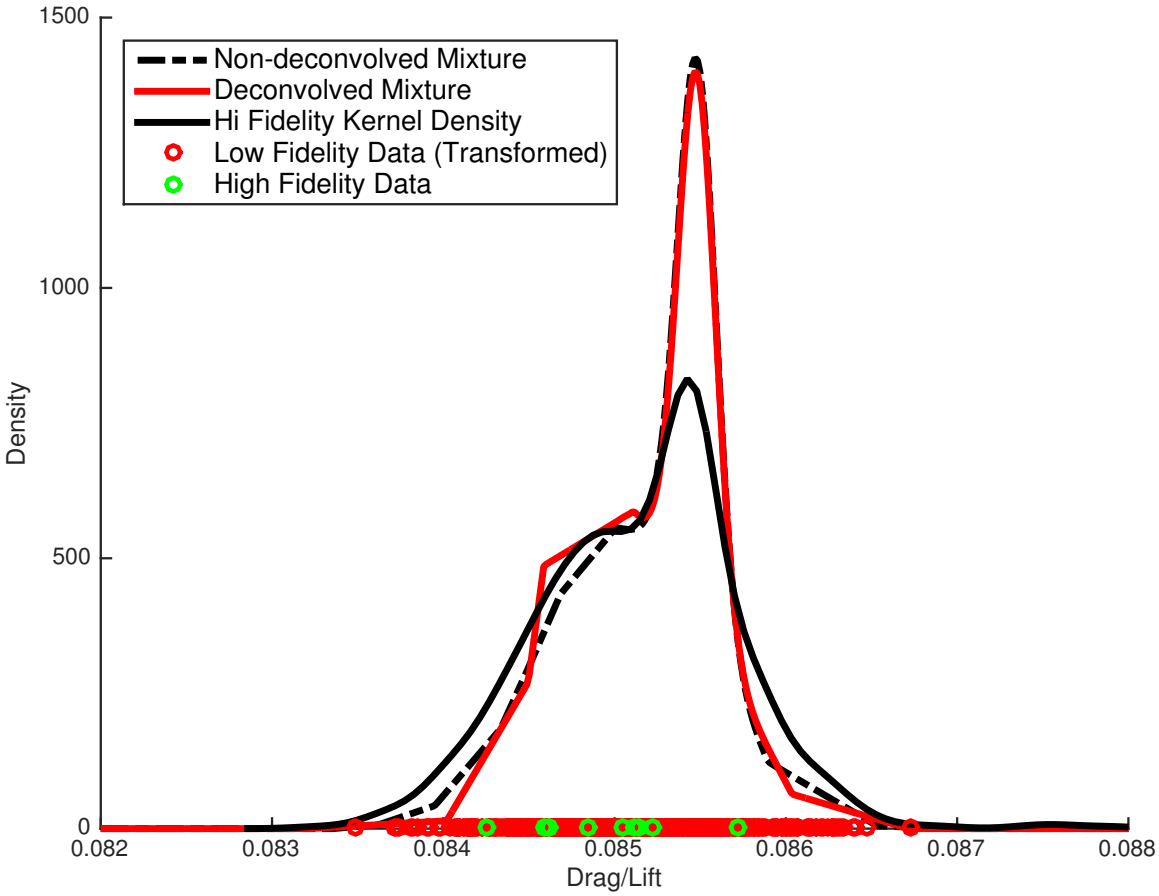
Recall that in Section 5.4 we recorded a weighing factor, p, with which to weight the component densities estimated in Figure 5.6. Now that we have identified the densities based on the horizontal and diagonal portions of the data the result is that

$$\hat{f}_X = p\hat{f}_{Xh} + (1-p)\hat{f}_{Xd}. \tag{5.5}$$

We plot the mixture density when $p = .28$ against the kernel smoothing density based exclusively on high-fidelity data. The result is shown in Figure 5.7. The result is a bimodal estimate whose modes closely approximate the modes shown in the kernel smoothing estimate.

