# Semi-Parametric Inference with Density Ratio Model Fitted to Distributed Data using Alternating Direction Method of Multipliers

With Applications to Big Data Problems

**Alexander Imbrogno**

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

Master of Science
in
Mathematics with Concentration in Statistics

Carleton University
Ottawa Ontario

# Abstract

With the sheer volume and complexity of modern day data sets, there has become a need for new techniques and methodologies to handle problems related to "big data". One such problem of interest arrises when data is distributed and stored in various locations. For reasons of confidentially, complexity or volume, we are unable to have access to the entire dataset in one centralized location. Alongside the presence of distributed data, we are also interested in carrying out inference using all of the data from each local storage unit.

This thesis presents the application of the Alternating Direction Method of Multipliers (ADMM) algorithm to fit a semi-parametric Density Ratio Model (DRM) to a collection of distributed independent samples. The model parameters obtained using ADMM were found to be comparable to those fit from the same data present in a non-distributed setting.

This thesis also develops methodologies for carrying out inference using the semi-parametric DRM in the presence of distributed data. Techniques were developed for carrying out the dual empirical likelihood ratio test, which allows for the testing of composite hypothesis about DRM model parameters. This thesis also develops theories for the estimation of the baseline and marginal distribution functions for each sample alongside providing a method for estimating the quantiles of each distribution function.

Applying the ADMM algorithm in the presence of distributed data, we have successfully fit a DRM and carried out inference which arrive to the same statistical conclusions as if the model was fit and inference was carried out using the same data in a non-distributed setting.

# Acknowledgements

I am truly grateful for my supervisor, Dr. Song Cai, for being an exceptional mentor, motivator and friend throughout the course of my academic career. He has encouraged me to explore new ideas and chase my curiosities shaping me into a better researcher. Thank you Song for all the help and support you have given me throughout my academic career.

Any length of writing will not be able to quantify my appreciation for my mother. Growing up, she would always tell me "You can do anything you set your mind to". The attitudes and beliefs these words have cultivated for me have been an extraordinary tool for helping me through the trials and tribulations of my research and thesis. She has lifted me up when times were tough and I know I can always count on her when needed. Thank you for everything that you have done and continue to do.

I would like to thank my colleagues/ friends, Josh Miller and Marc Lapointe for always offering their thoughts, perspectives, and suggestions on my research and various works.

Last but certainly not least, I would like to thank my best friends Thomas and Randy. Thank you guys for always supporting me on my long academic journey. You've always believed in me and offered up support when needed.

# Contents

# List of Tables

# List of Figures

# List of Abbreviations

**DRM:** Density Ratio Model

**ADMM:** Alternating Direction Method of Multipliers

**CDF:** Cumulative Distribution Function

**EL:** Empirical likelihood

**DEL:** Dual Empirical likelihood

**MELE** Maximum Empirical likelihood Estimate

**PDF:** Probability Density Function

**DELRT:** Dual Empirical likelihood Ratio Test

# Chapter 1

# Introduction

## 1.1 Application background

With advancements in computers and computing capabilities in recent years, it is now possible to collect, store, process, and analyze enormous quantities of data on a dailey basis. Much of this data is available in real time and can be quite complex in nature. Numerous challenges arrive when handling such data sets, referred to as "big data" in industry, because of the data volume and complexity. It often arises, due to confidentiality, data volume and network structure, that data is distributed across multiple machines instead of being stored in one central location. In this scenario direct communication between local machines is limited in both bandwith and data volume which becomes problematic for model fitting and inference using all the data. To enable communication between the local machines, a central unit is required which can exchange information with each local machine thus connecting the network.

Currently, much of the literature assumes the data is independent and identically distributed (*i.i.d.*) across the local machines (Zhang et al., 2013), however, *i.i.d* data may not always be reasonable. For example, assume the monthly sales for a major chain restaurant are stored on various local machines, one in each province. It may not be reasonable to assume the distribution of sales are equal across all provinces. In this situation, and situations like it, a semi-parametric approach to modelling the data is desirable since it adds the flexibility needed to relax the assumption of *i.i.d.* data.

This thesis will illustrate the application of the alternating direction method of multipliers (ADMM) algorithm to fit a semi-parametric density ratio model (DRM) to a collection of independent (non-identically distributed) samples. The samples are assumed to be stored on separate local machines with the machines being connected through a central server/ unit. In addition we also assume that only a short summary of data can be transferred between the global and local machines. Throughout this thesis, we will refer to this scenario as the "distributed data setting". Here distributed refers to data being distributed and stored in seperate locations, not to be confused with the statistical definition. Theories on how to carry out an empirical likelihood based inference framework in the presence of distributed data under DRM model assumptions will also be studied.

## 1.2 Density ratio models: a brief introduction with examples

As a tool for introducing the density ratio model let us consider a real world application of the DRM. With lumber being an important tool for most construction jobs, monitoring and maintaining lumber strength is of great value for both producers and consumers. Specifically, we may be interested in monitoring the strength of lumber populations from various regions, species, etc, over time.

In the lumber application, assume that we have multiple independent samples of lumber strengths from distinct populations over various years. With the intent of monitoring and comparing lumber strengths over those years, we require the methods used to be statistically efficient in order to reduce the sample sizes needed for each sample. Data collection in this context is both time consuming and costly so the use of efficient statistical methods is highly desirable. On the notion of efficiency, the DRM allows us to leverage an inherent feature about the data, that is distinct populations of lumber over years, species, regions, and so on, share some common latent strength characteristics. Although the sample sizes may be small, efficient inference can be made by pooling the data together and carrying out inference based on the pooled sample. With the data sharing some common characteristics across the samples, it is then reasonable to assume the underlying lumber populations for

each sample connect in some way. In particular, the DRM assumes the populations of lumber strengths are connected via their density functions. Suppose we have $m+1$ independent samples each coming from a population with CDF $F_k(x),\ k = 0, 1, ..., m$. The DRM assumes that

$$dF_k(x) = \exp\{\alpha_k + \boldsymbol{\beta}_k^T \boldsymbol{q}(x)\} dF_0(x),\ \text{for } k = 1, ..., m,$$

where $\boldsymbol{q}(x)$ is a prescribed $d$-dimensional basis function, $(\alpha_k, \boldsymbol{\beta}_k^T)$ , $k = 1, ..., m$ are model parameters and $F_0(x)$ is the baseline distribution which remains unspecified.

The above DRM assumption serves as a tool for pooling data across multiple samples leading to efficient methods of inference. In the context of monitoring lumber strengths, the DRM allows us to compare the strength of lumber from various populations over time while reducing the sample sizes needed for a desired level of precision. The unspecified baseline function maintains the semi-parametric nature of the model allowing for added flexibility.

### 1.2.1 The exponential family of distributions

The DRM encompasses a large range of distribution families, however this thesis will focus on the exponential families of distributions, specifically highlighting the family of normal distributions.

It can be shown that every member of the family of exponential distributions satisfies the assumptions of the DRM. A distribution is said to be a member of the exponential family of distributions if it's density function can be written as

$$f(x, \boldsymbol{\theta}) = k(x)\ \exp\{\boldsymbol{\eta}^T(\boldsymbol{\theta})\boldsymbol{t}(x) - A(\boldsymbol{\theta})\},\ x \in S,$$

where $\boldsymbol{\theta}$ is a paramater vector, $k(.)$, $\boldsymbol{\eta}(.)$, $A(.)$ and $\boldsymbol{t}(.)$ are prescribed functions and the support of $x$, $S$, does not depend on $\boldsymbol{\theta}$. We can now show that all densities $f_k$ with paramater vector $\boldsymbol{\theta}_k$ coming from the same exponential family, satisfy the assumptions of the DRM. Consider the density function

$$f_k(x) = \exp\{[\boldsymbol{\eta}^T(\boldsymbol{\theta}_k) - \boldsymbol{\eta}^T(\boldsymbol{\theta}_0)]\boldsymbol{t}(x) + [A(\boldsymbol{\theta_0}) - A(\boldsymbol{\theta}_k)]\} f_0(x).$$

By the above definition $f_k$ is a member of the exponential family and also satisfies the assumptions of the DRM with

$$f_0(x) = k(x),\ \boldsymbol{q}(x) = \boldsymbol{t}(x),\ \alpha_k = A(\boldsymbol{\theta_0}) - A(\boldsymbol{\theta}_k),\ \boldsymbol{\beta}_k = \boldsymbol{\eta}^T(\boldsymbol{\theta}_k) - \boldsymbol{\eta}^T(\boldsymbol{\theta}_0).$$

In order for the above density to fully define an exponential family, the function $k(x)$ needs to be fully defined. The counterpart of $k(x)$ is $f_0(x)$ and under DRM assumptions is left un-specified as the non-parametric component of the model. This illustrates how the parametric exponential family of distributions is a special case of the semi-parametric DRM.

This thesis will focus primarily on distributed data coming from the family of normal distributions. For illustration purposes this thesis will only present how the family of normal distributions satisfies the DRM assumptions.

### 1.2.2 The DRM and the family of normal distributions

The family of normal distributions with parameters $\mu$ and $\sigma$ is a member of the exponential family with

$$\boldsymbol{\eta}(\mu, \sigma) = (\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2})^T,\ \boldsymbol{t}(x) = (x, x^2)^T,\ A(\mu, \sigma) = \frac{\mu^2}{2\sigma^2} + ln\sigma.$$

Therefore normal families of distributions with parameters $(\mu_k, \sigma_k)$ satisfy the DRM assumptions with basis function $\boldsymbol{q}(x) = (x, x^2)^T$ and model parameters

$$\alpha_k = ln\frac{\sigma_0}{\sigma_k} + \frac{\mu_0^2}{2\sigma_0^2} - \frac{\mu_k^2}{2\sigma_k^2},\ \boldsymbol{\beta}_k = (\frac{\mu_k}{\sigma_k^2} - \frac{\mu_0}{\sigma_0^2}, \frac{1}{2\sigma_0^2} - \frac{1}{2\sigma_k^2})^T.$$

### 1.2.3 The relationship between logistic regression models and the DRM

We observe a close relationship between the two-sample DRM and logistic regression models in the context of case-control studies (Qin and Zhang, 1997). The aim of a

case-control study is to identify key factors which contribute to an individual having a certain disease. This is done through the comparison of two groups; the first group (control group) who does not have the disease and second group (case group) who does have the disease. Let $\boldsymbol{Y} = \begin{bmatrix} y_1, & y_2, & ... & y_n \end{bmatrix}$ be a random vector of indicator variables where $y_i = 1$ indicates membership in the case group for the $i^{th}$ individual in the study and $y_i = 0$ indicates membership in the control group. The vector $\boldsymbol{X}_i$ is a vector of covariates for the the $i^{th}$ individual.

A classical model for the case-control data is the *logistic regression model* used by Prentice and Pyke (1979), Farewell (1979) and Mantel (1973). The model is given by

$$Pr[\boldsymbol{Y}_i = 1 | \boldsymbol{X}_i = x] = \frac{\exp(a + \boldsymbol{b}^t x)}{1 + \exp(a + \boldsymbol{b}^t x)}$$

with parameters $a$ and $\boldsymbol{b}$ to be estimated. Let us assume the covariates, $\boldsymbol{X}$, have some unspecified distribution given by $f(x)$. We also denote the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{Y} = 1$ as $f_1(x)$ and the conditional distribution of $\boldsymbol{X}$ given $\boldsymbol{Y} = 0$ as $f_0(x)$. Now applying Baye's rule we have

$$\begin{aligned} f_0(x) &= \frac{Pr[\boldsymbol{Y} = 0 | \boldsymbol{X} = x] f(x)}{Pr[\boldsymbol{Y} = \boldsymbol{0}]} = \frac{f(x)}{Pr[\boldsymbol{Y} = \boldsymbol{0}](1 + \exp(a + \boldsymbol{b}^t x))}, \\ f_1(x) &= \frac{Pr[\boldsymbol{Y} = 1 | \boldsymbol{X} = x] f(x)}{Pr[\boldsymbol{Y} = \boldsymbol{1}]} = \frac{\exp(a + \boldsymbol{b}^t x) f(x)}{Pr[\boldsymbol{Y} = \boldsymbol{1}](1 + \exp(a + \boldsymbol{b}^t x))}. \end{aligned}$$

Using the equations above, we can see that the conditional distributions of $\boldsymbol{X}$ given $\boldsymbol{Y}$ constitute a two sample DRM with basis function $\boldsymbol{q}(x) = x$, $\alpha = a + \log(Pr[\boldsymbol{Y} = 0]/Pr[\boldsymbol{Y} = 1])$, and $\boldsymbol{\beta} = \boldsymbol{b}$. Thus in the scenario of binary case-control data we have

$$f_1(x) = \exp\Big(\{a + \log\{Pr[\boldsymbol{Y} = 0]/Pr[\boldsymbol{Y} = 1]\}\} + \boldsymbol{b}^T x\Big) f_0(x).$$

In the case of categorical case-control data with $\boldsymbol{Y} = 0, 1, ..., m$, and covariates $\boldsymbol{X}$ satisfying the multinomial logit model assumptions given by

$$\frac{Pr[\boldsymbol{Y} = k] | \boldsymbol{X} = x}{Pr[\boldsymbol{Y} = 0] | \boldsymbol{X} = x} = \exp(a_k + \boldsymbol{b}^T x), k = 1, 2, ..., m,$$

we have that $f_k(x)$, the conditional distributions of $\boldsymbol{X}$ given $\boldsymbol{Y} = k$, fulfill the DRM assumptions with $\boldsymbol{q}(x) = x$, $\alpha_k = a_k + \log(Pr[\boldsymbol{Y} = 0]/Pr[\boldsymbol{Y} = 1] + \boldsymbol{b}^t x)$, and $\boldsymbol{\beta}_k = \boldsymbol{b}_k$:

$$f_k(x) = \exp(\{a_k + \log(Pr[\boldsymbol{Y} = 0]/Pr[\boldsymbol{Y} = k])\} + \boldsymbol{b}^T x) f_0(x).$$

### 1.2.4 Empirical likelihood inference under the DRM: recent and past developments

One of the earliest traces of the DRM in the literature is a paper by Anderson (1972) on logistic discrimination. Despite this earlier publication, the DRM did not gain popularity until a series of papers published by Qin et al. (Qin (1993), Qin and Zhang (1997), Qin (1998)) which discussed inference problems under two-sample DRMs implementing the empirical likelihood (EL). Due to the non-parametric nature of EL inference, it emerged as a natural inference framework for the semi-parametric DRM. The formal introduction of the EL framework to the DRM in a two-sample case was established by Qin (1998) in which the establishment of the asymptotic normality of the maximum EL estimator of the DRM parameters was made.

Shortly thereafter, the EL became the standard inference methodology under the DRM with papers regarding various aspects of inference using the EL under the DRM being published. Cheng and Chu (2004) and Fokianos (2004) explored density estimation under the two-sample and multi-sample DRM respectively. Quantile Estimation under the two-sample and multi-sample DRM were studied by Zhang (2000) and Chen and Liu (2013), respectively. To test linear hypotheses regarding the parameters of multi-sample DRMs, Fokianos et al. (2001) proposed a simple Wald-type test. Keziou and Leoni-Aubin (2008) examined the EL ratio test in order to evaluate the equality of two distributions satisfying a two-sample DRM.

Zou and Fine (2002) made the discovery that the DRM is not a regular model. When the DRM parameter $\boldsymbol{\beta} = 0$ it follows that $\boldsymbol{\alpha} = 0$. Thus when the true value of the DRM parameter is $(\boldsymbol{\beta}^*, \boldsymbol{\alpha}^*) = 0$, the EL function is not well defined in a neighbourhood surrounding the true parameter value. This violates an important regularity condition for any likelihood type inference. More on this will be discussed

in the next chapter, where we introduce the dual empirical likelihood (DEL) as a solution for this non-regularity.

## 1.3 A brief introduction to alternating direction method of multipliers (ADMM)

It so happens that various fields and domains are approaching problems through the use of data analysis, more specifically through the application of statistical models and machine learning algorithms. A common feature of these approaches is the optimization of some sort of function. In the statistical model setting, we commonly see optimization present in the maximization of likelikehood functions, or consequently the minimization of the negative likelihood. In terms of machine learning, we often observe the minimization of a convex loss function. With the size and dimensionality of datasets increasing, it is important to explore optimization techniques and algorithms which can handle both the complexity of modern datasets but which are also scable to handle the problem in a distributed or fully de-centralized manner.

The alternating direction method of multipliers, an optimization algorithm intended to solve linear constrained convex optimization problems, was first introduced in the mid 1970's by Glowinski and Marroco (1975) and Gabay and Mercier (1976). A cornerstone property of the ADMM is its capacity to solve problems in a distributed manner when the main objective to be minimized can be split into the sum of smaller functions. The separability of the objective function allows the ADMM to break the main optimization into smaller sub-problems to be solved. There have been various other important papers analyzing the properties and applications of the ADMM, some of which include Fortin and Glowinski (2000), Fukushima (1993), Eckstein and Fukushima (1994), and Tseng (1991), to name a few. More specifically, the convergence properties of the ADMM have been studied by Gabay (1983) and Eckstein and Bertsekas (1992).

The application of the ADMM has been explored and considered for a range of problems including statistical problems. ADMM for constrained sparse regression was discussed in Bioucas-Dias and Figueiredo (2010). Bioucas-Dias and Figueiredo (2010) analyzed ADMM for trace norm regularized least squares minimization and

sparse inverse covariance selection was studied by Yuan (2009), among others. The ADMM will be technically presented in Chapter 3.

## 1.4 Outline of thesis

The remainder of this thesis will be organized in the following manner. Chapter 2 will introduce the framework of the dual empirical likelihood (DEL) inference under the DRM. This chapter will highlight hypothesis testing, quantile estimation, and distribution function estimation using the DRM under appropriate conditions. Chapter 3 will discuss pre-cursor algorithms which will be used as "stepping stones" and motivation for understanding and developing the ADMM. This chapter will also present the ADMM algorithm itself along with some common variations which will be used in the subsequent chapters. Chapter 4 will implement a variation of the ADMM in order to fit the DRM to a collcetion of independent and "distributed" samples. Techniques for carrying out inference using the DRM under the distributed data setting will also be developed in this section. These will again include hypothesis testing, quantile estimation, and distribution function estimation. Chapter 5 will present the results of a simulation where the theories and methodoliges presented in chapter 4 will be carried out on a set of independent "distributed" samples generated from the family of normal distribtions. The same inference will be carried out on the same set of samples instead using data which is not distributed, i.e we have all the data stored in one dataset/location. The results from both approaches will be compared to evaluate the preformance of the presented distributed methods.

# Chapter 2

# Dual Empirical Likelihood Inference under the DRM

This chapter will present the framework of the dual empirical likelihood (DEL) proposed by Keziou and Leoni-Aubin (2008) and Cai et al. (2017). The DEL will be the methodology adopted to carry out inference under the DRM. We will first present some basics regarding empirical likelihood (EL) extended to the multi-sample setting under the DRM. We will then review the non regularity of the DRM under this EL framework, first spotted by Zou and Fine (2002). This will motivate and propel the use of DEL when preforming inference under the DRM, which will be discussed alongside the EL. We will present the DEL function, and discuss it's optimization for finding the likilihood based estimator of the DRM parameter $\boldsymbol{\theta}$. The subsequent sections will cover the dual empirical likelihood ratio test for testing hypothesis about DRM parameters introduced by Cai et al. (2017). Quantile estimation under multiple samples, as proposed by Chen and Liu (2013) and the estimation of sample distribution functions, including the baseline distribution, will also be presented in this chapter.

## 2.1 EL function and non-regularity under DRM

Suppose we have $m+1$ independent samples denoted:

$$\{x_{kj} : j = 1, ..., n_k\}_{k=0}^{m}$$

with $n_k > 0$ being the size of the $k^{th}$ sample and $F_k$ being the distribution function of the population from which the $k^{th}$ sample has been drawn. Let us assume the $F_k$ satisfy the DRM assumptions, that is:

$$dF_k(x) = \exp(\alpha_k + \boldsymbol{\beta}_k^T \boldsymbol{q}(x)) dF_0(x), k = 1, ..., m.$$

This assumption also implies

$$\int dF_k(x) = \int \exp\{\alpha_k + \boldsymbol{\beta}_k^T \boldsymbol{q}(x)\} dF_0(x) = 1. \tag{2.1}$$

We have:

$$\boldsymbol{\alpha} = (\alpha_1, ..., \alpha_m)^T,\ \boldsymbol{\beta} = (\boldsymbol{\beta}_1^T, ..., \boldsymbol{\beta}_m^T)^T,\ \boldsymbol{\theta}_k = (\alpha_k, \boldsymbol{\beta}_k^T),\ \text{and } \boldsymbol{\theta} = (\boldsymbol{\alpha}^T, \boldsymbol{\beta}^T),$$

and we also put $\alpha_0 = 0$ and $\boldsymbol{\beta}_0^T = \mathbf{0}$ by definition.

The development of the EL inference results in the establishment of the profile log EL function of the $F_k$ given by

$$\tilde{l}_n(\boldsymbol{\theta}) = -\sum_{k=0}^{m}\sum_{j=1}^{n_k} \log\Big\{1+\sum_{r=1}^{m} \lambda_r[\exp\{\alpha_r+\boldsymbol{\beta}_r^T \boldsymbol{q}(x_{kj})\}-1]\Big\}+\sum_{k=0}^{m}\sum_{j=1}^{n_k} \Big\{\alpha_k+\boldsymbol{\beta}_k^T \boldsymbol{q}(x_{kj})\Big\}. \tag{2.2}$$

With Lagrange multiplier $\lambda_r$, the maximum empirical likelihood estimator (MELE) for the DRM parameter $\boldsymbol{\theta}$ is found by maximizing $\tilde{l}_n(\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$. The notation used for the MELE will be $\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}}\ \tilde{l}_n(\boldsymbol{\theta})$.

In order for any likelihood type inference to be preformed, certain regularity conditions surrounding the EL function must hold. Specifically, we require the likelihood function to be well-defined and differentiable in a neighbourhood around the true value of $\boldsymbol{\theta}$. As Zou and Fine (2002) noticed, when the true value is 0, i.e. $\boldsymbol{\theta} = \mathbf{0}$, the EL function is not well- defined in a neighbourhood around $\boldsymbol{\theta}$. It can then be said that the DRM is not regular at this value of $\boldsymbol{\theta}$. We also observe a violation of these regularity conditions when $\boldsymbol{\theta}_k = \boldsymbol{\theta}_j$, $k \neq j$.

In the application of applying a DRM to the monitoring of lumber strengths, detecting a change in lumber strength from one year to another translates to a statistical hypothesis test regarding the DRM paramater $\boldsymbol{\theta}$. Let $F_k$ , $k = 1, ..., m$ be the population of lumber at year $k$. We note that $F_k = F_j$ corresponds to $\boldsymbol{\theta}_k = \boldsymbol{\theta}_j$, hence the lumber quality remains stable between year $k$ and year $j$. Similarily we find $F_k = F_0$ equivalent to $\boldsymbol{\theta}_k = \mathbf{0}$. In a statistical setting under the DRM, distributional differences are detected by carrying out the following hypothesis test:

$$H_0 : F_k = F_j \text{ for all } k, j \in \{0, ..., m\} \text{ vs } H_a : F_k \neq F_j \text{ for some } k \neq j,$$

or equivalently,

$$H_0 : \boldsymbol{\theta}_k = \boldsymbol{\theta}_j \text{ for all } k, j \in \{0, ..., m\} \text{ vs } H_a : \boldsymbol{\theta}_k \neq \boldsymbol{\theta}_j \text{ for some } k \neq j.$$

To carry out this test, under the EL framework, the EL ratio test would normally be applied. However, the non regularity of the DRM in this instance forbids the simple application of the EL ratio test to this important hypothesis of interest. As a solution, let us now consider the *dual empirical likelihood*.

## 2.2 The DEL function

As a way to mitigate the non-regularity of the DRM under the EL framework, and to enable likelihood type inference, Keziou and Leoni-Aubin (2008) proposed to use a "dual" form of the EL in the case of two samples. The theory has been extended by Cai et al. (2017) to the case of multiple samples and is referred to as the *dual empirical likelihood* . The dual counterpart to the profile log EL function in (2.2) is the DEL function, given by

$$l_n(\boldsymbol{\theta}) = -\sum_{k=0}^{m}\sum_{j=1}^{n_k}\log\Big\{\sum_{r=0}^{m}\hat{\lambda}_r\exp\{\alpha_r+\boldsymbol{\beta}_r^T\boldsymbol{q}(x_{kj})\}\Big\}+\sum_{k=0}^{m}\sum_{j=1}^{n_k}\Big\{\alpha_k+\boldsymbol{\beta}_k^T\boldsymbol{q}(x_{kj})\Big\}. \quad (2.3)$$

Where $\hat{\lambda}_r = \frac{n_r}{n}$, $n_r$ is the size of the $r^{th}$ sample and $n = \sum_{r=0}^{m} n_r$ is the total pooled sample size. This DEL function is constructed by replacing $\lambda_r$ in equation 2.2 with $\hat{\lambda}_r$.

It holds that the MELE, $\hat{\boldsymbol{\theta}}$, which maximized the profile log EL function also maximizes the DEL (2.3). So we have that

$$\hat{\boldsymbol{\theta}} = \text{argmax}_{\boldsymbol{\theta}}\ l_n(\boldsymbol{\theta}) = \text{argmax}_{\boldsymbol{\theta}}\ \tilde{l}_n(\boldsymbol{\theta})$$

and both functions have the same maximal values $l_n(\hat{\boldsymbol{\theta}}) = \tilde{l}_n(\hat{\boldsymbol{\theta}})$.

The DEL has the following appealing properties: (i) it is well-defined for all values of $\boldsymbol{\theta}$ in the corresponding parameter space which allows the DEL to be used without worry to conduct likelihood type inference, especially when $\boldsymbol{\theta}_k = \boldsymbol{\theta}_j$ and $\boldsymbol{\theta} = \mathbf{0}$; (ii) it has a much simpler analytical form since the $\lambda_r$ are replace with their data value independent $\lambda_r$ ; (iii) the DEL is a smooth concave function which allows for the computation of the MELE to be preformed with added ease.

With the motivation to preform inference using the DRM, the DEL will be the framework adopted in the remainder of this thesis. Methods to apply DEL inference under the assumption of distributed data will be discussed further in Chapter 4.

## 2.3 Estimation of the baseline distribution: computing $\hat{F}_0(x)$ and $\hat{p}_{kj}$

Once the MELE of the DRM parameter is computed, by maximizing the DEL function (2.3), estimation of the distribution and density functions for the baseline distribution ($f_0(x)$ and $F_k(x)$) are ready to be carried out. We will use the notation $p_{kj} = dF_0(x_{kj})$, to denote the density function of the baseline distribution evaluated at the $j^{th}$ data point from sample $k$. The EL inference for multiple samples under the DRM gives us the equation for the MELE of $p_{kj}$, a function of $\hat{\boldsymbol{\theta}}$,

$$\hat{p}_{kj} = n^{-1}\Big\{1 + \sum_{r=1}^{m} \hat{\lambda}_r[\exp\{\hat{\alpha}_r + \hat{\boldsymbol{\beta}}_r^T \boldsymbol{q}(x_{kj})\} - 1]\Big\}^{-1}. \tag{2.4}$$

After computing the $\hat{p}_{kj}$ for all $k, j$ combinations, the MELE for the baseline CDF, $\hat{F}_0(x)$ is given by

$$\hat{F}_0(x) = \sum_{r=0}^{m} \sum_{j=1}^{n_r} \hat{p}_{rj} \mathbb{1}(x_{rj} \leq x), \tag{2.5}$$

where $\mathbb{1}(.)$ is the indicator function which takes a value of 1 if $x_{rj} \leq x$ and a value of 0 otherwise.

## 2.4 Estimation of non-baseline distribution functions, computing : $\hat{\boldsymbol{F}}_1, ..., \hat{\boldsymbol{F}}_m$

In a similar fashion to estimating the baseline CDF, we will employ the MELEs $\hat{p}_{kj}$ to the estimation of the non baseline CDFs, $F_1, ..., F_k$.

The MELE for $F_k(x)$ is given by

$$\hat{F}_k(x) = \sum_{r=0}^{m} \sum_{j=1}^{n_r} \exp\{\hat{\alpha}_k + \hat{\boldsymbol{\beta}}_k^T \boldsymbol{q}(x_{rj})\} \hat{p}_{rj} \mathbb{1}(x_{rj} \leq x). \tag{2.6}$$

It is important to note that $\hat{F}_k(x)$ is a function of $\boldsymbol{\theta}_k$, $x$, and $\hat{p}_{rj}$, $r = 0, .., m$, $j = 1, ..., n_r$. This fact will be exploited when introducing methods to compute $\hat{F}_k(x)$ given distributed samples.

## 2.5 Quantile estimation

In the context of our lumber application example, knowledge of the quantiles of lumber strengths is of high value to ensure lumber products meet a desired strength standard. That is, determing if a certain quantile of the population of lumber in a given year meets a prescribed strength standard. For instance we may require that 90 percent of the distribution of lumber attains a given industry set standard for strength testing. This leads us to computing quantile estimates for each of the $k = 0, .., m$ lumber distributions.

Prior to the computation of the quantile estimates for the $k^{th}$ lumber distribution, the $\hat{F}_k(x_{rj})$ need to be computed using all the $x_{rj}$ $\forall$ $r = 0, .., m$ and $j = 1, .., n_r$.

For $\delta \in [0, 1]$ the $\delta^{th}$ quantile of $F_k(x)$ is defined as

$$\omega_{k,\delta} = \min\Big\{x : F_k(x) \geq \delta\Big\}.$$

The EL-based estimator of $\omega_{k,\delta}$ is then given by

$$\hat{\omega}_{k,\delta} = \min\Big\{x_{obs} : \hat{F}_k(x_{obs}) \geq \delta\Big\}, \tag{2.7}$$

where $x_{obs} \in \{x_{rj} : r = 0, .., m, j = 1, .., n_r\}$ are coming from the observed data, hence the need to compute $\hat{F}_k(x_{rj})\ \forall\ r, j$.

## 2.6 Dual empirical likelihood ratio test: hypothesis testing regarding DRM parameters

As discussed in (2.1), detecting distributional differences is an important task in monitoring the strength of lumber from year to year. Recall, we detect distributional differences under the DRM by testing a hypothesis about the DRM parameters.

$$H_0 : \boldsymbol{\theta}_k = \boldsymbol{\theta}_j \text{ for all } k, j \in \{0, ..., m\} \text{ vs } H_a : \boldsymbol{\theta}_k \neq \boldsymbol{\theta}_j \text{ for some } k \neq j.$$

Let $\boldsymbol{\theta}_k = (\alpha_k, \boldsymbol{\beta}_k^T)^T$ be the DRM parameter corresponding to the lumber distribution at year $k$, $f_k$. Under DRM assumptions, $f_k$ is a valid probability distribution function (PDF) hence we have

$$\int dF_k(x) = \int \exp\{\alpha_k + \boldsymbol{\beta}_k^T \boldsymbol{q}(x)\} dF_0(x) = 1.$$

Solving for $\alpha_k$ we get

$$\alpha_k = -\log \int \exp\{\boldsymbol{\beta}_k^T \boldsymbol{q}(x)\} dF_0(x).$$

Hence the value of $\alpha_k$ is a function of $\boldsymbol{\beta}_k$, meaning we can simplify the above hypothesis to a hypothesis testing problem involving only $\boldsymbol{\beta}$. An equivalent hypothesis is

$$H_0 : \boldsymbol{\beta}_k = \boldsymbol{\beta}_j \text{ for all } k, j \in \{0, ..., m\} \text{ vs } H_a : \boldsymbol{\beta}_k \neq \boldsymbol{\beta}_j \text{ for some } k \neq j. \tag{2.8}$$

The above hypothesis can be abstracted to a more general linear hypothesis of the form

$$H_0 : \boldsymbol{g}(\boldsymbol{\beta}) = \mathbf{0} \text{ vs } H_a : \boldsymbol{g}(\boldsymbol{\beta}) \neq \mathbf{0}, \tag{2.9}$$

where $\boldsymbol{g} : \mathbb{R}^{md} \to \mathbb{R}^q$ with $q \leq md$ where $md$ is the length of $\boldsymbol{\beta}$ and $\boldsymbol{g}$ is assumed to be thrice differentiable with a full rank Jacobian matrix. To illustrate, assume we are interested in testing if the first and second lumber populations are the same. If both populations are normally distributed then the dimension of the basis function, $\boldsymbol{q}(x)$, is $d = 2$. The hypothesis written in the form given by (2.8) would be

$$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 \text{ vs } H_a : \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2. \tag{2.10}$$

Using the form given in (2.9) we have

$$\boldsymbol{g}(\boldsymbol{\beta}) = A\boldsymbol{\beta} = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2.$$

To carry out similar tests, including detecting distributional differences, Cai et al. (2017) developed the *dual empirical likelihood ratio test* (DELRT). The DELR test statistic is given by

$$R_n = 2\{l_n(\hat{\boldsymbol{\theta}}) - l_n(\tilde{\boldsymbol{\theta}})\}, \tag{2.11}$$

where $l_n(\hat{\boldsymbol{\theta}})$ is the DEL function evaluated at the MELE and $\tilde{\boldsymbol{\theta}}$ is the value of $\boldsymbol{\theta}$ which maximizes the DEL under the constraint given in the null hypothesis. Referring to the hypothesis stated in (2.10), the constraint on $\boldsymbol{\theta}$ is $\boldsymbol{\theta_1} = \boldsymbol{\theta_2}$.

The null limiting distribution of the DELR test statistic is in fact the same as a classical likelihood ratio test statistic. Under the null hypothesis, $\boldsymbol{g}(\boldsymbol{\beta}) = \mathbf{0}$, $R_n \to \chi^2_q$ as $n \to \infty$, with $q$ being the dimension of the image of $\boldsymbol{g}$. Exploiting the null limiting distribution of $R_n$, we attain a rejection region for the DELR test of

$(\chi^2_{q,1-\alpha}, \infty)$, corresponding to a test at $\alpha$ level of significance. Here, $\chi^2_{q,1-\alpha}$ is the $1-\alpha^{th}$ quantile of the $\chi^2_q$ distribution.

# Chapter 3

# Precursor Algorithms, ADMM, and Useful ADMM Variations

This chapter will serve as a technical extension and continuation from the introduction given in Chapter 1. In this chapter we review the literature surrounding some pre-cursor algorithms which contribute to the development of the ADMM. The precursor algorithms discussed will be the dual ascent, dual decomposition, and the augmented Lagrangian method of multipliers. Next we will present the ADMM algorithm along with some convergence properties discussed in the literature. Alongside the ADMM, we will also study a few useful variations of the algorithm. These variations will be important when solving for the MELE, $\hat{\boldsymbol{\theta}}$, and the MELE under a null constraint, $\tilde{\boldsymbol{\theta}}$. The useful variations to be covered are the consensus optimization and consensus optimization with regularization. We will also mention the implementation of a varying penalty parameter term in the algorithm which has been shown to speed up convergence (He et al., 2000).

## 3.1 Precursor algorithms

### 3.1.1 Dual ascent

The dual ascent algorithm aims to solve the convex linear-equality constrained minimization problem given by

$$\begin{aligned} &\text{minimize } f(\boldsymbol{x}) \\ &\text{subject to } A\boldsymbol{x} = \boldsymbol{b}, \end{aligned} \tag{3.1}$$

with variable $\boldsymbol{x} \in \mathbb{R}^n$, matrix $A \in \mathbb{R}^{m \times n}$, vector $\boldsymbol{b} \in \mathbb{R}^m$, and convex function $f : \mathbb{R}^n \to \mathbb{R}$. The Lagrangian for the problem given in (3.1) is

$$L(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}) + \boldsymbol{y}^T(A\boldsymbol{x} - \boldsymbol{b}),$$

where $\boldsymbol{y} \in \mathbb{R}^n$ is called the Lagrange multiplier or dual variable. The dual function for the same problem is

$$g(\boldsymbol{y}) = \inf_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{y}).$$

The dual problem for (3.1) is then

$$\text{maximize } g(\boldsymbol{y}).$$

Let $\boldsymbol{y}^*$ be an optimal point for the dual problem. When strong duality holds, the optimal value of the primal problem (3.1) is the same as that for the dual problem. This allows us to retrieve $\boldsymbol{x}^*$, the optimizer of $f$ subject to the constraints, by

$$\boldsymbol{x}^* = argmin_{\boldsymbol{x}} L(\boldsymbol{x}, \boldsymbol{y}^*),$$

provided there is only one minimizer of $L(\boldsymbol{x}, \boldsymbol{y^*})$, thus strong convexity and finiteness of $f$ is needed. We solve the dual problem using gradient ascent. For a fixed value $\boldsymbol{y}_0$ we obtain $\boldsymbol{x}^+ = \text{argmin}_x L(\boldsymbol{x}, \boldsymbol{y}_0)$. Assuming $g$ is differentiable, the gradient of $g$ is, $\nabla g(\boldsymbol{y}_0) = A\boldsymbol{x}^+ - b$, which is the residual of the equality constraint. Using this we obtain the dual ascent by iterating between the following steps

**Algorithm**

$$\boldsymbol{x}^{i+1} = \text{argmin}_x L(\boldsymbol{x}, \boldsymbol{y}^i), \tag{3.2}$$

$$\boldsymbol{y}^{i+1} = \boldsymbol{y}^i + \gamma^i(A\boldsymbol{x}^{i+1} - \boldsymbol{b}), \tag{3.3}$$

The superscript refers to the iteration index and $\gamma^i > 0$ is the stepsize at iteration $i$. The algorithm borrows it's name from the fact that with appropriate choice of $\alpha^i$, the optimal value of $g$ at iteration $i$ is increasing, i.e. , $g(\boldsymbol{y}^{i+1}) > g(\boldsymbol{y}^i)$. The dual ascent converges almost surely to a global minimum when the stepsize in (3.3) decrease at an appropriate rate subject to the relatively mild assumptions Bottou (1998) Kiwiel (2001).