In Figure 5.7, we see that the mixture density in red overestimates the higher mode significantly and we compare mixtures produced through different partitions in Figure 5.8. We see that the density at the higher mode is reduced at a cost of an increased density at the lower mode. The sharp peaks in the density may be reduced by imposing bounds in change of gradients as shown in Chapter 3 and Chapter 4.

Figure 5.7: Mixture and Kernel Estimates



The dashed line shows the mixture density produced without deconvolution in the diagonal portion. The red line shows the deconvolved mixture density. The black line shows the kernel smoothing estimate of the high fidelity data. The red samples show observations of low-fidelity drag/lift transformed by regression. We see that a deconvolved mixture density captures the bimodal behavior of the data, but overestimates the density in the higher mode.

The difference between the epi-spline mixtures and the kernel smoothing estimate must be considered in light of the difference in computation time for producing simulation

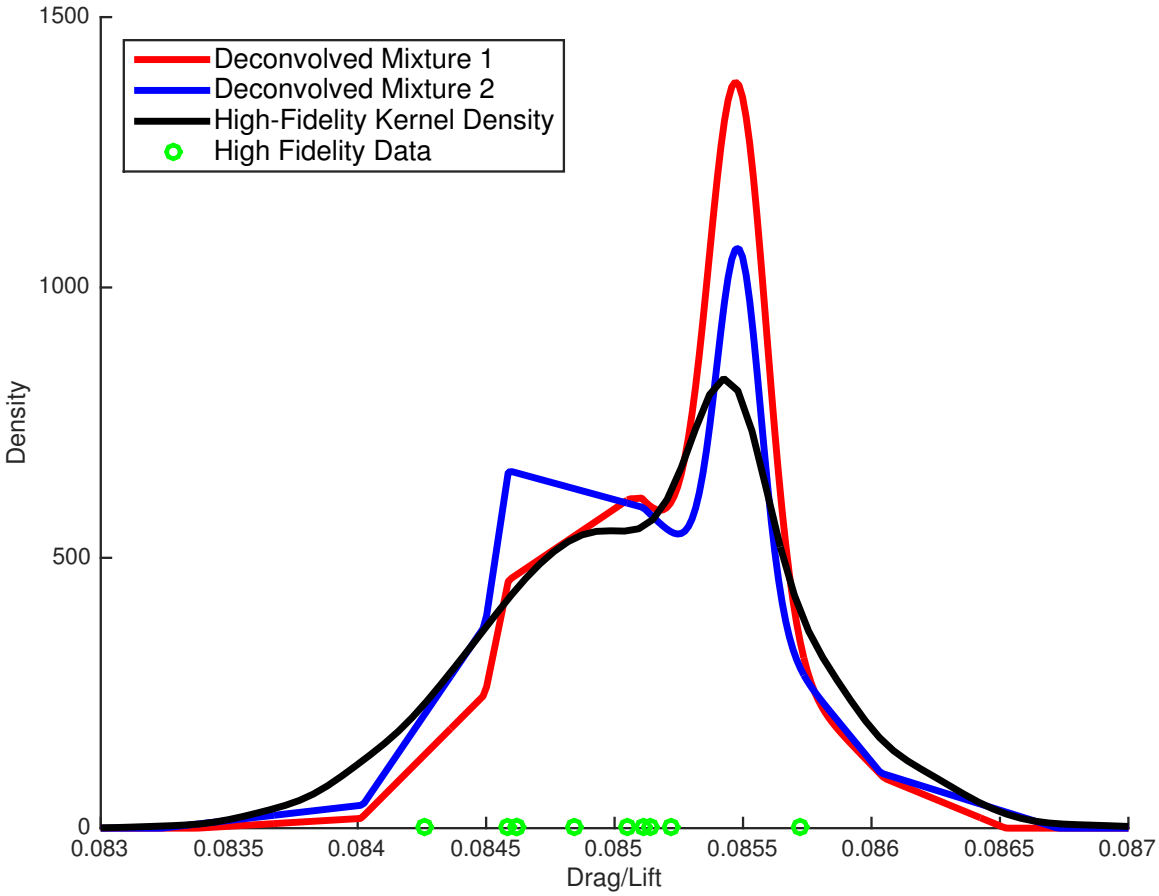
results. Table 5.4 shows how much computational expense would be required for each method of producing an estimate. Given the same number of cores, using an epi-spline mixture to acquire a density would allow for a reduction in computation time of over 87%.