### 3.1.2 Dual decomposition

The idea of dual decomposition, which can be traced back to Everett III (1963), stems from a nice property of the dual ascent. When the objective function $f$ and it's variable $\boldsymbol{x}$ are separable, the dual ascent can lead to a fully decentralized algorithm. In this regard, a function is said to be separable if it can be decomposed into a sum of smaller objective functions which are functions of subsets of the main variable $\boldsymbol{x}$. If $f$ is separable, then

$$f(\boldsymbol{x}) = \sum_{k=1}^{N} f_k(\boldsymbol{x}_k),$$

with variable $\boldsymbol{x} = (\boldsymbol{x}_1, \boldsymbol{x}_2, ..., \boldsymbol{x}_N)$ and the sub-variables $\boldsymbol{x}_i \in \mathbb{R}^{n_i}$ are sub-vectors of $\boldsymbol{x}$. We also partition the matrix $A$ as $A = [A_1 A_2 ... A_N]$ such that $A\boldsymbol{x} = \sum_{i=1}^{N} A_i \boldsymbol{x}_i$. Assuming $f$ is separable in $\boldsymbol{x}$ and we wish to solve the same constrained minimization problem given in (3.1), the Lagrangian for the problem can be written as

$$L(\boldsymbol{x}, \boldsymbol{y}) = \sum_{k=1}^{N} L_k(\boldsymbol{x}_k, \boldsymbol{y}) = \sum_{k=1}^{N} \{f_k(\boldsymbol{x}_k) + \boldsymbol{y}^T A_k \boldsymbol{x}_k - \frac{1}{N} \boldsymbol{y}^T \boldsymbol{b}\}.$$

Due to the separability of the objective and Lagrangian, the update step for each $\boldsymbol{x_k}$ can be carried out in parallel leading to a de-centralized algorithm. The dual decomposition algorithm consists of iterating between the following update steps:

**Algorithm**

$$\boldsymbol{x}_k^{i+1} = \text{argmin}_{x_k} L_k(\boldsymbol{x}_k, \boldsymbol{y}^i) \text{ for } k = 1, .., N. \tag{3.4}$$

$$\boldsymbol{y}^{i+1} = \boldsymbol{y}^i + \gamma^i(A\boldsymbol{x}^{i+1} - \boldsymbol{b}). \tag{3.5}$$

We can think about each iteration as containing 2 main actions. The first is a collection step where the locally computed values of $\boldsymbol{x}_k^i$ are gathered in one central location to compute the residual, $A\boldsymbol{x}^{i+1} - \boldsymbol{b}$, and update the dual variable $\boldsymbol{y}$. The second step is a broadcast step where the globally computed $\boldsymbol{y}^{i+1}$ is broadcasted to the $N$ worker nodes which will then locally update their $\boldsymbol{x_i}$, by solving the $k$ minimization problems given in (3.4).

### 3.1.3 Augmented Lagrangian and the method of multipliers

The augmented Lagrangian method of multipliers (ALMM) arose with the aim to bring robustness to the dual ascent algorithm, specifically to allow convergence when assumptions like strict convexity or finiteness of the objective function are not met. With the addition of an augmented penalty term, the augmented Lagrangian for the problem given in (3.1) is

$$L_p(\boldsymbol{x}, \boldsymbol{y}) = f(\boldsymbol{x}) + \boldsymbol{y}^T(A\boldsymbol{x} - \boldsymbol{b}) + (p/2)\left\|A\boldsymbol{x} - \boldsymbol{b}\right\|_2^2, \tag{3.6}$$

with penalty parameter $p > 0$. The associated dual function is

$$g_p(\boldsymbol{y}) = \text{inf}_{\boldsymbol{x}} L_p(\boldsymbol{x}, \boldsymbol{y}). \tag{3.7}$$

The major benefit of including the penalty term $p$, is that $g_p$ can be shown to be differentiable under mild conditions on the original problem, thus adding some robustness to the algorithm. Applying the dual ascent iterations we get the following algorithm:

**Algorithm**

$$\boldsymbol{x}^{i+1} = \text{argmin}_x L_p(\boldsymbol{x}, \boldsymbol{y}^i), \tag{3.8}$$

$$\boldsymbol{y}^{i+1} = \boldsymbol{y}^i + p(A\boldsymbol{x}^{i+1} - \boldsymbol{b}), \tag{3.9}$$

Note that the gradient of $g_p$ is found the same way as in the above algorithms. We first minimize over $\boldsymbol{x}$ and then compute the residual for the equality constraint, $A\boldsymbol{x}^{i+1} - \boldsymbol{b}$. We also replace the stepsize $\alpha^i$ with the fixed penalty parameter $p$. The ALMM algorithm is essentially the dual ascent method but using the augmented Lagrangian $L_p$ during the $\boldsymbol{x}$ update and stepsize $p$ in the $\boldsymbol{y}$ update.

One thing to consider is how to choose the penalty parameter $p$. It holds that as long as $p > 0$, the conditions for primal and dual feasibility for (3.1) are met. The primal feasibility conditions ensures that for a given optimal point the equality constraint is met. Mathematically we require $A\boldsymbol{x}^* - \boldsymbol{b} = \boldsymbol{0}$ for an optimal point $\boldsymbol{x}^*$. Dual feasibility requires that for a pair of optimal points $(\boldsymbol{x}^*, \boldsymbol{y}^*)$ the Lagrangian is in fact optimized. This translates to $\nabla L(\boldsymbol{x}^*, \boldsymbol{y}^*) = \nabla f(\boldsymbol{x}^*) + A^t\boldsymbol{y}^* = \boldsymbol{0}$. It will now be shown that $p > 0$ is suffice for ensuring primal and dual feasibility are met when using the method of multipliers to solve (3.1).

The optimality conditions for (3.1) are met when

$$A\boldsymbol{x}^* - \boldsymbol{b} = \boldsymbol{0} \text{ and } \nabla L(\boldsymbol{x}^*, \boldsymbol{y}^*) = \nabla f(\boldsymbol{x}^*) + A^t\boldsymbol{y}^* = \boldsymbol{0}.$$

By Definition, $\boldsymbol{x}^{i+1}$ minimizes $L_p(\boldsymbol{x}, \boldsymbol{y}^i)$. Therefore,

$$\begin{aligned}
\boldsymbol{0} &= \nabla_x L_p(\boldsymbol{x}^{i+1}, \boldsymbol{y}^i) \\
\boldsymbol{0} &= \nabla_x f(\boldsymbol{x}^{i+1}) + A^T\Big(\boldsymbol{y}^i + p(A\boldsymbol{x}^{i+1} - \boldsymbol{b})\Big) \\
\boldsymbol{0} &= \nabla_x f(\boldsymbol{x}^{i+1}) + A^T\boldsymbol{y}^{i+1}
\end{aligned}$$

We can see that for any stepsize $p > 0$, the iterate $(\boldsymbol{x}^{i+1}, \boldsymbol{y}^{i+1})$ is dual feasible and as the iterations proceed, the residual $A\boldsymbol{x}^{i+1} - \boldsymbol{b}$ converges to 0 thus attaining

optimality. The added robustness of the method of multipliers does come at a price. When the objective function $f$ is separable, it holds that the augmented Lagrangian, $L_p$ is not. This means the $\boldsymbol{x}$ update step cannot be carried out in parallel as done in dual decomposition. This will now lead us into the ADMM algorithm which attempts to marry the robustness of ALMM with the parallelization of dual decomposition.

## 3.2 ADMM and useful variations

### 3.2.1 Alternating direction method of multipliers

The ADMM, first introduced by Glowinski and Marroco (1975) and Gabay and Mercier (1976), presents an algorithm which allows for decomposability as seen in dual decomposition but with the added convergence properties seen in ALMM. The ADMM aims to solve the convex liner constrained optimization problem given below, by breaking the problem into smaller, easier to handle, sub-problems (as seen in dual decomposition). ADMM solves the following

$$\begin{aligned} &\text{minimize } f(\boldsymbol{x}) + g(\boldsymbol{z}) \\ &\text{subject to } A\boldsymbol{x} + B\boldsymbol{z} = \boldsymbol{c}, \end{aligned} \tag{3.10}$$

with variables $\boldsymbol{x} \in \mathbb{R}^n$ and $\boldsymbol{z} \in \mathbb{R}^m$, matrices $A \in \mathbb{R}^{n\times p}$ and $B \in \mathbb{R}^{m\times p}$ and vector $c \in \mathbb{R}^p$. Let $\boldsymbol{v}^* = inf\{f(\boldsymbol{x}) + g(\boldsymbol{z}) | A\boldsymbol{x} + B\boldsymbol{z} = \boldsymbol{c}\}$ be the optimal value of the above problem. $f$ and $g$ are assumed to be convex functions, $f : \mathbb{R}^n \to \mathbb{R}$, $g : \mathbb{R}^m \to \mathbb{R}$. The above problem can be viewed as the same problem given in (3.1), but with the the variable $\boldsymbol{x}$ in (3.1) separated into $\boldsymbol{x}$ and $\boldsymbol{z}$ in (3.10). Thus we have objective function $f(\boldsymbol{x})$ in (3.1) being separable into $f(\boldsymbol{x}) + g(\boldsymbol{z})$ given in (3.10). Just like the method of multipliers we construct the augmented Lagrangian.

$$L_p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{y}) = f(\boldsymbol{x}) + g(\boldsymbol{z}) + \boldsymbol{y}^T \big\{ A\boldsymbol{x} + B\boldsymbol{z} - \boldsymbol{c} \big\} + \frac{p}{2} \left\| A\boldsymbol{x} + B\boldsymbol{z} - \boldsymbol{c} \right\|_2^2 . \tag{3.11}$$

Applying the ideas from the update steps seen in the ALMM and dual decomposition, the ADMM is obtained by iterating between the following steps:

**Algorithm**

$$\boldsymbol{x}^{i+1} = \text{argmin}_x L_p(\boldsymbol{x}, \boldsymbol{z}^i, \boldsymbol{y}^i), \tag{3.12}$$

$$\boldsymbol{z}^{i+1} = \text{argmin}_z L_p(\boldsymbol{x}^{i+1}, \boldsymbol{z}, \boldsymbol{y}^i), \tag{3.13}$$

$$\boldsymbol{y}^{i+1} = \boldsymbol{y}^i + p(A\boldsymbol{x}^{i+1} + B\boldsymbol{z}^{i+1} - \boldsymbol{c}). \tag{3.14}$$

Similar to the method of multipliers, the ADMM uses a stepsize of $p$ during the gradient ascent step when updating the dual variable, $\boldsymbol{y}$. We define the residual for the equality constraint at iteration $i$ as $\boldsymbol{r}^i = A\boldsymbol{x}^i + B\boldsymbol{z}^i - \boldsymbol{c}$. The **ALMM** iterations for solving (3.10) is

**Algorithm** (ALMM)

$$(\boldsymbol{x}^{i+1}, \boldsymbol{z}^{i+1}) = \text{argmin}_{(x,z)} L_p(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{y}^i), \tag{3.15}$$

$$\boldsymbol{y}^{i+1} = \boldsymbol{y}^i + p(A\boldsymbol{x}^{i+1} + B\boldsymbol{z}^{i+1} - \boldsymbol{c}). \tag{3.16}$$

In the method of multipliers, the augmented Lagrangian is updated jointly with respect to both $\boldsymbol{x}$ and $\boldsymbol{z}$ (3.15) whereas the ADMM algorithm updates $\boldsymbol{x}$ (3.12) and $\boldsymbol{z}$ (3.13) in a sequential alternating manner, hence the term "alternating direction". The partitioning of the $\boldsymbol{x}$ and $\boldsymbol{z}$ update steps are what allows for decomposition using the method of multipliers when the objective function is separable into $f$ and $g$.

### 3.2.2 Convergence properties of ADMM

The current literature contains a rich documentation of the convergence properties of ADMM however we will limit ourselves to discussing only a few key properties. First, we make the following assumptions:

**Assumption 1:**
The extended real value functions $f : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$ and $g : \mathbb{R}^m \to \mathbb{R} \cup +\{\infty\}$ are closed, proper and convex.

A convex function, $f$, is said to be proper if $f(x) < +\infty$ for at least one $x \in \text{dom}\{f\}$ and $f(x) > -\infty$ for every $x \in \text{dom}\{f\}$. $f$, is said to be closed if $\forall\ \alpha \in \mathbb{R}$, the sub-level set $\{x \in \text{dom}\{f\} | f(x) < \alpha\}$ is a closed set. Assumption 1 is key in implying that the minimization problems given in (3.12) and (3.13) are in fact solvable. I.e. there exists an $\boldsymbol{x}^*$ and $\boldsymbol{z}^*$, not necessarily unique, which minimizes the augmented Lagrangian. It is also worth noting that Assumption 1 implies that $f$ and $g$ need not be differentiable, seeing as they can take values of $+\infty$.

**Assumption 2**

The unaugmented Lagrangian, $L_0$ has a saddle point, meaning there exists $(\boldsymbol{x}^*, \boldsymbol{z}^*, \boldsymbol{y}^*)$ for which

$$L_0(\boldsymbol{x}^*, \boldsymbol{z}^*, \boldsymbol{y}) \leq L_0(\boldsymbol{x}^*, \boldsymbol{z}^*, \boldsymbol{y}^*) \leq L_0(\boldsymbol{x}, \boldsymbol{z}, \boldsymbol{y}^*).$$

By Assumption 1, we have that $L_0(\boldsymbol{x}^*, \boldsymbol{z}^*, \boldsymbol{y}^*)$ is indeed finite for any saddle point $(\boldsymbol{x}^*, \boldsymbol{z}^*, \boldsymbol{y}^*)$. With this holding true, it follows that $(\boldsymbol{x}^*, \boldsymbol{z}^*)$ is a solution to (3.10) such that $A\boldsymbol{x}^* + B\boldsymbol{z}^* = \boldsymbol{c}$, $f(\boldsymbol{x}^*) < \infty$ and $g(\boldsymbol{z}^*) < \infty$. This assumption also implies that $\boldsymbol{y}^*$ is a dual optimal point and the optimal values of the primal and dual problems are equal (strong duality holds).

Under Assumptions 1 and 2, we have a few key convergence properties noticed by Eckstein and Bertsekas (1992) and Gabay (1983).

- **Residual Convergence:** $r^i \to 0$ as $i \to \infty$, the iterates approach feasibility as the algorithm proceeds
- **Objective Convergence:** $f(\boldsymbol{x}^i) + g(\boldsymbol{z}^i) \to v^*$ as $i \to \infty$, i.e. the objective function evaluated at the iterates approaches the optimal value
- **Dual Optimal Point** $\boldsymbol{y}^i \to \boldsymbol{y}^*$ as $i \to \infty$, the iterate $\boldsymbol{y}^i$ approaches a dual optimal point

The ADMM is relatively slow to converge with high accuracy, however the algorithm converges with modest accuracy generally after a few tens of iterations. For the kinds of distributed large scale problems we are considering, this is sufficient in many applications (Boyd et al., 2011).

### 3.2.3 Optimality conditions and stopping criteria

The sufficient optimality conditions needed for the problem given in (3.10) are: *primal feasibility*:

$$A\boldsymbol{x}^* + B\boldsymbol{z}^* - \boldsymbol{c} = \boldsymbol{0}, \tag{3.17}$$

and *dual feasibility*:

$$\boldsymbol{0} \in \partial f(\boldsymbol{x}^*) + A^T\boldsymbol{y}^*, \tag{3.18}$$
$$\boldsymbol{0} \in \partial g(\boldsymbol{z}^*) + B^T\boldsymbol{y}^*. \tag{3.19}$$

where $\partial$ is the sub-differential operator. When $f$ and $g$ are differentiable, the sub-differentials $\partial f$ and $\partial g$ are replaced with their gradients $\nabla f$ and $\nabla g$ and $\in$ is replaced with "=".

From the $\boldsymbol{z}$ update step, we have that $\boldsymbol{z}^{i+1}$ minimizes $L_p(\boldsymbol{x}^{i+1}, \boldsymbol{z}, \boldsymbol{y}^i)$, which implies

$$\begin{aligned}
\boldsymbol{0} &\in \partial_z L_p(\boldsymbol{x}^{i+1}, \boldsymbol{z}^{i+1}, \boldsymbol{y}^i), \\
\boldsymbol{0} &\in \partial g(\boldsymbol{z}^{i+1}) + B^T\boldsymbol{y}^i + pB^T(A\boldsymbol{x}^{i+1} + B\boldsymbol{z}^{i+1} - \boldsymbol{c}), \\
\boldsymbol{0} &\in \partial g(\boldsymbol{z}^{i+1}) + B^T\boldsymbol{y}^i + pB^T r^{i+1}, \\
\boldsymbol{0} &\in \partial g(\boldsymbol{z}^{i+1}) + B^T\boldsymbol{y}^{i+1}.
\end{aligned}$$

This shows that the iterates $(\boldsymbol{z}^i, \boldsymbol{y}^i)$ will always satisfy the dual feasibility optimlaity condition (3.19). Attaining optimality will then come down to conditions (3.17) and (3.18) being met. Using a similar argument, if $\boldsymbol{x}^{i+1}$ minimizes $L_p(\boldsymbol{x}, \boldsymbol{z}^i, \boldsymbol{y}^i)$ we have

$$\begin{aligned}
0 &\in \partial f(\boldsymbol{x}^{i+1}) + A^T\boldsymbol{y}^i + pA^T(A\boldsymbol{x}^{i+1} + B^T\boldsymbol{z}^i - \boldsymbol{c}), \\
0 &\in \partial f(\boldsymbol{x}^{i+1}) + A^T(\boldsymbol{y}^i + pr^{i+1} + pB(\boldsymbol{z}^i - \boldsymbol{z}^{i+1})), \\
&\in \partial f(\boldsymbol{x}^{i+1}) + A^T\boldsymbol{y}^{i+1} + pA^TB(\boldsymbol{z}^i - \boldsymbol{z}^{i+1}),
\end{aligned}$$

which leads us to

$$pA^TB(\boldsymbol{z}^{i+1} - \boldsymbol{z}^i) \in \partial f(\boldsymbol{x}^{i+1}) + A^T\boldsymbol{y}^{i+1}$$

so the quantity

$$\boldsymbol{s}^i = pA^TB(\boldsymbol{z}^{i+1} - \boldsymbol{z}^i)$$

can be viewed as the residual for the dual feasibility condition given in (3.18). We will refer to $\boldsymbol{s}^i$ as the *dual residual* at iteration $i$. We then have $\boldsymbol{r}^i = A\boldsymbol{x}^i + B\boldsymbol{z}^i - \boldsymbol{c}$ as the *primal residual* (for optimality condition 3.17) at iteration $i$.

As the algorithm proceeds, both $\boldsymbol{s}^i$ and $\boldsymbol{r}^i$ converge to 0 allowing a user-defined stopping rule to be implemented. A reasonable stopping criterion would be to terminate the algorithm when both the primal and dual residuals are small, i.e.,

$$\left\|\boldsymbol{r}^i\right\|_2 \leq \epsilon^{pri} \text{ and } \left\|\boldsymbol{s}^i\right\|_2 \leq \epsilon^{dual}$$

for some small tolerances $\epsilon^{pri} > 0$ and $\epsilon^{dual} > 0$.

### 3.2.4 Self-adaptive penalty parameter

One factor affecting the convergence speed of the ADMM is the penalty parameter and stepsize for the dual variable update, $p$. He et al. (2000) suggests an approach to improve convergence speed by replacing the fixed $p$ with a self adaptive term, $p^i$, whos value is updated after every iteration. This assists in convergence speed and also makes the algorithm less dependant on initial values of $p$. He et al. (2000) give a simple scheme for updating $p^i$:

$$p^i = \begin{cases} p^{i-1}(1+\tau^{i-1}), & \text{if } \left\|\boldsymbol{r}^{i-1}\right\|_2 > u\left\|\boldsymbol{s}^{i-1}\right\|_2, \\ p^{i-1}(1+\tau^{i-1})^{-1} & \text{if } \left\|\boldsymbol{s}^{i-1}\right\|_2 > u\left\|\boldsymbol{r}^{i-1}\right\|_2, \\ p^{i-1} & \text{otherwise.} \end{cases}$$

where $u > 0$ and $\tau^{(i)}$ are parameters. Common choices for these parameters are $u = 10$ and $\tau^{(i)} = 1$ for all iterations $i$. The main idea here is to keep the norms

of the primal and dual residuals to within some factor, $u$, of each other while they both converge to 0. When $p$ is large, a heavier penalty is placed on constraint violations thus producing smaller values of the primal residual. At the same time however, the value of the dual residuals will tend to be high since by definition $s^i = pA^TB(\boldsymbol{z}^{i+1} - \boldsymbol{z}^i)$ which will increase as $p$ increases. The above scheme attempts to find $p$ such that $\boldsymbol{s}^i$ and $\boldsymbol{r}^i$ are "relatively" close.

### 3.2.5 Global variable consensus optimization

Let us consider the optimization problem

$$\text{minimize } \sum_{k=1}^{N} f_k(\boldsymbol{x}), \tag{3.20}$$

with global variable, $\boldsymbol{x} \in \mathbb{R}^n$, and $f_k : \mathbb{R}^n \to \mathbb{R} \cup \{+\infty\}$. With the $f_k$ being the $i^{th}$ term in the objective, these functions can also encode constraints by assigning $f_k$ to be $+\infty$. With the presence of the global variable $\boldsymbol{x}$, the problem cannot currently be decomposed into smaller sub problems, since the sum is not separable in $\boldsymbol{x}$. We aim to solve (3.20) such that each term in the objective can be handled by it's own local machine/ processing element i.e, solved in a distributed manner like dual decomposition or ADMM. This problem appears in many contexts and applications. For instance, we can view the above as a model fitting problem with parameter vector $\boldsymbol{x}$ and $f_k(\boldsymbol{x})$ being the loss function associated with the $k^{th}$ block of data. With the aim of developing a distributed algorithm, we can re-express the above problem with local variables $\boldsymbol{x}_k \in \mathbb{R}^n$ and global variable $\boldsymbol{z} \in \mathbb{R}^n$. Then (3.20) can be rewritten as:

$$\begin{aligned} \text{minimize } & \sum_{k=1}^{N} f_k(\boldsymbol{x}_k) \\ \text{subject to } & \boldsymbol{x}_k - \boldsymbol{z} = \boldsymbol{0}, k = 1, ..., N. \end{aligned} \tag{3.21}$$

The re-expression of the problem allows for a distributed optimization approach which the objective function now being separable across its variable $\boldsymbol{x}$. The constraint term $\boldsymbol{x}_k - \boldsymbol{z} = \boldsymbol{0}, \forall k$, ensures the local variables stay in "consensus" i.e, all

the local variables take the same value. We call (3.21) the *global consensus problem* since all the local variables must "agree" and be in consensus. This idea of the global consensus problem can be viewed as a technique for transforming non-separable additive objective functions (3.20) into a sum that is separable (3.21).

Applying a distributed ADMM method to solve (3.21), we first start with the augmented Lagrangian

$$L_P(\boldsymbol{x}_1, ..., \boldsymbol{x}_N, \boldsymbol{z}, \boldsymbol{y}) = \sum_{k=1}^{N} \{ f_k(\boldsymbol{x}_k) + \boldsymbol{y}_k^T(\boldsymbol{x}_k - \boldsymbol{z}) + \frac{p}{2} \|\boldsymbol{x}_k - \boldsymbol{z}\|_2^2 \}.$$

The resulting ADMM algorithm update steps are:

**Algorithm**

$$\boldsymbol{x}_k^{i+1} := \text{argmin}_x \big( f_k(\boldsymbol{x}_k) + \boldsymbol{y}_k^{iT}(\boldsymbol{x}_k - \boldsymbol{z}^i) + \frac{p}{2} \left\|\boldsymbol{x}_k - \boldsymbol{z}^i\right\|_2^2 \big) \tag{3.22}$$

$$\boldsymbol{z}^{i+1} = \frac{1}{N} \sum_{k=1}^{N} \big\{ \boldsymbol{x}_k + \frac{1}{p} \boldsymbol{y}_k \big\} \tag{3.23}$$

$$\boldsymbol{y}_k^{i+1} := \boldsymbol{y}_k^i + p(\boldsymbol{x}_k^{i+1} - \boldsymbol{z}^{i+1}) \tag{3.24}$$

Here, we write $(\boldsymbol{y}_k^i)^T = \boldsymbol{y}_k^{iT}$ to lighten the notation. The first and last steps of the algorithm are carried out in parallel for $k = 1, ..., N$, as desired. Similar to other algorithms we've discussed, global consensus optimization alternates between collection and broadcast steps. The algorithm can be further simplified. First, re-writing $\boldsymbol{z}^{i+1}$ we get

$$\boldsymbol{z}^{i+1} = \frac{1}{N} \sum_{k=1}^{N} \big\{ \boldsymbol{x}_k + \frac{1}{p} \boldsymbol{y}_k \big\} = \bar{\boldsymbol{x}}^{i+1} + \frac{1}{p} \bar{\boldsymbol{y}}^{i+1}.$$

Similarly, averaging over $\boldsymbol{y}$ we get;

$$\bar{\boldsymbol{y}}^{i+1} = \bar{\boldsymbol{y}}^i + p(\bar{\boldsymbol{x}}^{i+1} - \boldsymbol{z}^{i+1})$$

Then substituting the first equation into the second, we have that $\boldsymbol{y}^{i+1}$ have average value 0 after the first iteration, meaning that the $\boldsymbol{z}$ update step can be

replaced by $\boldsymbol{z}^{i+1} = \bar{\boldsymbol{x}}^{i+1}$, making the simplified algorithm:

**Algorithm**

$$\boldsymbol{x}_k^{i+1} := \text{argmin}_x\big(f_k(\boldsymbol{x}_k) + \boldsymbol{y}_k^{iT}(\boldsymbol{x}_k - \bar{\boldsymbol{x}}^i) + \frac{p}{2}\left\|\boldsymbol{x}_k - \bar{\boldsymbol{x}}^i\right\|_2^2\big), \tag{3.25}$$

$$\boldsymbol{y}_k^{i+1} := \boldsymbol{y}_k^i + p(\boldsymbol{x}_k^{i+1} - \bar{\boldsymbol{x}}^{i+1}). \tag{3.26}$$

Under this algorithm, the $k^{th}$ processing unit can handle their own respective $\boldsymbol{x}$ and $\boldsymbol{y}$ update step while communicating with a global machine to compute the $\bar{\boldsymbol{x}}^i$ after each iteration. Again this allows for both the $\boldsymbol{x}$ and $\boldsymbol{y}$ update steps to be carried out in parallel. The dual variables are updated driving the local variables into consensus while the quadratic regularization (linear term in the quadratic) helps to pull the variables towards their average value while minimizing the local $f_k$.

For global consensus ADMM, the primal and dual residuals are;

$$\boldsymbol{r}^i = (\boldsymbol{x}_1^i - \bar{\boldsymbol{x}}^i, ..., \boldsymbol{x}_N^i - \bar{\boldsymbol{x}}^i) \text{ and } \boldsymbol{s}^i = -p(\bar{\boldsymbol{x}}^i - \bar{\boldsymbol{x}}^{i-1}, ..., \bar{\boldsymbol{x}}^i - \bar{\boldsymbol{x}}^{i-1})$$

with squared norms,

$$\left\|\boldsymbol{r}^i\right\|_2^2 = \sum_{k=1}^{N}\left\|\boldsymbol{x}_k^i - \bar{\boldsymbol{x}}^i\right\|_2^2 \text{ and } \left\|\boldsymbol{s}^i\right\|_2^2 = Np^2\left\|\bar{\boldsymbol{x}}^i - \bar{\boldsymbol{x}}^{i-1}\right\|_2^2.$$

The first term $\|\boldsymbol{r}^i\|_2^2$ is N times the standard deviation of the points $\boldsymbol{x}_1, ..., \boldsymbol{x}_N$, a natural measure for the lack of consensus.

In the context of a model fitting problem, using a likelihood based approach, the $f_k$ can be viewed as the negative log-likelihood contribution, given the observed data present on the $k^{th}$ processor or machine. By now we can begin to see the appeal of applying a consensus ADMM algorithm for a DRM model fitting problem under a distributed data setting. We will find great use of the global consensus optimization when solving for the MELE, $\hat{\boldsymbol{\theta}}$.

Recalling back to the testing of hypotheses regarding distributional differences, we compute the DELR test statistic using $\tilde{\boldsymbol{\theta}}$, the MELE under the null constraint.

This requires us to solve a global consensus problem with an additional constraint added by the null hypothesis. This will lead us into a slight variation of the above global variable consensus in which a regularizer is added to the objective sum, enforcing an additional null constraint.

### 3.2.6 Global variable consensus with regularization

The global variable consensus with regularization starts with the global variable consensus problem given in (3.21), and adds an additional term, $g(\boldsymbol{z})$, to the objective sum. This newly added term often represents an additional constraint or regularization. The global variable consensus with regularization problem is written as

$$\begin{aligned} &\text{minimize } \sum_{k=1}^{N} f_k(\boldsymbol{x}_k) + g(\boldsymbol{z}) \\ &\text{subject to } \boldsymbol{x}_k - \boldsymbol{z} = \boldsymbol{0}, k = 1, ..., N. \end{aligned} \tag{3.27}$$

Extending the theory from the "vanilla" global variable consensus, we attain the ADMM iterations for the regularization variant:

**Algorithm**

$\boldsymbol{x}_k^{i+1} := \text{argmin}_{x_k}\big(f_k(\boldsymbol{x}_k) + \boldsymbol{y}_k^{iT}(\boldsymbol{x}_k - \boldsymbol{z}^i) + \frac{p}{2}\left\|\boldsymbol{x}_k - \boldsymbol{z}^k\right\|_2^2)\big),$

$\boldsymbol{z}^{i+1} := \text{argmin}_z\big(g(\boldsymbol{z}) + \sum_{k=1}^{N}(-\boldsymbol{y}_k^{iT}\boldsymbol{z} + \frac{p}{2}\left\|\boldsymbol{x}_k^{i+1} - \boldsymbol{z}\right\|_2^2)\big),$

$\boldsymbol{y}_k^{i+1} := \boldsymbol{y}_k^i + p(\boldsymbol{x}_k^{i+1} - \boldsymbol{z}^{i+1}).$

When $g(\boldsymbol{z})$ is non- zero, we do not have $\bar{\boldsymbol{y}}^{i+1} = \boldsymbol{0}$ so the $\boldsymbol{z}$ update step cannot be dropped from the algorithm as seen in vanilla global consensus. In the application to the DELRT, the choice of $g(\boldsymbol{z})$ will depend on the hypotheses of interest being tested, specifically the null constraint. Deciding on the term $g(\boldsymbol{z})$ will be studied in

the subsequent chapter.

# Chapter 4

# Adapting DEL Inference Under the DRM to Distributed Data

In this chapter, which contains the main contribution of this thesis, we present methods for carrying out DEL inference using the DRM in the presence of distributed samples. First and foremost we will discuss the application of the ADMM global variable consensus when solving for the MELE of the DRM parameter, $\boldsymbol{\theta}$. Next, methods for computing $\hat{F}_0$ and $\hat{p}_{kj} = d\hat{F}_0(x_{kj})$ for the baseline distribution will be given. An extension of these methods will then be applied to estimating the non-baseline distribution functions, $F_k(x)$ , $k = 1, ..., m$. Then, methods to estimate the quantiles for each of the $m + 1$ samples, using the previously computed $\hat{F}_k$, will be presented. Finally we look into detecting differences amongst the distributions $f_k$ by carrying out the dual empirical likelihood ratio test. To carry out this test with distributed data, we propose the application of the global variable consensus with regularization algorithm to compute $\tilde{\boldsymbol{\theta}}$ used in the test statistic.

## 4.1 Solving for the MELE using ADMM global variable consensus

To solve for the MELE of $\boldsymbol{\theta}$ we must, by definition, find $\boldsymbol{\theta}$ which maximizes the DEL function or equivalently, minimizes the negative DEL. We write the MELE as

$$\hat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}} \ -l_n(\boldsymbol{\theta}), \tag{4.1}$$

where the negative DEL is

$$-l_n(\boldsymbol{\theta}) = \sum_{k=0}^{m}\sum_{j=1}^{n_k}\log\Big\{\sum_{r=0}^{m}\hat{\lambda}_r\exp\{\alpha_r+\boldsymbol{\beta}_r^T\boldsymbol{q}(x_{kj})\}\Big\} - \sum_{k=0}^{m}\sum_{j=1}^{n_k}\{\alpha_r+\boldsymbol{\beta}_r^T\boldsymbol{q}(x_{kj})\}.$$

We will now re-write the objective function, $-l_n(\boldsymbol{\theta})$, into a form more recognizable to the global variable census problem given in (3.20).

rewriting $l_n(\boldsymbol{\theta})$ we get

$$\begin{aligned}
-l_n(\boldsymbol{\theta}) &= \sum_{k=0}^{m}\sum_{j=1}^{n_k}\log\Big\{\sum_{r=0}^{m}\hat{\lambda}_r\exp\{\alpha_r+\boldsymbol{\beta}_r^T\boldsymbol{q}(x_{kj})\}\Big\} - \sum_{k=0}^{m}\sum_{j=1}^{n_k}\{\alpha_r+\boldsymbol{\beta}_r^T\boldsymbol{q}(x_{kj})\} \\
&= \sum_{k=0}^{m}\Big\{\sum_{j=1}^{n_k}\Big\{\log\{\sum_{r=0}^{m}\hat{\lambda}_r\exp\{\alpha_r+\boldsymbol{\beta}_r^T\boldsymbol{q}(x_{kj})\}\} - \{\alpha_r+\boldsymbol{\beta}_r^T\boldsymbol{q}(x_{kj})\Big\}\Big\} \\
&= \sum_{k=0}^{m} g_k(\boldsymbol{\theta}),
\end{aligned}$$

where $g_k(\boldsymbol{\theta}) = \sum_{j=1}^{n_k}\Big\{\log\{\sum_{r=0}^{m}\hat{\lambda}_r\exp\{\alpha_r+\boldsymbol{\beta}_r^T\boldsymbol{q}(x_{kj})\}\} - \{\alpha_r+\boldsymbol{\beta}_r^T\boldsymbol{q}(x_{kj})\Big\}$ is a function of the $k^{th}$ distributed sample and the shared DRM paramater $\boldsymbol{\theta}$.

Using this re-expression we can see the MELE re-written in terms of a problem solvable using the global variable consensus optimization:

$$\hat{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}}\ \sum_{k=0}^{m} g_k(\boldsymbol{\theta})$$

Transforming the problem into the global variable consensus problem given in (3.21) we have:

$$\begin{aligned}
&\text{minimize}_{\boldsymbol{\theta},\boldsymbol{z}}, \quad \sum_{k=0}^{m} g_k(\boldsymbol{\theta}_k), \\
&\text{subject to} \quad \boldsymbol{\theta}_k - \boldsymbol{z} = \mathbf{0}, \text{for } k = 0, ..., m,
\end{aligned} \tag{4.2}$$

with local variables $\boldsymbol{\theta}_k$ and global variable $\boldsymbol{z}$. The augmented Lagrangian for (4.2) is

$$L_p(\boldsymbol{\theta}_0, ..., \boldsymbol{\theta}_m, \boldsymbol{z}, \boldsymbol{y}_0, ..., \boldsymbol{y}_m) = \sum_{k=0}^{m} \left\{ g_k(\boldsymbol{\theta}_k) + \boldsymbol{y}_k^T(\boldsymbol{\theta}_k - \boldsymbol{z}) + \frac{p}{2} \|\boldsymbol{\theta}_k - \boldsymbol{z}\|_2^2 \right\}.$$

Applying the global consensus ADMM iterations, we obtain the update steps for the local and global variables along with Lagrange multiplier at iteration $i+1$.

**Algorithm**,

$$\boldsymbol{\theta}_k^{i+1} := \text{argmin}_{\theta_k} \left( g_k(\boldsymbol{\theta}_k) + \boldsymbol{y}_k^{iT}(\boldsymbol{\theta}_k - \boldsymbol{z}^i) + \frac{p}{2} \left\|\boldsymbol{\theta}_k - \boldsymbol{z}^i\right\|_2^2 \right), \text{ for } k = 0, ..., m, \tag{4.3}$$

$$\boldsymbol{z}^{i+1} := \bar{\boldsymbol{\theta}}^{i+1} = \frac{1}{m} \sum_{k=0}^{m} \boldsymbol{\theta}_k^{i+1}, \text{ Computed Globally}, \tag{4.4}$$

$$\boldsymbol{y}_k^{i+1} := \boldsymbol{y}_k^i + p(\boldsymbol{\theta}_k^{i+1} - \boldsymbol{z}^{i+1}), \text{ for } k = 0, ..., m. \tag{4.5}$$

In the above algorithm, we first update the local variables $\boldsymbol{\theta}_k^{i+1}$, which is carried out independently on each local machine. The minimization problems for this update step can be solved in parallel due to the separability of the Lagrangian. Next we communicate the summary values $\boldsymbol{\theta}_k^{i+1}$ from each local machine to a global machine. On the global machine, the global variable $\boldsymbol{z}^{i+1}$ is updated by averaging over the local $\boldsymbol{\theta}_k^{i+1}$, to obtain $\boldsymbol{z}^{i+1} = \bar{\boldsymbol{\theta}}^{i+1}$. We showed in section 3.2.5 that the global variable $\boldsymbol{z}$ can be replaced by the average of local variables, therefore simplifying the algorithm. Next the globally computed $\bar{\boldsymbol{\theta}}^{i+1}$ is communicated back to the local machines to be used in updating $\boldsymbol{y}_k$. The update steps (4.3) to (4.5) are repeated until the primal and dual residuals are smaller than some pre-set values.

Various algorithms and methods can be applied to solve the $k$ unconstrained optimizations in (4.3). Some methods, such as the Newton-Raphson method, require the gradient of the $k^{th}$ term in the augmented Lagrangian (function to be minimized) to solve the $k^{th}$ minimization problem. We will call the $k^{th}$ term in the augmented Lagrangian $h_k(\boldsymbol{\theta}_k)$, i.e.,

$$h_k(\boldsymbol{\theta}_k) = g_k(\boldsymbol{\theta}_k) + \boldsymbol{y}_k^{iT}(\boldsymbol{\theta}_k - \boldsymbol{z}^i) + \frac{p}{2}\left\|\boldsymbol{\theta}_k - \boldsymbol{z}^i\right\|_2^2. \text{ for } k = 0, ..., m$$

and the $\boldsymbol{\theta}_k$ update step (4.3) is

$$\boldsymbol{\theta}_k^{i+1} := \text{argmin}_{\boldsymbol{\theta}_k} h_k(\boldsymbol{\theta}_k).$$

### 4.1.1 Gradient of $h_k(\boldsymbol{\theta}_k)$

In an attempt to ease confusion between $\boldsymbol{\theta}_k$, the $k^{th}$ local variable in problem (4.2) and $\boldsymbol{\theta}_k = (\alpha_k, \boldsymbol{\beta}_k^T)^T$, the parameters of the $k^{th}$ distribution in the DRM, an additional notation change will now be made. Let $\boldsymbol{\theta}_k$ be the $k^{th}$ local variable and $\boldsymbol{\theta}_{k,s} = (\alpha_{k,s}, \boldsymbol{\beta}_{k,s}^T)^T$ be the model parameters for the $s^{th}$ distribution belonging to the $k^{th}$ local variable.

We can express the gradient of $h_k(\theta_k)$ as

$$\nabla h_k(\boldsymbol{\theta_k}) = \nabla g_k(\boldsymbol{\theta}_k) + \boldsymbol{y}_k^{(i)} + p(\boldsymbol{\theta_k} - \boldsymbol{z}_k{}^i). \tag{4.6}$$

We will now focus on the term $\nabla g_k(\boldsymbol{\theta}_k)$, to complete the expression of $\nabla h_k$. Let us first introduce some notation which will assist in developing a compact expression for $\nabla g_k$.