Table 5.4: Computational Expense of Differing Methods

Method	X Count	Cores	Hours	Y Count	Cores	Hours	Core-Hours
Kernel	898	8	4	0	1	1.39×10^{-3}	28736
Mixture	110	8	4	898	1	1.39×10^{-3}	3521

An epi-spline mixture using both high-fidelity and low-fidelity data can identify the characteristics of a bimodal distribution of high-fidelity data. Additionally, using such a mixture technique can save an enormous amount of computational expense when the difference in computation time between high and low-fidelity simulations is vast.

Figure 5.8: Mixture Estimates Generated From Different Partitions



The red line shows the mixture estimate that results from assuming the ambiguous points of a sample fall in the horizontal pattern. The blue line shows the mixture estimate that results from assuming the ambiguous points of a sample fall in the diagonal pattern. The black line shows the kernel smoothing estimate. The red line overestimates the density of the higher mode and the blue line overestimates the density of the lower mode. Both capture the bimodal behavior of the density. The sharp peaks can be eliminated with bounds on changes in gradient.

THIS PAGE INTENTIONALLY LEFT BLANK

CHAPTER 6:
Conclusion

We have presented a new method of deconvolution using first-order epi-splines. To the best of the author's knowledge, no other available deconvolution technique blends multiple data sets of varying fidelity and incorporates soft information to produce a single density estimate. All information available should be used for uncertainty quantification, and especially in deconvolution, since highly accurate data is often extremely scarce.

Initial results are promising for deconvolution via epi-splines. We show highly accurate results in the Gamma example for a single trial with all soft information included. Replications show that this level of accuracy is not achieved by chance. The SBP scenario gives an example where epi-spline deconvolution can produce meaningful results without any high-fidelity data. Our example with the hydrofoil data shows the potential for epi-splines to be used in data collection resource decisions. Epi-spline deconvolution provides an alternative to analyzing the results of high and low fidelity observations in a compartmentalized fashion. Provided the user can obtain some knowledge of the accuracy of a low fidelity data set showing that measurement errors are Gaussian, epi-spline deconvolution can blend high and low fidelity data sets together for a viable estimate. Regression can serve as a tool for identifying an estimate of measurement error.

Challenges remain for the development of first-order epi-splines for density estimates. Widely available software makes use of automatic bandwidth selection in kernel smoothing density estimation, but we provide no such tool for selecting a hyperparameter Δ in the maximum change in gradient constraint described in Section 2.3.8. Similarly, the convolution constraint of Section 2.3.3 is well defined, but the tolerance and number of low fidelity density values to compare during deconvolution must be set manually. We recommend further analysis to obtain a reasonable selection for automatically identifying a well set δ before attempting to expand the scope of epi-spline deconvolution to multivariate densities or non-Gaussian errors. We remain confident that these challenges can be overcome, enabling epi-spline deconvolution for further use and widespread distribution.