Let,

$$\begin{aligned}
\rho_s(\boldsymbol{\theta}_k, x) &= \exp(\alpha_{k,s} + \boldsymbol{\beta}_{k,s}^T \boldsymbol{q}(x)) \text{ for } s = 0, .., m, \\
\boldsymbol{h}(\boldsymbol{\theta}_k, x) &= (\hat{\lambda}_1 \rho_1(\boldsymbol{\theta}_k, x), ..., \hat{\lambda}_m \rho_m(\boldsymbol{\theta}_k, x)^T, \\
s(\boldsymbol{\theta}_k, x) &= \hat{\lambda}_0 + \sum_{s=1}^{m} \hat{\lambda}_s \rho_s(\boldsymbol{\theta}_k, x), \\
\mathcal{L}_k(\boldsymbol{\theta}_k, x_{kj}) &= -\log(\sum_{r=0}^{m} \hat{\lambda}_r \rho_r(\boldsymbol{\theta_k}, x_{kj})) + \alpha_{k,k} + \boldsymbol{\beta}_{k,k}^T \boldsymbol{q}(x_{kj}).
\end{aligned}$$

We notice that $g_k(\boldsymbol{\theta}_k) = -\sum_{j=1}^{n_k} \mathcal{L}_k(\boldsymbol{\theta}_k, x_{kj})$. So we have the gradient of $g_k(\boldsymbol{\theta}_k)$ as,

$$\nabla g_k(\boldsymbol{\theta}_k) = -\sum_{j=1}^{n_k} \nabla \mathcal{L}_k(\boldsymbol{\theta}_k, x_{kj}). \tag{4.7}$$

Using the above notation, we will now present the gradient of $\mathcal{L}_k(\boldsymbol{\theta}_k, x_{kj})$. Let $\boldsymbol{e}_k$ be a vector of length $m$ with the $k^{th}$ entry being 1 and all others being 0, and let $\delta_{ij} = 1$ when $i = j$ and 0 otherwise. We can then express of $\mathcal{L}_k(\boldsymbol{\theta}_k, x_{kj})$ as

$$\nabla \mathcal{L}_k(\boldsymbol{\theta}, x_{kj}) = \begin{bmatrix} \frac{\partial \mathcal{L}_k(\boldsymbol{\theta}_k, x_{kj})}{\partial \boldsymbol{\alpha}} \\ \\ \frac{\partial \mathcal{L}_k(\boldsymbol{\theta}_k, x_{kj})}{\partial \boldsymbol{\beta}} \end{bmatrix}$$

where,

$$\frac{\partial \mathcal{L}_k(\boldsymbol{\theta}_k, x_{kj})}{\partial \boldsymbol{\alpha}} = (1 - \delta_{k0})\boldsymbol{e}_k - \frac{\boldsymbol{h}(\boldsymbol{\theta}_k, x_{kj})}{s(\boldsymbol{\theta}_k, x_{kj})}$$

and

$$\frac{\partial \mathcal{L}_k(\boldsymbol{\theta}_k, x_{kj})}{\partial \boldsymbol{\beta}} = \frac{\partial \mathcal{L}_k(\boldsymbol{\theta}_k, x_{kj})}{\partial \boldsymbol{\alpha}} \otimes \boldsymbol{q}(x_{kj}) = \left[(1 - \delta_{k0})\boldsymbol{e}_k - \frac{\boldsymbol{h}(\boldsymbol{\theta}_k, x_{kj})}{s(\boldsymbol{\theta}_k, x_{kj})}\right] \otimes \boldsymbol{q}(x_{kj})$$

with $\otimes$ being the Kronecker product. Using the definition of $\nabla g_k(\boldsymbol{\theta}_k)$ in (4.7) and the above results, we obtain

$$\nabla g_k(\boldsymbol{\theta}_k) = \begin{bmatrix} -\sum_{j=1}^{n_k}(1 - \delta_{k0})\boldsymbol{e}_k - \frac{\boldsymbol{h}(\boldsymbol{\theta}_k, x_{kj})}{s(\boldsymbol{\theta}_k, x_{kj})} \\ \\ -\sum_{j=1}^{n_k}[(1 - \delta_{k0})\boldsymbol{e}_k - \frac{\boldsymbol{h}(\boldsymbol{\theta}_k, x_{kj})}{s(\boldsymbol{\theta}_k, x_{kj})}] \otimes \boldsymbol{q}(x_{kj}) \end{bmatrix}$$

Substituting the above expression into (4.6), we obtain a closed form, compact expression for $\nabla h_k(\boldsymbol{\theta}_k)$.

### 4.1.2 Primal and dual residuals

The squared primal and dual residuals for the global variable consensus algorithm, applied to solving for $\hat{\boldsymbol{\theta}}$, are

$$\begin{aligned} \left\|\boldsymbol{r}^{(i)}\right\|_2^2 &= \sum_{k=0}^{m} \left\|\boldsymbol{\theta}_k^{(i)} - \boldsymbol{z}^{(i)}\right\|_2^2, \\ \left\|\boldsymbol{s}^{(i)}\right\|_2^2 &= (m+1)p^2 \left\|\boldsymbol{z}^{(i)} - \boldsymbol{z}^{(i-1)}\right\|_2^2. \end{aligned}$$

The algorithm can be then terminated when

$$\left\|\boldsymbol{r}^{i}\right\|_2^2 \leq \epsilon^{pri} \text{ and } \left\|\boldsymbol{s}^{i}\right\|_2^2 \leq \epsilon^{dual},$$

for some small tolerances $\epsilon^{pri} > 0$ and $\epsilon^{dual} > 0$.

## 4.2 Estimating $F_0(x)$ and $p_{kj}$, of the baseline distribution

Once the DRM has been fit to the distributed data via solving for $\hat{\boldsymbol{\theta}}$, components of the baseline distribution can then be estimated. Theories will now be presented for estimating the distribution function, $\hat{F}_0(x)$, and the density function $f_0(x_{kj})$, for the baseline distribution under the setting of distributed data.

We will begin by estimating the baseline probabilities, $p_{kj} = dF_0(x_{kj})$, for each $x_{kj}$ $k = 0, ..., m$, $j = 1, .., n_k$. Recall that the likelihood based estimate of $p_{kj}$ given in (2.3) is

$$\hat{p}_{kj} = n^{-1}\Big\{1 + \sum_{r=1}^{m} \hat{\lambda}_r[\exp\{\hat{\alpha}_r + \hat{\boldsymbol{\beta}}_r^T \boldsymbol{q}(x_{kj})\} - 1]\Big\}^{-1},$$

which is a function of $\hat{\boldsymbol{\theta}}$. To assist in the development of the theory let us first introduce some additional useful notation.

Let

$$\hat{\boldsymbol{p}} = [\hat{p}_{01}, .., \hat{p}_{0n_0}, ..., \hat{p}_{m1}, ..., \hat{p}_{mn_m}]^T$$

be a vector containing the $\hat{p}_{kj}$ $\forall k = 0, ..., m$ and $j = 1, ..., n_k$. We also have

$$\hat{\boldsymbol{p}}_k = [\hat{p}_{k1}, ..., \hat{p}_{kn_k}]^T = \begin{bmatrix} n^{-1}\Big\{1 + \sum_{r=1}^{m} \hat{\lambda}_r[\exp\{\hat{\alpha}_r + \hat{\boldsymbol{\beta}}_r^T \boldsymbol{q}(x_{k1})\} - 1]\Big\}^{-1} \\ . \\ . \\ . \\ n^{-1}\Big\{1 + \sum_{r=1}^{m} \hat{\lambda}_r[\exp\{\hat{\alpha}_r + \hat{\boldsymbol{\beta}}_r^T \boldsymbol{q}(x_{kn_k})\} - 1]\Big\}^{-1} \end{bmatrix},$$

a vector of the $\hat{p}_{kj}$ for the $k^{th}$ distributed sample. It is then clear that

$$\hat{\boldsymbol{p}} = [\hat{\boldsymbol{p}}_0^T, ..., \hat{\boldsymbol{p}}_m^T]^T.$$

It follows that $\hat{\boldsymbol{p}}_k$ is a function of $\hat{\boldsymbol{\theta}}$ and the $k^{th}$ sample. Similar to update steps in the ADMM iterations, this allows for the $\hat{\boldsymbol{p}}$ vector to be computed in a decentralized fashion where the $k^{th}$ local machine is responsible for computing $\hat{\boldsymbol{p}}_k$. A simple distributed approach for computing $\hat{\boldsymbol{p}}$ consists of a broadcast step where the MELE $\hat{\boldsymbol{\theta}}$ is shared with the $m + 1$ local machines. Each local machine then proceeds to compute, in parallel, their respective $\hat{\boldsymbol{p}}_k$. Note that there is no need to broadcast the locally computed $\hat{\boldsymbol{p}}_k$ to a central computer since it is also possible to carry out the estimation of $F_k$ $k = 1, ..., m$, in a similar parallel fashion.

Once the $\hat{\boldsymbol{p}}_k$ have been computed for $k = 0, ..., m$, we are then able to compute estimates for the baseline distribution function, $\hat{F}_0(x)$, a function of $\hat{\boldsymbol{p}}$ and the data. Recall the MELE for $F_0(x)$, given in equation (2.5) is

$$\hat{F}_0(x) = \sum_{k=0}^{m} \sum_{j=1}^{n_k} \hat{p}_{kj} \mathbb{1}(x_{kj} \leq x).$$

The structure of $\hat{F}_0(x)$ is appealing in the distributed data setting since like the $\hat{\boldsymbol{p}}$ vector, it is seperable and can be computed in a distributed fashion. We start with re-writing $\hat{F}_0(x)$ as

$$\hat{F}_0(x) = \sum_{k=0}^{m} \gamma_k(x, \underline{\boldsymbol{x}}_k, \hat{\boldsymbol{p}}_k),$$

where $\gamma_k(x, \underline{\boldsymbol{x}}_k, \hat{\boldsymbol{p}}_k) = \sum_{j=1}^{n_k} \hat{p}_{kj} \mathbb{1}(x_{kj} \leq x)$, a function of $x$, $\hat{\boldsymbol{p}}_k$, and $\underline{\boldsymbol{x}}_k$ ($k^{th}$ sample of data). With each $\gamma_k$ being a function of $x$ and the information contained on the $k^{th}$ local machine , $(\underline{\boldsymbol{x}}_k, \hat{\boldsymbol{p}}_k)$, we obtain a distributed algorithm as follows:

**Algorithm**

Broadcast $x$ to each of the $m+1$ local machines. (4.8)

Compute $\gamma_k(x, \underline{\boldsymbol{x}}_k, \hat{\boldsymbol{p}}_k)$ for $k = 0, ..., m$. (4.9)

Transfer the values of $\gamma_k(x, \underline{\boldsymbol{x}}_k, \hat{\boldsymbol{p}}_k)$ to the global machine. (4.10)

Compute $\hat{F}_0(x) = \sum_{k=0}^{m} \gamma_k(x, \underline{\boldsymbol{x}}_k, \hat{\boldsymbol{p}}_k)$ on the global machine. (4.11)

with (4.9) being carried out in parallel across the $m+1$ local machines. The above algorithm is represented visually in Figure 4.1 below.

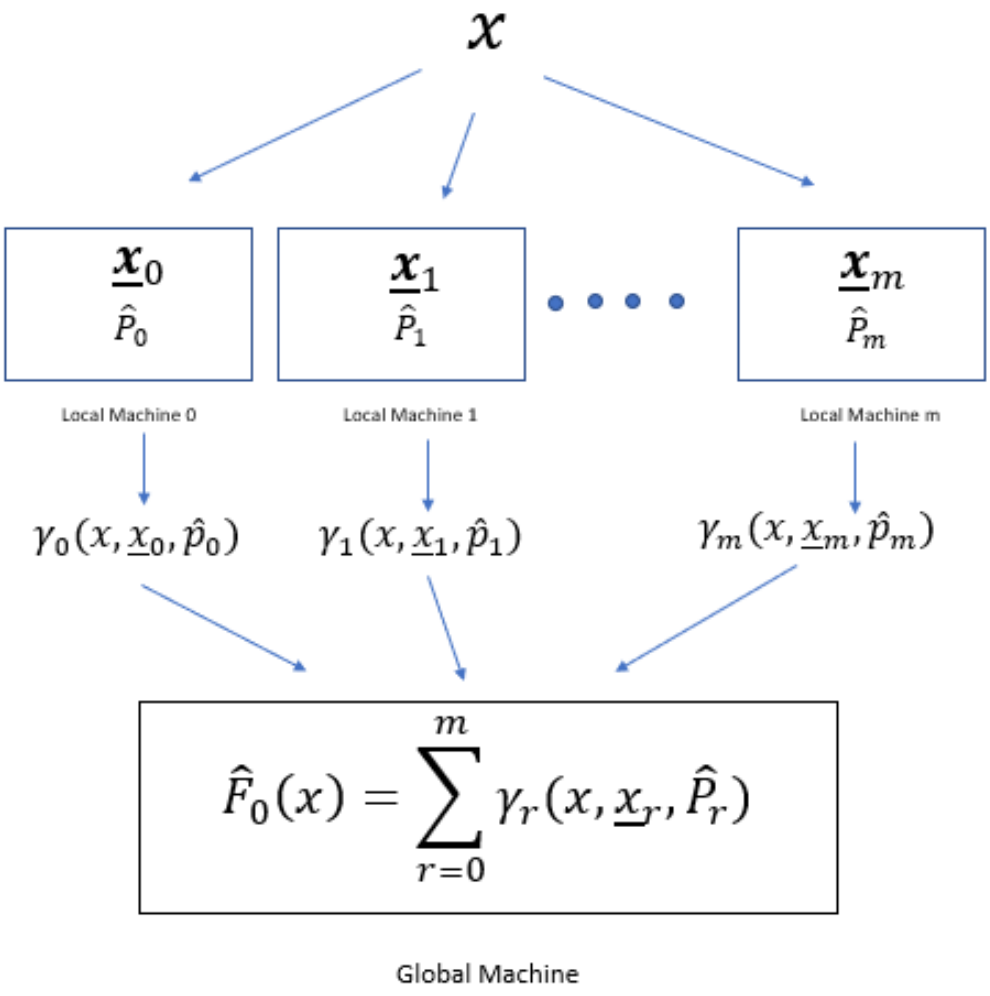


Figure 4.1: Visual depiction of the distributed approach to estimating the baseline distribution function of the DRM.

In the next section we, extend the above methods to estimating the non-baseline distribution functions, $\hat{F}_k(x)$, $k = 1, ..., m$.

## 4.3 Estimating the non-baseline distribution functions, $\hat{F}_1(x), ..., \hat{F}_m(x)$

The above distributed algorithm for computing $\hat{F}_0(x)$ will now be extended to encompass the non-baseline distributions. Recall the MELE for $F_k(x)$, $k = 1, ..., m$ is given by

$$\hat{F}_k(x) = \sum_{r=0}^{m} \sum_{j=1}^{n_r} \exp\{\hat{\alpha}_k + \hat{\boldsymbol{\beta}}_k^T \boldsymbol{q}(x_{rj})\} \hat{p}_{rj} \mathbb{1}(x_{rj} \leq x).$$

Similar to $\hat{F}_0(x)$, we can re-write $\hat{F}_k(x)$ into a separable sum which makes the distributed computation possible:

$$\hat{F}_k(x) = \sum_{r=0}^{m} \psi_r(x, \underline{\boldsymbol{x}}_r, \hat{\boldsymbol{\theta}}_k, \hat{\boldsymbol{p}}_r).$$

Where $\psi_r(x, \underline{\boldsymbol{x}}_r, \hat{\boldsymbol{\theta}}_k, \hat{\boldsymbol{p}}_r) = \sum_{j=1}^{n_r} \exp[\hat{\alpha_k} + \hat{\boldsymbol{\beta}}_k^T \boldsymbol{q}(x_{rj})] \hat{p}_{rj} \mathbb{1}(x_{rj} \leq x)$ is a function of $x$, $\boldsymbol{\theta}_k$ and information stored on the $r^{th}$ local machine $(\underline{\boldsymbol{x}}_r, \hat{\boldsymbol{p}}_r)$. The distributed algorithm for computing $\hat{F}_k(x)$ for $k = 1, ..., m$ is as follows:

**Algorithm**

Brodcast $x$ and $\hat{\boldsymbol{\theta}}_k$ to each of the $m + 1$ local machines. (4.12)

Compute $\psi_r(x, \underline{\boldsymbol{x}}_r, \hat{\boldsymbol{\theta}}_k, \hat{\boldsymbol{p}}_r)$ for $r = 0, ..., m$ (4.13)

Brodcast the values of $\psi_r(x, \underline{\boldsymbol{x}}_r, \hat{\boldsymbol{\theta}}_k, \hat{\boldsymbol{p}}_r)$ to the global machine. (4.14)

Compute $\hat{F}_k(x) = \sum_{r=0}^{m} \psi_r(x, \underline{\boldsymbol{x}}_r, \hat{\boldsymbol{\theta}}_k, \hat{\boldsymbol{p}}_r)$ on the global machine. (4.15)

with the $\psi_r$ computation in (4.12) being carried out in paralell for $r = 0, ..., m$.

## 4.4 Quantile estimation for the distributions $F_0, ..., F_m$

Once the estimates for the $m + 1$ distribution functions have been computed, we can proceed with estimating the quantiles for each distribution. Referring back to

the definition given in (2.6), the EL-based estimator for the $\delta^{th}$ quantile of $F_k$ is

$$\hat{\omega}_{k,\delta} = \min\{x_{obs} : \hat{F}_k(x_{obs}) \geq \delta\}.$$

In the scenario when we have all the data present in one centralized location, the computation of $\hat{\omega}_{k,\delta}$ is relatively straight forward. A simple approach would be to iterate over the data, point by point, updating the quantile estimate when we find a point whose cumulative probability is closer to the quantile of interest. However when the data is distributed and stored across multiple machines, we do not have such a straightforward computation since we cannot iterate over all the data in one pass. The solution we will propose is to first compute the estimated quantile based on the data present on each local machine. This means we first find local estimates $\hat{\omega}_{k,\delta}^{(r)} = \min\{x_r : \hat{F}_k(x_r) \geq \delta\}$, $r = 0, ..., m$, where $x_r$ is a data point from the $r^{th}$ sample. We then obtain the final quantile estimate by taking the minimum of the $m+1$ local estimates, i.e. $\hat{\omega}_{k,\delta} = \min\{\hat{\omega}_{k,\delta}^{(r)} : r = 0, ..., m\}$. A distributed algorithm for estimating the $\delta^{th}$ quantile of the $k^{th}$ distribution under the distributed data setting is then given by

**Algorithm**

Broadcast $\delta$ to each of the $m+1$ local machines. (4.16)

Compute $\hat{\omega}_{k,\delta}^{(r)} = \min\{x_r : \hat{F}_k(x_r) \geq \delta\}$ , done in paralell for $r = 0, ..., m$. (4.17)

Transfer the values of $\hat{\omega}_{k,\delta}^{(r)}$ to the global machine. (4.18)

Compute $\hat{\omega}_{k,\delta} = \min(\hat{\omega}_{k,\delta}^{(r)} : r = 0, ..., m)$ on the global machine. (4.19)

A visual representation of this algorithm is shown in figure 4.2.

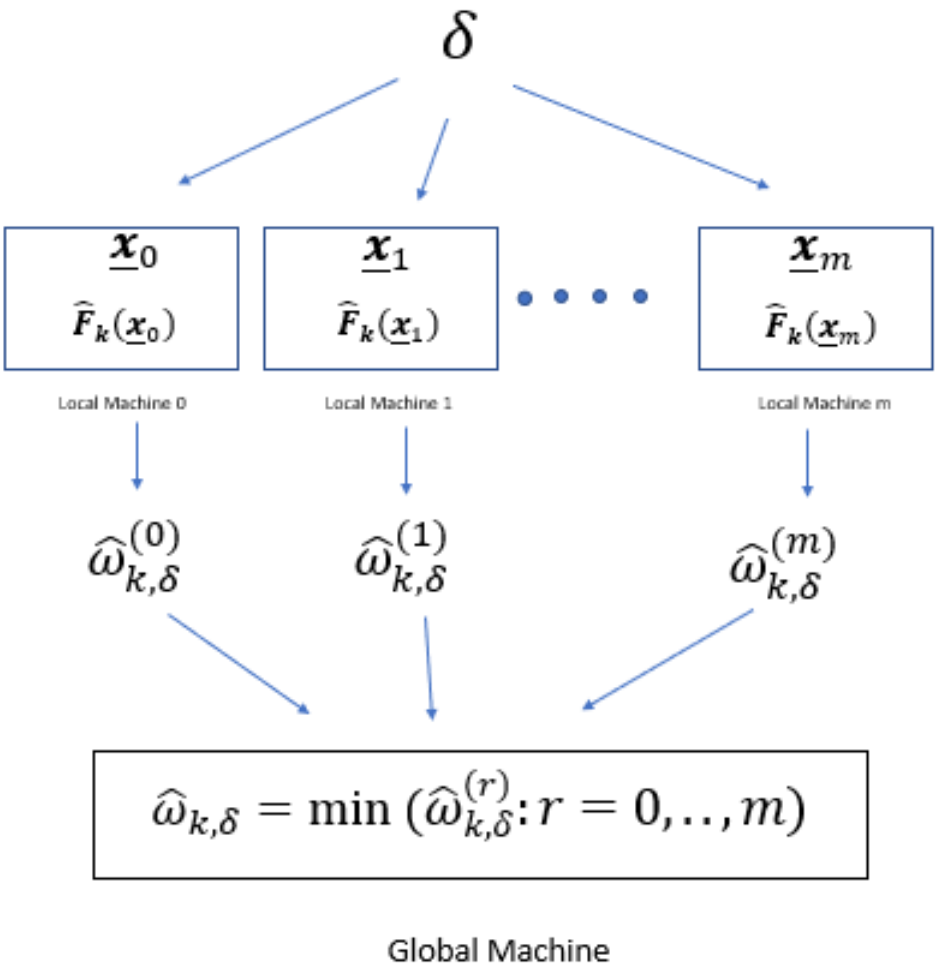


Figure 4.2: A visual depiction of an algorithm for estimating the $\delta^{th}$ quantile of $F_k$ using distributed data

## 4.5 DELRT for testing composite hypothesis about $\boldsymbol{\theta}$ using distributed data

As previously discussed, the DELRT can be employed to carry out various hypothesis tests regarding the model parameters of the DRM. As discussed in section 2.6, one important test of interest is to detect distribution differences amongst the $f_k$ in the DRM. This is done by testing the equality of DRM parameters, i.e. $\boldsymbol{\theta}_k = \boldsymbol{\theta}_j$, for some $k \neq j$.

For illustration purposes, let us continue with the same hypothesis of interest given in equation (2.10), testing the equality of the first and second distributions in the DRM. The relevant hypothesis is given by

$$H_0 : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 \text{ vs } H_a : \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2.$$

or equivalently

$$H_0 : \boldsymbol{g}(\boldsymbol{\beta}) = \mathbf{0} \text{ vs } H_a : \boldsymbol{g}(\boldsymbol{\beta}) \neq \mathbf{0}$$

with

$$\boldsymbol{g}(\boldsymbol{\beta}) = A\boldsymbol{\beta} = \begin{bmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} \boldsymbol{\beta}_1 \\ \boldsymbol{\beta}_2 \end{bmatrix} = \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2.$$

In order to compute the DELR test-statistic, $\tilde{\boldsymbol{\theta}}$, the MELE of $\boldsymbol{\theta}$ subject to the null constraint must first be computed. Using the re-expression of $-l_n(\boldsymbol{\theta})$ shown in (4.1), we must solve the following constrained convex optimization given by

$$\tilde{\boldsymbol{\theta}} = \text{argmin}_{\boldsymbol{\theta}} \sum_{k=0}^{m} g_k(\boldsymbol{\theta}_k)$$

$$\text{subject to } \boldsymbol{\theta}_k - \boldsymbol{z} = \mathbf{0} \text{ , } k = 0, ..., 1 \text{ and } \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2.$$

again rewriting $-l_n(\boldsymbol{\theta})$ as its separable sum, $\sum_{k=0}^{m} g_k(\boldsymbol{\theta})$. Computing $\tilde{\boldsymbol{\theta}}$ requires us to solve the above global variable consensus problem with an additional constraint on $\boldsymbol{\theta}$, that is $\boldsymbol{\beta_1} = \boldsymbol{\beta}_2$. In order to handle the additional constraint placed upon $\boldsymbol{\theta}$, we will employ the global variable consensus with regularization. As shown in (3.25), we can re-write the above problem as

$$\text{minimize } \sum_{k=0}^{m} g_k(\boldsymbol{\theta}_k) + \Omega(\boldsymbol{z}), \tag{4.20}$$

$$\text{subject to } \boldsymbol{\theta}_k - \boldsymbol{z} = \mathbf{0}, k = 0, ..., 1 \tag{4.21}$$

with local variables $\boldsymbol{\theta}_k$, global variable $\boldsymbol{z}$ and regularization term $\Omega(\boldsymbol{z})$ given by

$$\Omega(\boldsymbol{z}) = \lambda \left\| B\boldsymbol{z} \right\|_2 = \lambda \left\| \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 & 0 & -1 \end{bmatrix} \begin{bmatrix} \boldsymbol{\alpha} \\ \boldsymbol{\beta} \end{bmatrix} \right\|_2 = \lambda \left\| \begin{bmatrix} \alpha_1 - \alpha_2 \\ \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 \end{bmatrix} \right\|_2,$$

where $\lambda$ is a positive tuning parameter. The additional term $\Omega(\boldsymbol{z})$ added to the objective function encodes the constraint given by the null hypothesis, $H_0 : \boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$. When the constraint is violated it is clear that $\Omega(\boldsymbol{z}) > 0$ and when it is met, we have $\Omega(\boldsymbol{z}) = 0$, leaving us with the original objective. Although we test hypotheses involving strictly $\boldsymbol{\beta}$ (since $\boldsymbol{\alpha}$ is a function of $\boldsymbol{\beta}$) we include both $\boldsymbol{\alpha}$ and $\boldsymbol{\beta}$ in $\Omega(\boldsymbol{z})$ to help speed up convergence.

Solving for $\tilde{\boldsymbol{\theta}}$ using global variable consensus with regularization, we obtain the following iterations:

**Algorithm**

$$\boldsymbol{\theta}_{\boldsymbol{k}}^{(i+1)} := \text{argmin}_{\boldsymbol{\theta_k}} \Big( g_k(\boldsymbol{\theta}_k) + \boldsymbol{y}_{\boldsymbol{k}}^{iT}(\boldsymbol{\theta_k} - \boldsymbol{z}^{\boldsymbol{i}}) + (\frac{p}{2}) \left\| \boldsymbol{\theta_k} - \boldsymbol{z}^{\boldsymbol{i}} \right\|_2^2 \Big), \tag{4.22}$$

$$\boldsymbol{z}^{(i+1)} := \text{argmin}_{\boldsymbol{z}} \Big( \Omega(\boldsymbol{z}) + \sum_{k=0}^{m} [-\boldsymbol{y}_{\boldsymbol{k}}^{iT}\boldsymbol{z} + \frac{p}{2} \left\| \boldsymbol{\theta}_{\boldsymbol{k}}^{(i+1)} - \boldsymbol{z} \right\|_2^2] \Big), \tag{4.23}$$

$$\boldsymbol{y}_{\boldsymbol{k}}^{(i+1)} := \boldsymbol{y}_{\boldsymbol{k}}^{i} + p(\boldsymbol{\theta}_{\boldsymbol{k}}^{(i+1)} - \boldsymbol{z}^{(i+1)}). \tag{4.24}$$

again, the $\boldsymbol{\theta}_k$ are updated in paralell for $k = 0, ..., m$. The squared norms for the primal and dual residuals are

$$\left\|\boldsymbol{r}^i\right\|_2^2 = \sum_{k=0}^{m} \left\|\boldsymbol{\theta}_{\boldsymbol{k}}^i - \boldsymbol{z}^i\right\|_2^2 \text{ and } \left\|\boldsymbol{s}^i\right\|_2^2 = (m+1)p^2 \left\|\boldsymbol{z}^i - \boldsymbol{z}^{(i-1)}\right\|_2^2.$$

The algorithm can be terminated when

$$\sqrt{\sum_{k=0}^{m} \|\boldsymbol{\theta}_{\boldsymbol{k}}^i - \boldsymbol{z}^i\|_2^2} < \epsilon^{pri} \text{ and } \sqrt{(m+1)p^2 \left\|\boldsymbol{z}^i - \boldsymbol{z}^{(i-1)}\right\|_2^2} < \epsilon^{dual}$$

for some small $\epsilon^{pri}$ and $\epsilon^{dual}$, chosen by the user. After convergence of the algorithm, we use the newly computed $\tilde{\boldsymbol{\theta}}$ to calculate the DELR test-statistic, $R_n = 2\Big(l_n(\hat{\boldsymbol{\theta}} - l_n(\tilde{\boldsymbol{\theta}})\Big)$, which has a $\chi^2_{(2)}$ distribution under the above null hypothesis.

This concludes the presentation of methods to carry out semi-parametric empirical inference with the DRM under the distributed data setting. In the next chapter we will focus on a numerical example in which will carry out the methods described in this chapter to a set of distributed independent samples.

# Chapter 5

# Numerical Example

In this chapter we present the results of a numerical example, using R version 3.6.1, which illustrates the distributed approach to DEL inference under the DRM. A collection of $m+1$ independent random samples were selected from the family of normal distributions. The $k^{th}$ sample was drawn from a normal distribution with mean $\mu_k$ and variance $\sigma_k^2$. As studied in section 1.2.2, the above collection of samples satisfy the DRM model assumptions, rendering the model appropriate in this setting. We additionally assume the samples are stored on seperate local machines, with communication being enabled through a central unit. Furthermore, we assume only limited data transfer between the local and global machines. To achieve this, the samples were stored in independent vectors with iteration over the vectors not being possible, i.e. we can only access one sample/ vector at a time.

All of the methods developed in Chapter 4 for distributed DEL inference will be implemented using the simulated data. The R package *drmdel* was used to validate the inferences obtained. This package, written by Song Cai, has many functionalities with respect to DRMs. Specifically, we will be using the package to compute the MELE of the DRM parameter, perform the DELR test, estimate population distribution functions, estimate the quantiles of the population distributions, compare quantiles from different distributions, and estimate densities of different populations amongst other uses. The package and its manual can be download from The Comprehensive R Archive Network (CRAN) at http://cran.r-project.org/web/packages/drmdel/index.html. The inference obtained using the distributed data and methods will be compared to the inferences generated from

*drmdel* using the same data except stored in one central location/vector. Throughout this chapter, we will commonly refer to these two approaches as the distributed and non-distributed approach. Although we only present one set of results, the example was run many times with different sample seeds and similar successful results were achieved.

In addition to the numerical example we also present a short Monte Carlo simulation used to investigate how to optimally choose the value of the ADMM penalty parameter $p$, in order to minimize the number of iterations needed for convergence.

### 5.0.1 Samples

In this example, we selected $m+1=5$ samples of size $n_k = 75$ , $k = 0, ..., m$, from the family of normal distributions with the $k^{th}$ sample being drawn as follows:

Sample 0 was drawn from a Normal(0,1). (Baseline Sample)

Sample 1 was drawn from a Normal(1,$2^2$).

Sample 2 was drawn from a Normal($2, 2^2$).

Sample 3 was drawn from a Normal($5, 4^2$).

Sample 4 was drawn from a Normal($8, 3^2$).

We will now move on to illustrating the first component, and a necessary step of inference, computing the MELE for the DRM parameter.

## 5.1 Fitting the DRM to observed distributed samples using global variable consensus

With the number of observed samples being 5, the paramater vector for the model is

$$\boldsymbol{\theta} = \begin{bmatrix} \boldsymbol{\alpha} & \boldsymbol{\beta}^T \end{bmatrix}^T = \begin{bmatrix} \alpha_1 & \alpha_2 & \alpha_3 & \alpha_4 & \boldsymbol{\beta_1}^T & \boldsymbol{\beta_2}^T & \boldsymbol{\beta_3}^T & \boldsymbol{\beta_4}^T \end{bmatrix}^T,$$

where $\alpha_k \in \mathbb{R}$, $\boldsymbol{\beta}_k \in \mathbb{R}^2$, $k = 1, 2, 3, 4$. With the true underlying distribution of each sample being known, the true value of $\boldsymbol{\theta}$ is

$$\boldsymbol{\alpha} = \begin{bmatrix} -0.818 & -1.193 & -2.168 & -4.654 \end{bmatrix}^T,$$

$$\boldsymbol{\beta} = \begin{bmatrix} 0.250 & 0.375 & 0.500 & 0.375 & 0.313 & 0.469 & 0.889 & 0.444 \end{bmatrix}^T.$$

### 5.1.1 Initialization values, penalty parameter and stopping criteria

Implementing the global variable consensus, we must first decide on initial values of $\boldsymbol{y}_k$, $\boldsymbol{\theta}_k$, and $\boldsymbol{z}$ used in the distributed algorithm. The algorithm was initialized with

$$\boldsymbol{y}_k^1 = \boldsymbol{\theta}_k^1 = \boldsymbol{z}^1 = \mathbf{0}_{12},$$

where $\mathbf{0}_{12}$ is a zero vector of length 12.

To assist in convergence speed and robustness, the varying penalty parameter scheme given in section 3.2.4 was implemented in the algorithm. The choices of $p^1$, $u$, and $\tau^i$ used were

$$p^1 = 1, \ u = 10, \ \text{and } \tau^i = 1, \ \text{for all iterations } i.$$

During the local variable update step, we are required to solve an unconstrained minimization of the augmented Lagrangian. For this particular problem, the Brayden-Fletcher-Goldfarb-Shanno (BFGS) algorithm was applied using the R function *optim* to carry out the optimization. Initialization values for these problems must be determined at each iteration of the global consensus. For the minimization problem given in the update of $\boldsymbol{\theta}_k^i$, the value of $\boldsymbol{\theta}_k^{i-1}$ was used for initialization, i.e. the value of the local variable at the previous iteration.

The algorithm was terminated when

$$\sqrt{\sum_{k=0}^{m} \|\boldsymbol{\theta}_{\boldsymbol{k}}^{i} - \boldsymbol{z}^{i}\|_2^2} < 0.02 \text{ and } \sqrt{(m+1)p^2 \|\boldsymbol{z}^{i} - \boldsymbol{z}^{(i-1)}\|_2^2} < 0.05.$$

Given the above stopping criteria, the algorithm converged in $i = 1,113$ iterations.

### 5.1.2 Comparing $\hat{\boldsymbol{\theta}}^{\boldsymbol{Dist}}$ to $\hat{\boldsymbol{\theta}}^{\boldsymbol{drmdel}}$

We will refer to the MELE found using the ADMM global variable consensus as $\hat{\boldsymbol{\theta}}^{Dist}$ and the MELE obtained from *drmdel* as $\hat{\boldsymbol{\theta}}^{drmdel}$.

From the data, we obtained

$$\hat{\boldsymbol{\alpha}}^{Dist} = \begin{bmatrix} -0.812 & -1.093 & -1.899 & -5.095 \end{bmatrix}^T,$$
$$\hat{\boldsymbol{\alpha}}^{drmdel} = \begin{bmatrix} -0.810 & -1.093 & -1.905 & -4.943 \end{bmatrix}^T,$$

and

$$\hat{\boldsymbol{\beta}}^{Dist} = \begin{bmatrix} 0.468 & 0.234 & 0.733 & 0.205 & 0.545 & 0.289 & 1.413 & 0.239 \end{bmatrix}^T,$$
$$\hat{\boldsymbol{\beta}}^{drmdel} = \begin{bmatrix} 0.468 & 0.233 & 0.733 & 0.204 & 0.537 & 0.291 & 1.334 & 0.250 \end{bmatrix}^T.$$

As a method of comparison, the Mean Absolute Percent Error (MAPE) of $\hat{\boldsymbol{\theta}}^{Dist}$ relative to $\hat{\boldsymbol{\theta}}^{drmdel}$ was computed. In general the MAPE value is given by

$$M = \frac{1}{n} \sum_{i=1}^{n} \left| \frac{A_i - F_i}{A_i} \right|$$

with $A_i$ being a reference value and $F_i$ being the value to be assessed. For the purpose of our application, $A_i$ will be replaced with the non-distributed parameter estimate of the $i^{th}$ distribution, and $F_i$ will be its distributed counterpart.

For our numerical example, the MAPE equation to compare $\hat{\boldsymbol{\theta}}^{Dist}$ to $\hat{\boldsymbol{\theta}}^{drmdel}$ is given by

$$M_{\hat{\boldsymbol{\theta}}} = \frac{1}{4}\sum_{i=1}^{4}\left|\frac{\hat{\boldsymbol{\theta}}_i^{drmdel} - \hat{\boldsymbol{\theta}}_i^{Dist}}{\hat{\boldsymbol{\theta}}_i^{drmdel}}\right|.$$

with the division being carried out element-wise between the top and bottom vectors. The MAPE of $\hat{\boldsymbol{\theta}}^{Dist}$ relative to $\hat{\boldsymbol{\theta}}^{drmdel}$ is

$$M_{\hat{\boldsymbol{\theta}}} = \begin{bmatrix} 0.009305 & 0.018499 & 0.013939 \end{bmatrix}^T.$$

The MAPE vector, along with the estimates themselves, illustrate that $\hat{\boldsymbol{\theta}}_{Dist}$ and $\hat{\boldsymbol{\theta}}_{drmdel}$ differ by very little. The distributed data and methodology seem to produce nearly identical estimates of $\boldsymbol{\theta}$ as the non-distributed setting does. We do however appear to observe a slightly higher deviation between the distributed MELE for the fourth distribution, and the non-distributed version. We can see this by looking at the contribution to the MAPE of $\hat{\boldsymbol{\theta}}_4^{Dist}$ and $\hat{\boldsymbol{\theta}}_4^{drmdel}$ :

$$M_{\hat{\boldsymbol{\theta}}_4} = \left|\frac{\hat{\boldsymbol{\theta}}_4^{drmdel} - \hat{\boldsymbol{\theta}}_4^{Dist}}{\hat{\boldsymbol{\theta}}_4^{drmdel}}\right| = \begin{bmatrix} 0.030864 & 0.058170 & 0.041953 \end{bmatrix}^T,$$

which is high relative to the overall MAPE but still low considering. We will continue to keep this in mind when evaluating distributed estimates for components of the fourth distribution.

To further evaluate the performance of the ADMM based approach, we will now compare the likelihood function evaluated at both $\hat{\boldsymbol{\theta}}_{Dist}$ and $\hat{\boldsymbol{\theta}}_{drmdel}$. This will allow us to compare the optimal value of the DEL found using the distributed approach to that of the non-distributed . The results are:

$$l_n(\hat{\boldsymbol{\theta}}^{Dist}) = -164.3493,\ l_n(\hat{\boldsymbol{\theta}}^{drmdel}) = -165.3607 \text{ and } M_{l_n} = 0.006116326.$$

Where $M_{l_n}$ is the MAPE of $l_n(\hat{\boldsymbol{\theta}}^{Dist})$ relative to $l_n(\hat{\boldsymbol{\theta}}^{drmdel})$. We can see that the optimal value of the DEL function found via global consensus is extremely close to

the *drmdel* value. The distributed optimal value differs from the non-distributed version by less than 1 percent. Overall, the distributed method for finding the MELE arrives to similar optimal values and points as those produced in the non-distributed *drmdel* approach.

Now that the DRM has been fit to the data, we will continue with estimating components of the baseline distribution, $p_{kj}$ and $F_0(x)$.

## 5.2 Estimating $p_{kj}$ and $F_0(x)$ for the baseline distribution

We will now use the value of $\hat{\boldsymbol{\theta}}^{Dist}$ found above to compute the estimates of $p_{kj}$ and $F_0(x)$, applying the distributed methodology discussed in section 4.2. We will again use the MAPE of the distributed estimates relative to the non-distributed estimates as a means of comparison. We will also present a plot of the distributed estimates against their non distributed counterpart for all $k, j$ to illustrate visually the similarities between the two methodologies.