THIS PAGE INTENTIONALLY LEFT BLANK

APPENDIX: Computation

A.1 Convolution Expression

We can find a closed form solution of the integral in the right hand side of

$$\hat{f}_y(y) = \sum_{k=1}^{N} \int_{m_{k-1}}^{m_k} \frac{1}{\sigma\sqrt{2\pi}} e^{-(y-x-\mu)^2 \big/ 2\sigma^2} (a_0^k + a^k x_k) dx \quad \forall y \tag{A.1}$$

by separating the expression into slope and intercept components and evaluating the integral over every interval. We expand this integral to

$$\int_{m_{k-1}}^{m_k} \frac{a^k x}{\sigma\sqrt{2\pi}} e^{-(y-x-\mu)^2 \big/ 2\sigma^2} dx + \int_{m_{k-1}}^{m_k} \frac{a_0^k}{\sigma\sqrt{2\pi}} e^{-(y-x-\mu)^2 \big/ 2\sigma^2} dx. \tag{A.2}$$

Through a combination of written methods and software such as the MATLAB Symbolic Toolbox [31] we find the closed form solutions of each addend in Equation (A.2). The slope component becomes

$$\begin{aligned} a^k \Bigg(& -\frac{\mu}{4\sigma} (-m_{k-1} - \mu + y) \operatorname{erf}\left(\sqrt{2}\right) + \frac{\mu}{2} \operatorname{erf}\left(\frac{-m_k - \mu + y}{\sigma\sqrt{2}}\right) \\ & - \frac{\sigma}{\sqrt{2\pi}} e^{-\frac{(-m_k - \mu + y)^2}{2\sigma^2}} + \frac{\sigma}{\sqrt{2\pi}} e^{-\frac{(-m_{k-1} - \mu + y)^2}{2\sigma^2}} \\ & + \frac{y}{2} \operatorname{erf}\left(\frac{-m_{k-1} - \mu + y}{\sigma\sqrt{2}}\right) - \frac{y}{2} \operatorname{erf}\left(\frac{-m_k - \mu + y}{\sigma\sqrt{2}}\right) \Bigg) \end{aligned} \tag{A.3}$$

and the intercept component, notably simpler, becomes

$$\frac{a_0^k}{2} \left(\operatorname{erf}\left(\frac{m_k + \mu - y}{\sigma\sqrt{2}}\right) - \operatorname{erf}\left(\frac{m_{k-1} + \mu - y}{\sigma\sqrt{2}}\right) \right). \tag{A.4}$$

We use the error function

$$\operatorname{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2} dt. \tag{A.5}$$

A.2 Artificial Data

First, we generate observations of random variables according to their respective densities using any capable software program. We store the sum of these two random variables in a .csv file which we use as input for a General Algebraic Modeling System (GAMS) program. In MLP-E, we produce an X data set. The GAMS program processes the data and produces output files which contain optimal solutions. Finally, these files are passed back to another program which plots the output and computes MSE.

Algorithm 1: Replication of Optimized Epi-Spline Estimates

Result: Mean Epi-Spline Estimate Error
foreach *Trial* **do**
  Generate 5000 sums of X and ε observations;
  Record each sum as Y data;
  if *Using Maximum Likelihood Estimate* **then** Record 3 new X samples;
  Input and process data into epi-spline estimating program;
  Input soft information for f_y and f_x;
  Solve for $\hat{f}_y$ using MLP-E;
  Record value of $\hat{f}_y(y)$ for 101 values of y;
  if *X data recorded* **then**
    Solve for $\hat{f}_x$ using MLP-E;
  else
    if *Maximum Entropy Desired* **then**
      Solve for $\hat{f}_x$ usingMEP-E;
    else
      Solve for $\hat{f}_x$ minimizing convolution tolerance, δ;
    end
  end
  Export $\hat{f}_y$ and $\hat{f}_x$ results to post-processing program;
  Compute MSE;
end
Compute mean MSE;
Display estimates and MSE as required;

A.3 Hydrofoil Concept

We generate a bimodal distribution in the following process. First we sample the data to identify patterns and establish a partition of the sample. Two univariate regressions provide information as inputs to produce distinct component densities. Abundant low fidelity data supplements scarce high-fidelity data for a deconvolved component estimate. The mixture density produced is comparable to a kernel smoothing estimate of exclusively high-fidelity data.

Algorithm 2: Developing an Epi-Spline Mixture Density

Result: Mixture Density Estimate
Collect 100 pairs of high-fidelity and low fidelity observations;
Establish interval $[a,b]$ in which horizontal portion falls;
$p =$ number of points for which $x \in [a,b]$;
foreach *Component c* **do**
  Calculate β_{0c}, β_c, and σ_c;
end
Set $\hat{f}_{Xh} \equiv \text{Normal}(\mu = \beta_{0h} + \beta_h \text{E}(Y), \sigma = \hat{\sigma}_h)$;
Collect 700 low fidelity observations;
Set $Y' \equiv \beta_{0d} + \beta_d Y$;
Generate $\hat{f}_Y d$ using epi-splines with Y' observations;
Collect 10 high-fidelity observations $X \notin [a,b]$;
Deconvolve $\hat{f}_{Yd}$ using high-fidelity observations and σ_d to generate $\hat{f}_{Xd}$;
$\hat{f}_X = p\hat{f}_{Xh} + (1-p)\hat{f}_{Xd}$;
return $\hat{f}_X$

A.4 Notes on Computation Time

We solved for all epi-spline estimates on a mid-2012 MacBook Air® laptop with a 1.8 Ghz Intel Core i5 processor and 4GB 1600 Mhz DDR3 RAM using GAMS and the CONOPT solver. All solves took less than 1-2 minutes. The numerical experiments of 30 replications in Section 3.3 took less than 8 minutes each.

THIS PAGE INTENTIONALLY LEFT BLANK

List of References

[1] S. Riad, "The deconvolution problem: An overview," *Proceedings of the IEEE*, vol. 74, no. 1, Jan 1986, pp. 82–85.

[2] E. A. Robinson, "Predictive decomposition of time series with application to seismic exploration," *Geophysics*, vol. 32, no. 3, 1967, pp. 418–484.

[3] E. A. Robinson, "Predictive decomposition of seismic traces," *Geophysics*, vol. 22, no. 4, 1957, pp. 767–778.

[4] J. Goutsias and J. M. Mendel, "Maximum-likelihood deconvolution: An optimization theory perspective," *Geophysics*, vol. 51, no. 6, 1986, pp. 1206–1220.

[5] J. M. Mendel and C. S. Burrus, *Maximum-likelihood Deconvolution: A Journey Into Model-based Signal Processing*. New York: Springer-Verlag, 1990.

[6] A. Ziolkowski, *Deconvolution*. Boston: International Human Resources Development Corporation, 1984.

[7] J. M. Mendel, *Optimal Seismic Deconvolution: An Estimation-based Approach*. New York: Academic Press, 1983.

[8] V. K. Arya and J. K. Aggarwal, *Deconvolution of Seismic Data*. Stroudsburg, PA: Hutchinson Ross Pub. Co., 1982.

[9] J. K. Kauppinen *et al.*, "Fourier self-deconvolution: A method for resolving intrinsically overlapped bands," *Applied Spectroscopy*, vol. 35, no. 3, 1981, pp. 271–276.

[10] W. E. Blass and G. W. Halsey, *Deconvolution of Absorption Spectra*. New York: Academic Press, 1981.

[11] F. N. Madden *et al.*, "A comparison of six deconvolution techniques," *Journal of Pharmacokinetics and Biopharmaceutics*, vol. 24, no. 3, 1996, pp. 283–299.

[12] D. Verotta, "Two constrained deconvolution methods using spline functions," *Journal of pharmacokinetics and biopharmaceutics*, vol. 21, no. 5, 1993, pp. 609–636.

[13] S. S. Haykin, *Blind Deconvolution*. Englewood Cliffs, N.J.: PTR Prentice Hall, 1994.