### 5.2.1 Comparing $\hat{\boldsymbol{p}}_{\boldsymbol{kj}}^{\mathbf{Dist}}$ to $\hat{\boldsymbol{p}}_{\boldsymbol{kj}}^{\boldsymbol{drmdel}}$

The MAPE of $\hat{p}_{kj}^{Dist}$ relative to $\hat{p}_{kj}^{drmdel}$ at all the $x_{kj}$ is

$$M_{\hat{p}_{kj}} = \frac{1}{375}\sum_{k=0}^{4}\sum_{j=1}^{n_k}\left|\frac{\hat{p}_{kj}^{Dist} - \hat{p}_{kj}^{drmdel}}{\hat{p}_{kj}^{drmdel}}\right| = 0.070866.$$

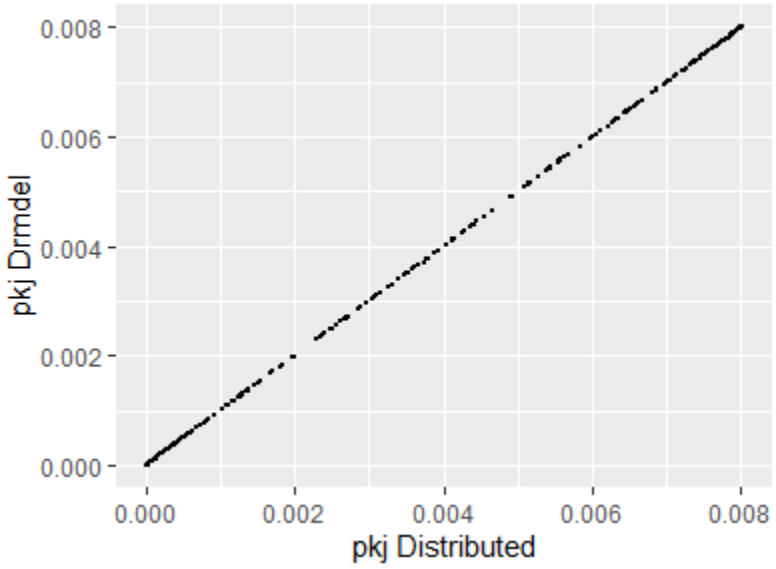


Figure 5.1: Distributed estimates of $p_{kj}$ plotted against drmdel counterpart.

The above plot shows the values of $\hat{p}_{kj}^{Dist}$ plotted against $\hat{p}_{kj}^{drmdel}$ calculated at the same value of $x_{kj}$. The points on the plot create a near perfect linear line of the functional form $y = x$, indicating small differences between the distributed and non-distributed estimates. Overall, the MAPE value obtained is low however we should exercise caution when interpreting. There were some estimates whose contribution to the MAPE seemd to "inflate" the overall measure. This happened when the value of the estimate themselves were extremely small, hence differencess between $\hat{p}_{kj}^{Dist}$ and $\hat{p}_{kj}^{drmdel}$ relative to $\hat{p}_{kj}^{drmdel}$ were magnified. The table below shows the 5 highest contributors to the MAPE of $\hat{p}_{kj}^{Dist}$ relative to $\hat{p}_{kj}^{drmdel}$.

Table 5.1: 5 Largest contributors to the MAPE of $\hat{\boldsymbol{p}}_{\boldsymbol{kj}}^{\boldsymbol{Dist}}$ relative to $\hat{\boldsymbol{p}}_{\boldsymbol{kj}}^{\boldsymbol{drmdel}}$

| $\boldsymbol{k}$ | $\boldsymbol{j}$ | $\boldsymbol{x_{kj}}$ | $\hat{\boldsymbol{p}}_{\boldsymbol{kj}}^{\boldsymbol{Dist}}$ | $\hat{\boldsymbol{p}}_{\boldsymbol{kj}}^{\boldsymbol{drmdel}}$ | **MAPE Contribution** |
|---|---|---|---|---|---|
| 4 | 39 | 14.98458 | 1.27E-33 | 6.60E-34 | 0.924242 |
| 3 | 19 | 13.79946 | 3.90E-29 | 2.08E-29 | 0.875 |
| 3 | 27 | 13.449 | 7.03E-28 | 3.80E-28 | 0.85 |
| 4 | 73 | 13.12532 | 9.52E-27 | 5.25E-27 | 0.813333 |
| 4 | 70 | 13.0733 | 1.44E-26 | 7.96E-27 | 0.809045 |

Although the MAPE contribution produced at these $x_{kj}$ are quite high, the estimates themselves and their relative differences are not meaningful since their values are at round-off level. To adjust for this, the MAPE was re-computed rounding the $\hat{p}_{kj}^{Dist}$ and $\hat{p}_{kj}^{Dist}$ to seven decimal places. This rounding adjustment was applied to all the MAPE calculations in the subsequent sections. After rounding, the MAPE was re-calculated to be

$$M_{\hat{p}_{kj}}^{round} = 0.001153.$$

Looking at both the MAPE value and pairwise plot of $\hat{p}_{kj}^{Dist}$ against $\hat{p}_{kj}^{drmdel}$, we see the distributed methodology and *drmdel* produce estimates which are identical. Using the values of $\hat{p}_{kj}^{Dist}$, we will now estimate the baseline distribution function and compare the results to the non-distributed values.

### 5.2.2 Comparing $\hat{\boldsymbol{F}}_{\boldsymbol{0}}(\boldsymbol{x_{kj}})^{\boldsymbol{Dist}}$ to $\hat{\boldsymbol{F}}_{\boldsymbol{0}}(\boldsymbol{x_{kj}})^{\boldsymbol{drmdel}}$

Below we find a plot of $F_0(x_{kj})^{Dist}$ plotted against $F_0(x_{kj})^{drmdel}$, both calculated at the same $x_{kj}$ values.

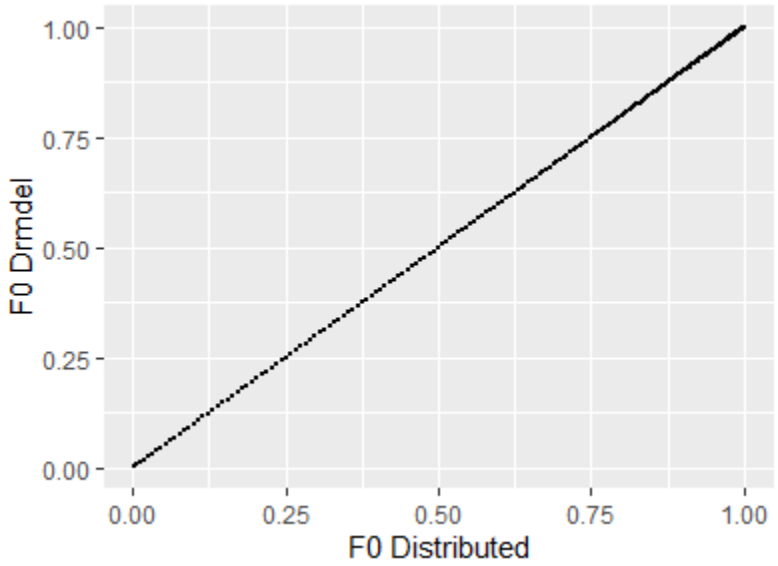


Figure 5.2: Distributed estimates of $F_0(x_{kj})$ plotted against drmdel counterpart.

The MAPE of $\hat{F}_0(x_{kj})^{Dist}$ relative to $\hat{F}_0(x_{kj})^{drmdel}$ is

$$M_{\hat{F}_0(x)} = \frac{1}{375}\sum_{k=0}^{4}\sum_{j=1}^{n_k}\left|\frac{\hat{F}_0(x_{kj})^{Dist} - \hat{F}_0(x_{kj})^{drmdel}}{\hat{F}_0(x_{kj})^{drmdel}}\right| = 0.000885467.$$

Looking at the MAPE value and the plot of $F_0(x_{kj})^{Dist}$ against $F_0(x_{kj})^{drmdel}$, we can see that the distributed methodology and *drmdel* produce identical estimates of the baseline distribution function.

Under the distributed data setting, using the methodologies presented in section 4.2, we were able to achieve identical results when estimating $p_{kj}$ and $F_0(x)$ compared to those obtained from *drmdel* in a non-distributed setting. Next we will estimate the other 4 non-baseline distribution functions and again compare the results to the non-distributed approach.

## 5.3 Estimating the non-baseline distribution functions, $\hat{F}_1(x), ..., \hat{F}_4(x)$

Following the methods given in section 4.3, we will now estimate the non-baseline distribution functions using the distributed data and compare the results to those obtained from *drmdel*. As a means of comparison, we will continue to use the MAPE of the distributed estimate relative to its non-distributed counterpart as well as a pairwse plot.

The MAPE of $\hat{F}_1(x_{kj})^{Dist}$ relative to $\hat{F}_1(x_{kj})^{drmdel}$ is

$$M_{\hat{F}_1(x)} = \frac{1}{375}\sum_{k=0}^{4}\sum_{j=1}^{n_k}\left|\frac{\hat{F}_1(x_{kj})^{Dist} - \hat{F}_1(x_{kj})^{drmdel}}{\hat{F}_1(x_{kj})^{drmdel}}\right| = 0.001633.$$

Below we find a plot of $F_1(x_{kj})^{Dist}$ plotted against $F_1(x_{kj})^{drmdel}$, both calculated at the same $x_{kj}$ value. Similar plots will be presented when comparing estimates for the other three non-baseline distributions.

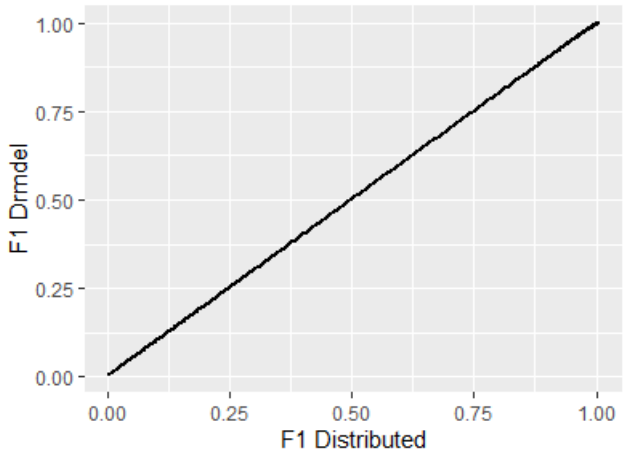


Figure 5.3: Distributed estimates of $F_2(x_{kj})$ plotted against drmdel counterpart.

The MAPE of $\hat{F}_2(x_{kj})^{Dist}$ relative to $\hat{F}_2(x_{kj})^{drmdel}$ is

$$M_{\hat{F}_2(x)} = \frac{1}{375}\sum_{k=0}^{4}\sum_{j=1}^{n_k}\left|\frac{\hat{F}_2(x_{kj})^{Dist} - \hat{F}_2(x_{kj})^{drmdel}}{\hat{F}_2(x_{kj})^{drmdel}}\right| = 0.002443.$$

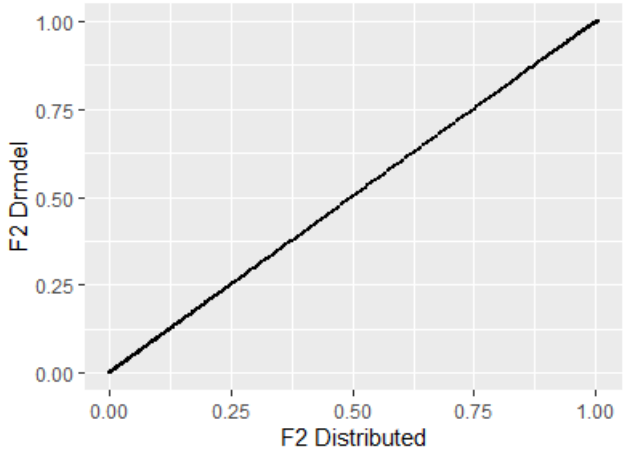


Figure 5.4: Distributed estimates of $F_2(x_{kj})$ plotted against drmdel counterpart.

The MAPE of $\hat{F}_3(x_{kj})^{Dist}$ relative to $\hat{F}_3(x_{kj})^{drmdel}$ is

$$M_{\hat{F}_3(x)} = \frac{1}{375}\sum_{k=0}^{4}\sum_{j=1}^{n_k}\left|\frac{\hat{F}_3(x_{kj})^{Dist} - \hat{F}_3(x_{kj})^{drmdel}}{\hat{F}_3(x_{kj})^{drmdel}}\right| = 0.010610.$$

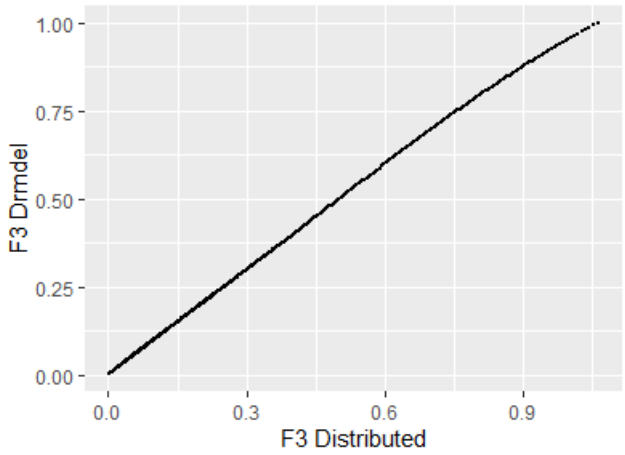


Figure 5.5: Distributed estimates of $F_3(x_{kj})$ plotted against drmdel counterpart.

The MAPE of $\hat{F}_4(x_{kj})^{Dist}$ relative to $\hat{F}_4(x_{kj})^{drmdel}$ is

$$M_{\hat{F}_4(x)} = \frac{1}{375}\sum_{k=0}^{4}\sum_{j=1}^{n_k}\left|\frac{\hat{F}_4(x_{kj})^{Dist} - \hat{F}_4(x_{kj})^{drmdel}}{\hat{F}_4(x_{kj})^{drmdel}}\right| = 0.107315.$$

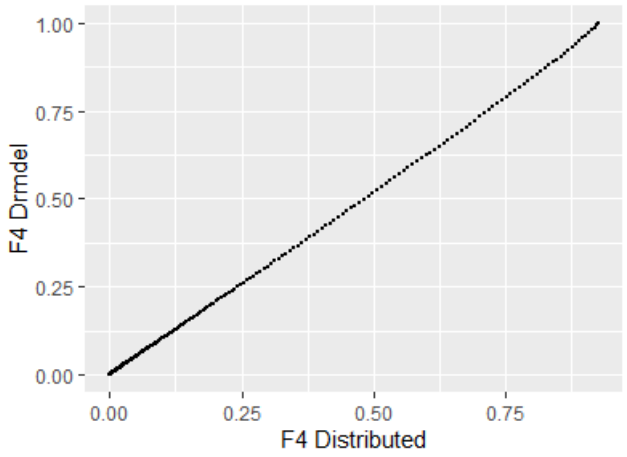


Figure 5.6: Distributed estimates of $F_4(x_{kj})$ plotted against drmdel counterpart.

Averaging the above MAPE values, we obtain an average value of

$$\bar{M}_{\hat{F}_k} = 0.0305.$$

The above numerical example illustrates that the distributed methodologies for estimating $F_k(x_{kj})$ provide identical estimates to *drmdel*. We obtained relatively small MAPE values across each of the $k = 1, 2, 3, 4$ distributions. Furthermore, a scatter plot of $F_k(x_{kj})^{Dist}$ against $F_k(x_kj)^{drmdel}$, $k = 1, 2, 3, 4$ displays a near perfect linear relationship of the form $y = x$. Across all of the samples, we obtain an average MAPE of 0.0305, meaning on average a distributed estimate of a non-baseline distribution function deviated by about 3 percent from its non-distributed counterpart. For the estimates of the first 3 distribution functions, we observed MAPE values of less than or equal to 1.1 percent, however we observe $M_{\hat{F}_4} = 0.10731$ for the estimates of the fourth distribution. Although not significantly large overall, we do observe what appears to be a larger MAPE value in comparison to the values for the first 3 distributions. However this is not concerning since the a large portion of the estimates themselves were again at round-off level, hence creating large relative differences. Adding to this, in 5.12 we noticed that $\boldsymbol{\theta}_4^{Dist}$ differed from its non-distributed counterpart by more in comparison to the other distributions. We can see this difference carrying forward in these results.

## 5.4 Distributed quantile estimation for $F_0, ..., F_4$

We now continue our inference using the DRM by estimating the quantiles of $F_k, k = 0, ..., 4$, applying the distributed methods given in section 4.4. Again, the MAPE was used to compare the distributed and non-distributed estimates.

Here we denote $\hat{\omega}_{k,\delta}^{Dist}$ to be the distributed estimate of the $\delta^{th}$ quantile of distribution $k$, and $\hat{\omega}_{k,\delta}^{drmdel}$ to be its non-distributed counterpart. In this section we will restrict the analysis of $\hat{\omega}_{k,\delta}$ to values of $\delta$ corresponding to $0.1, 0.2, 0.3, ...0.9$. The results for $\hat{\omega}_{k,\delta}^{Dist}$ and $\hat{\omega}_{k,\delta}^{drmdel}$ for $k = 0, ..., 4$ are given in Table 5.2- 5.6 and the MAPE values of $\hat{\omega}_{k,\delta}^{Dist}$ relative to $\hat{\omega}_{k,\delta}^{drmdel}$ $\forall k$ are given in Table 5.7.