[14] S. Chaudhuri *et al.*, "Blind deconvolution methods: A review," in *Blind Image Deconvolution*. Springer, 2014, pp. 37–60.

[15] A. Levin *et al.*, "Understanding and evaluating blind deconvolution algorithms," in *IEEE Conf. Computer Vision and Pattern Recognition*, 2009, pp. 1964–1971.

[16] A. Meister, *Deconvolution Problems in Nonparametric Statistics*. Berlin Heidelberg: Springer-Verlag, 2009.

[17] M. H. Neumann and O. Hössjer, "On the effect of estimating the error density in nonparametric deconvolution," *Journal of Nonparametric Statistics*, vol. 7, no. 4, 1997, pp. 307–330.

[18] X.-F. Wang and D. Ye, "The effects of error magnitude and bandwidth selection for deconvolution with unknown error distribution," *Journal of Nonparametric Statistics*, vol. 24, no. 1, 2012, pp. 153–167.

[19] A. Sarkar *et al.*, "Bayesian semiparametric density deconvolution in the presence of conditionally heteroscedastic measurement errors," *Journal of Computational and Graphical Statistics*, vol. 23, no. 4, 2014, pp. 1101–1125.

[20] J. Fan, "On the optimal rates of convergence for nonparametric deconvolution problems," *The Annals of Statistics*, vol. 19, no. 3, 1991, pp. 1257–1272.

[21] A. Delaigle and P. Hall, "Parametrically assisted nonparametric estimation of a density in the deconvolution problem," *Journal of the American Statistical Association*, vol. 109, no. 506, 2014, pp. 717–729.

[22] R. J. Carroll and P. Hall, "Optimal rates of convergence for deconvolving a density," *Journal of the American Statistical Association*, vol. 83, no. 404, 1988, pp. 1184–1186.

[23] R. J. Carroll *et al.*, *Measurement Error in Nonlinear Models A Modern Perspective*, 2nd ed., ser. Chapman & Hall/CRC Monographs on Statistics & Applied Probability. Boca Raton, FL: CRC Press, 2006.

[24] R Core Team, *R: A Language and Environment for Statistical Computing*, R Foundation for Statistical Computing, Vienna, Austria, 2015. [Online]. Available: https://www.R-project.org/

[25] X.-F. Wang and B. Wang, "Deconvolution estimation in measurement error models: the R package decon," *Journal of statistical software*, vol. 39, no. 10, 2011.

[26] J. O. Royset and R. J-B. Wets, "From data to assessments and decisions: Epi-spline technology," INFORMS Tutorial, INFORMS, 2014.

[27] J. O. Royset *et al.*, "Uncertainty quantification using exponential epi-splines," Proceedings of the International Conference on Structural Safety and Reliability, New York, 2013.

[28] J. O. Royset and R. J-B. Wets, "Fusion of hard and soft information in nonparametric density estimation," *European J. of Operational Research*, 2015, to appear.

[29] J. O. Royset and R. J-B. Wets, "Multivariate epi-splines and evolving function identification problems," *Set-Valued and Variational Analysis*, 2015, to appear.

[30] D. I. Singham *et al.*, "Density estimation of simulation output using exponential epi-splines," Proceedings of the 2013 Winter Simulation Conference Institute of Electrical and Electronics Engineers, Inc., Piscataway, NJ, 755-765, 2013.

[31] MATLAB, *version 8.6.0.267246 (R2015b)*. Natick, Massachusetts: The MathWorks Inc., 2015.

[32] R. Yang *et al.*, "A quadratic programming approach to density deconvolution with additive measurement errors," Northwestern University, Evanston, IL, 2015, unpublished.

THIS PAGE INTENTIONALLY LEFT BLANK

Initial Distribution List

1. Defense Technical Information Center
   Ft. Belvoir, Virginia

2. Dudley Knox Library
   Naval Postgraduate School
   Monterey, California