Table 5.2: **Comparison of $\hat{\omega}_{0,\delta}^{\mathbf{Dist}}$ and $\hat{\omega}_{0,\delta}^{drmdel}$ for a subset of $\delta$**

| $\delta$ | $\hat{\omega}_{0,\delta}^{\mathbf{Dist}}$ | $\hat{\omega}_{0,\delta}^{drmdel}$ | **MAPE** |
|---|---|---|---|
| 0.1 | -1.713008 | -1.713008 | 0 |
| 0.2 | -1.054161 | -1.054161 | 0 |
| 0.3 | -0.663634 | -0.663634 | 0 |
| 0.4 | -0.345796 | -0.345796 | 0 |
| 0.5 | -0.040169 | -0.040169 | 0 |
| 0.6 | 0.26076 | 0.26076 | 0 |
| 0.7 | 0.571307 | 0.571307 | 0 |
| 0.8 | 0.967355 | 0.967355 | 0 |
| 0.9 | 1.402369 | 1.402369 | 0 |

Table 5.3: **Comparison of $\hat{\omega}_{1,\delta}^{\mathbf{Dist}}$ and $\hat{\omega}_{1,\delta}^{drmdel}$ for a subset of $\delta$**

| $\delta$ | $\hat{\omega}_{1,\delta}^{\mathbf{Dist}}$ | $\hat{\omega}_{1,\delta}^{drmdel}$ | **MAPE** |
|---|---|---|---|
| 0.1 | -0.974372 | -0.974372 | 0 |
| 0.2 | -0.083211 | -0.083211 | 0 |
| 0.3 | 0.552449 | 0.552449 | 0 |
| 0.4 | 1.044497 | 1.044497 | 0 |
| 0.50 | 1.454469 | 1.454469 | 0 |
| 0.6 | 1.921345 | 1.921345 | 0 |
| 0.7 | 2.432355 | 2.432355 | 0 |
| 0.8 | 3.126737 | 3.051784 | 0.02456 |
| 0.9 | 4.300129 | 4.300129 | 0 |

Table 5.4: **Comparison of $\hat{\omega}_{2,\delta}^{\mathbf{Dist}}$ and $\hat{\omega}_{2,\delta}^{drmdel}$ for a subset of $\delta$**

| $\delta$ | $\hat{\omega}_{2,\delta}^{\mathbf{Dist}}$ | $\hat{\omega}_{2,\delta}^{drmdel}$ | **MAPE** |
|---|---|---|---|
| 0.1 | -0.175287 | -0.175287 | 0 |
| 0.2 | 0.6034542 | 0.603454 | 0 |
| 0.3 | 1.1258339 | 1.125834 | 0 |
| 0.4 | 1.5237917 | 1.523792 | 0 |
| 0.5 | 2.0298709 | 2.029871 | 0 |
| 0.6 | 2.4323553 | 2.432355 | 0 |
| 0.7 | 2.9478785 | 2.947879 | 0 |
| 0.8 | 3.6923414 | 3.692341 | 0 |
| 0.9 | 4.692817 | 4.692817 | 0 |

Table 5.5: **Comparison of $\hat{\omega}_{3,\delta}^{\mathbf{Dist}}$ and $\hat{\omega}_{3,\delta}^{drmdel}$ for a subset of $\delta$**

| $\delta$ | $\hat{\omega}_{3,\delta}^{\mathbf{Dist}}$ | $\hat{\omega}_{3,\delta}^{drmdel}$ | **MAPE** |
|---|---|---|---|
| 0.1 | 0.403981 | 0.403981 | 0 |
| 0.2 | 1.58874 | 1.59616 | 0.004649 |
| 0.3 | 2.660567 | 2.660567 | 0 |
| 0.4 | 3.918007 | 3.960301 | 0.01068 |
| 0.5 | 4.835626 | 4.850523 | 0.003071 |
| 0.6 | 5.997467 | 5.997467 | 0 |
| 0.7 | 7.176064 | 7.288239 | 0.015391 |
| 0.8 | 8.542563 | 8.612845 | 0.00816 |
| 0.9 | 9.591606 | 10.0519 | 0.045791 |

Table 5.6: **Comparison of $\hat{\omega}_{4,\delta}^{\mathbf{Dist}}$ and $\hat{\omega}_{4,\delta}^{drmdel}$ for a subset of $\delta$**

| $\delta$ | $\hat{\omega}_{4,\delta}^{\mathbf{Dist}}$ | $\hat{\omega}_{4,\delta}^{drmdel}$ | **MAPE** |
|---|---|---|---|
| 0.1 | 4.24018 | 4.232945 | 0.001709 |
| 0.2 | 5.41964 | 5.18674 | 0.044903 |
| 0.3 | 6.533286 | 6.502423 | 0.004746 |
| 0.4 | 7.288239 | 7.176064 | 0.015632 |
| 0.5 | 8.330926 | 7.939727 | 0.049271 |
| 0.6 | 8.664601 | 8.596094 | 0.00797 |
| 0.7 | 9.318201 | 9.20588 | 0.012201 |
| 0.8 | 10.78952 | 9.755482 | 0.105996 |
| 0.9 | 12.84484 | 11.54632 | 0.112462 |

Table 5.7: **MAPE of $\hat{\omega}_{k,\delta}^{\mathbf{Dist}}$ relative to $\hat{\omega}_{k,\delta}^{drmdel}$ for $k = 0, ..., 4$**

| $k$ | $M_{\hat{\omega}_{4,\delta}}$ |
|---|---|
| 0 | 0 |
| 1 | 0.002729 |
| 2 | 0 |
| 3 | 0.009749 |
| 4 | 0.039432 |

The average MAPE of $\hat{\omega}_{k,\delta}^{Dist}$ relative to $\hat{\omega}_{k,\delta}^{drmdel}$ across the 5 distributions is

$$\bar{M}_{\hat{\omega}_{k,\delta}} = \frac{1}{5}\sum_{k=0}^{4} M_{\hat{\omega}_{k,\delta}} = 0.010871.$$

The tables above show the quantile estimates of the $F_k$ produced from the distributed data and methodology are again nearly identical to those computed from the non-distributed data and approach.

## 5.5 Detecting differences in distributions using the DELR test

We will now employ the application of the DELR test under the circumstances of distributed samples to carry out a hypothesis test regarding the pairwise equality of certain DRM distributions. In the spirit of the lumber application, let us assume Sample 1 and Sample 3 are the lumber strengths of oak and spruce at year 1 and Sample 2 and Sample 4 are the strength of oak and spruce at year 2, respectively.

To asses if the lumber quality in terms of strength have changed between year 1 and year 2, we can carry out the DELR test with the following null and alternate hypotheses.

$$\boldsymbol{H_0} : \boldsymbol{\beta}_1 = \boldsymbol{\beta}_2 \text{ and } \boldsymbol{\beta}_3 = \boldsymbol{\beta}_4 \text{ vs } \boldsymbol{H_a} : \boldsymbol{\beta}_1 \neq \boldsymbol{\beta}_2 \text{ or } \boldsymbol{\beta}_3 \neq \boldsymbol{\beta}_4$$

To compute the DELR test statistic and carry out the test using the distributed data, we will implement the global variable consensus with regularization and the procedures given in section 4.5. We refer to $\tilde{\boldsymbol{\theta}}^{Dist}$ as the value of $\boldsymbol{\theta}$ which maximizes $l_n(\boldsymbol{\theta})$ subject to the constraint under $H_0$. The non-distributed counter part will be referred to as $\tilde{\boldsymbol{\theta}}^{drmdel}$.

### 5.5.1 Initialization values, penalty parameter, stopping criteria and regularizer

The initial values used for the local variables, lagrange multipliers, and global variable were set to

$$\boldsymbol{\theta}_k^1 = \boldsymbol{y}_k^1 = \boldsymbol{z}^1 = \mathbf{0}_{12} \text{ for } k = 0, ..., 4,$$

with $\mathbf{0}_{12}$ being the zero vector of length 12. To maintain simplicity, a constant penalty parameter was used instead of varying the value at each iteration. The fixed penalty parameter $p = 10$ was used at each iteration.

To initialize the minimization problems for the local variable update steps, the

value of the local variable at the previous iteration was used. The algorithm was terminated when

$$\sqrt{\sum_{k=0}^{m} \|\boldsymbol{\theta}_{\boldsymbol{k}}^{i} - \boldsymbol{z}^{i}\|_2^2} < 0.05 \text{ and } \sqrt{(m+1)p^2 \|\boldsymbol{z}^{i} - \boldsymbol{z}^{(i-1)}\|_2^2} < 0.035.$$

Given the above stopping criteria, the algorithm converged in $i = 734$ iterations.

Referring back to global variable consensus with regularization in section 4.5, a regularization term was added, $\Omega(\boldsymbol{\theta})$, to encode the constraint given in the null hypothesis. For the null hypothesis given above, the corresponding regularizer is

$$\Omega(\boldsymbol{\theta}) = \lambda \|B\boldsymbol{\theta}\|_2 = 15 \left\| \begin{bmatrix} \alpha_1 - \alpha_2 \\ \alpha_3 - \alpha_4 \\ \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 \\ \boldsymbol{\beta}_3 - \boldsymbol{\beta}_4 \end{bmatrix} \right\|_2 \text{ with tuning parameter } \lambda.$$

The matrix B is given by

$$B = \begin{bmatrix} 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & -1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 & -1 \end{bmatrix}.$$

This regularization term encodes the constraint $F_1 = F_2$ and $F_3 = F_4$.

### 5.5.2 Comparison of $\tilde{\boldsymbol{\theta}}^{\boldsymbol{Dist}}$ to $\tilde{\boldsymbol{\theta}}^{\boldsymbol{drmdel}}$

We will now present and compare the values of $\tilde{\boldsymbol{\theta}}^{Dist}$ to it's non-distributed counterpart, $\tilde{\boldsymbol{\theta}}^{drmdel}$. The MAPE of $\tilde{\boldsymbol{\theta}}^{dist}$ relative to $\tilde{\boldsymbol{\theta}}^{drmdel}$ will again be computed for comparison.

The computed parameter estimates are

$$\tilde{\boldsymbol{\alpha}}^{Dist} = \begin{bmatrix} -0.9466 & -0.9466 & -2.5253 & -2.5253 \end{bmatrix}^T,$$
$$\tilde{\boldsymbol{\alpha}}^{drmdel} = \begin{bmatrix} -0.9363 & -0.9364 & -2.5733 & -2.5734 \end{bmatrix}^T,$$
$$\tilde{\boldsymbol{\beta}}^{Dist} = \begin{bmatrix} 0.5613 & 0.2310 & 0.5613 & 0.2310 & 0.6277 & 0.2915 & 0.6277 & 0.2915 \end{bmatrix}^T,$$
$$\tilde{\boldsymbol{\beta}}^{drmdel} = \begin{bmatrix} 0.5726 & 0.2282 & 0.5900 & 0.3089 & 0.5726 & 0.22827 & 0.5900 & 0.3089 \end{bmatrix}^T.$$

The MAPE vector obtained was

$$M_{\tilde{\boldsymbol{\theta}}} = \frac{1}{4}\sum_{k=1}^{4}\left|\frac{\tilde{\boldsymbol{\theta}}_k^{Dist} - \tilde{\boldsymbol{\theta}}_k^{drmdel}}{\tilde{\boldsymbol{\theta}}_k^{drmdel}}\right| = \begin{bmatrix} 0.014793 & 0.041897 & 0.034493 \end{bmatrix}^T.$$

Furthermore, we can compare the optimal values of the DEL function at both $\tilde{\boldsymbol{\theta}}^{Dist}$ and $\tilde{\boldsymbol{\theta}}^{drmdel}$:

$$l_n(\tilde{\boldsymbol{\theta}}^{Dist}) = 147.2706 \text{ and } l_n(\tilde{\boldsymbol{\theta}}^{drmdel}) = 148.2719,$$

with the MAPE of $l_n(\tilde{\boldsymbol{\theta}}^{Dist})$ relative to $l_n(\tilde{\boldsymbol{\theta}}^{drmdel})$ given by

$$M_{l_n} = 0.006753.$$

The MAPE vector $M_{\tilde{\boldsymbol{\theta}}}$ shows us that the distributed and non-distributed estimates of $\tilde{\boldsymbol{\theta}}$ are very close to one another. Furthermore, inspecting the optimal values of both the points $\tilde{\boldsymbol{\theta}}^{Dist}$ and $\tilde{\boldsymbol{\theta}}^{drmdel}$, we see they are nearly identical, differing from the non-distributed value by less than one percent.

Next, we carry out the remainder of the DELRT test. The test statistic(s) will be computed using $\tilde{\boldsymbol{\theta}}^{Dist}$ and $\tilde{\boldsymbol{\theta}}^{drmdel}$, and a conclusion will be made by comparing the test-statistic to a critical region.

### 5.5.3 Test statistic, rejection region, and conclusion

Recall, the DELR test-statistic is defined as

$$R_n = 2(l_n(\hat{\boldsymbol{\theta}}) - l_n(\tilde{\boldsymbol{\theta}})).$$

For the computation of $R_n^{Dist}$, $\hat{\boldsymbol{\theta}}^{Dist}$ and $\tilde{\boldsymbol{\theta}}^{Dist}$ will be used while $R_n^{drmdel}$ will implement the non-distributed counterparts, $\hat{\boldsymbol{\theta}}^{drmdel}$ and $\tilde{\boldsymbol{\theta}}^{drmdel}$.

The values of the test statistics were found to be

$$R_n^{Dist} = 34.157 \text{ and } R_n^{drmdel} = 34.178,$$

with the MAPE between the two being

$$M_{R_n} = 0.000594.$$

As was discussed in section 2.3.6, the null limiting distribution of $R_n$ is $\chi_q^2$ where $q$ is the dimension of the image of $g(\boldsymbol{\beta})$. For the hypotheses being tested $g(\boldsymbol{\beta})$ is given by

$$g(\boldsymbol{\beta}) = \begin{bmatrix} \boldsymbol{\beta}_1 - \boldsymbol{\beta}_2 \\ \boldsymbol{\beta}_3 - \boldsymbol{\beta}_4 \end{bmatrix}$$

.

Hence $q = 4$ under this hypothesis and $R_n \sim \chi_4^2$. Using a significance level of $\epsilon = 0.01$, the critical region for the test is

$$\chi_{4,0.99}^2 = 13.2767.$$

Since both $R_n^{Dist} > 13.2767$ and $R_n^{drmdel} > 13.2767$, both methodologies conclude at $\epsilon = 0.01$ level of significance that at least one of $F_1 = F_2$ or $F_3 = F_4$ is not true. Since we know the true underlying distributions of the original data, is it true that $F_1 \neq F_2$ and $F_3 \neq F_4$ hence the tests both arrive to the same correct statistical conclusion.

## 5.6 Monte Carlo Simulation

In this section we present the results of a Monte Carlo simulation in an attempt to provide some insight into optimally choosing $p$ with respect to algorithm convergence speed. In each iteration of the simulation we drew two independant samples from two seperate normal distributions. The distribution and sizes of each sample were:

$$\text{Sample}_0 \sim \text{Normal}(0, 1),\ n_0 = 100.$$
$$\text{Sample}_1 \sim \text{Normal}(1, 2^2),\ n_1 = 100.$$

Using the observed data, we then implemented the ADMM global variable consensus to fit the model by computing the MELE. In each iteration of the simulation the value of the ADMM penalty parameter $p$ was altered. The number of iterations needed for convergence was then recorded. The simulation consisted of 15 iterations where the values of $p_i$ at iteration $i$ were

$$\boldsymbol{p} = [p_1, ..., p_{15}] = [0.5, 1, 10, 20, 30, 40, 50, 100, 200, 300, 400, 500, 600, 700, 800].$$

For each iteration, the ADMM was initialized with the initial values

$$\boldsymbol{z}^1 = \boldsymbol{y}_k^1 = \boldsymbol{\theta}_k^1 = \mathbf{0}_3,\ k = 0, 1,$$

and the local variable update optimizations for $\boldsymbol{\theta}_k^{i+1}$ were initilaized with the value of the local variable at the previous iterations, i.e. $\boldsymbol{\theta}_k^i$. The algorithm was terminated when

$$\left\|\boldsymbol{r}^i\right\|_2^2 < 10^{-3} \text{ and } \left\|\boldsymbol{s}^i\right\|_2^2 < 10^{-3}.$$

Below we plot the values of $p^i$ against the number of ADMM iterations needed for convergence (associated with the $p^i$).

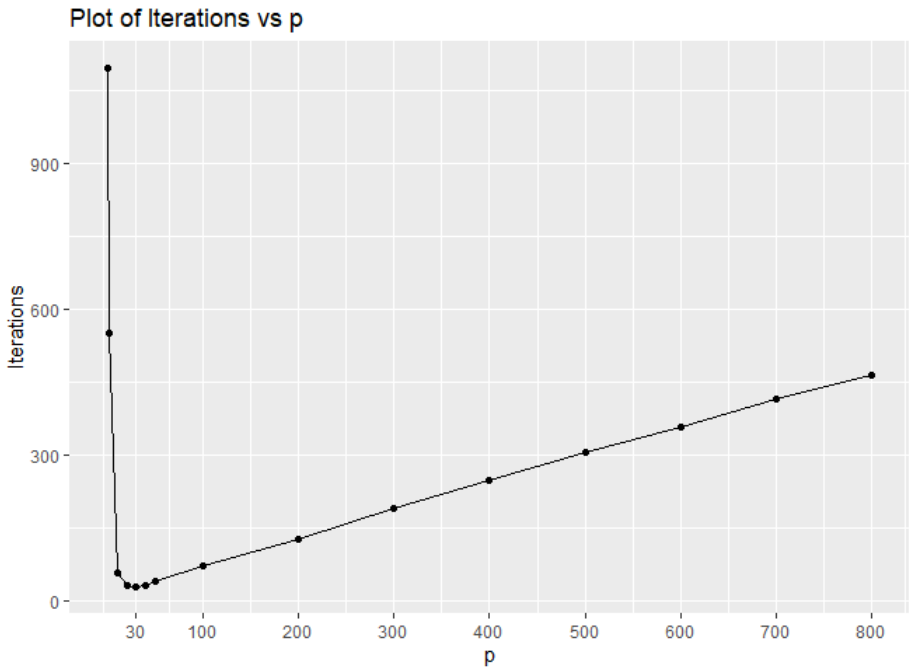


Figure 5.7: A plot of $p^i$ vs number of iterations needed for convergence.

The first choice of $p$ on the plot is $p = 0.5$ which requires the largest number of iterations to converge. We observe a sharp decrease as $p$ approaches $p = 30$, in which case we only require 29 iterations for converge (a local minimum). The number of iterations needed to converge seems to then increase linearly when $p > 30$. Based on this small simulation, there is evidence to support the existance of a relationship between $p$ and convergence speed when considering the ADMM for this problem. It may prove valuable to explore this area further in the future.

# Chapter 6

# Summary and Future Works

## 6.1 Summary of present work

This thesis has presented techniques to fit a semi-parametric DRM to multiple distributed samples, stored across various local machines. Variants of the ADMM algorithm were implemented to find the MELE of the DRM parameter, as well as to compute the DELR test-statistic. Methods to carry out various components of DEL inference using the DRM were developed and presented in the presence of distributed samples. The statistical conclusions of the distributed-data based DEL inference were compared to the same inferences preformed on the same data but in a non-distributed setting. Both sets of inference arrived at the same statistical conclusions illustrating the reliability of the distributed methods. The types of inference for which distributed methodologies were developed were estimating DRM model parameters, estimating the baseline distribution function and probabilities, estimating non-baseline distribution functions, estimating quantiles for each distribution, and carrying out the DELR test to detect distributional differences.

## 6.2 Future works and improvements

### 6.2.1 More rigorous convergence properties

Throughout the research involved in this thesis, it was found that as the number of samples and size of the pooled sample increases, the computation speed of each

ADMM iteration decreases. Thus, a motivation to explore more rigorous convergence properties (of the ADMM with respect to $\hat{\boldsymbol{\theta}}$ and $\tilde{\boldsymbol{\theta}}$) would allow for the possibly to implement the minimum number of ADMM iterations needed to achieve a desired level of precision on our estimate of the DRM parameter. These upgraded convergence properties could serve as a tool for saving computing power through avoiding extra iterations which do not necessarily offer a return in terms of the minimization of the DEL.

### 6.2.2 Other Distributed Algorithms

It would be worthwhile to consider other algorithms to apply for solving the distributed optimization of the DEL under DRM assumptions. To go hand in hand with more rigorous convergence properties, it would be valuable to consider different distributed/ parallel techniques in an attempt to possibly find a more efficient approach.

# Bibliography


Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika*, 59(1):19–35.

Bioucas-Dias, J. M. and Figueiredo, M. A. T. (2010). Alternating direction algorithms for constrained sparse regression: Application to hyperspectral unmixing. In *2010 2nd Workshop on Hyperspectral Image and Signal Processing: Evolution in Remote Sensing*, pages 1–4.

Bottou, L. (1998). Online learning and stochastic approximations. *On-line learning in neural networks*, 17(9):142.

Boyd, S., Parikh, N., Chu, E., Peleato, B., and Eckstein, J. (2011). Distributed optimization and statistical learning via the alternating direction method of multipliers. *Found. Trends Mach. Learn.*, 3(1):1–122.

Cai, S., Jiahua, and Zidek, J. V. (2017). Hypothesis testing in the presence of multiple samples under density ratio models. *Statistica Sinica*, 27(2):761–783.

Chen, J. and Liu, Y. (2013). Quantile and quantile-function estimations under density ratio model. *Ann. Statist.*, 41(3):1669–1692.

Cheng, K. and Chu, C. (2004). Semiparametric density estimation under a two-sample density ratio model. *Bernoulli*, 10(4):583–604.

Eckstein, J. and Bertsekas, D. (1992). On the douglas-rachford splitting method and the proximal point algorithm for maximal monotone operators. *Mathematical Programming*, 55(3):293–318.

Eckstein, J. and Fukushima, M. (1994). Some reformulations and applications of the alternating direction method of multipliers.

Everett III, H. (1963). Generalized lagrange multiplier method for solving problems of optimum allocation of resources. *Operations research*, 11(3):399–417.

Farewell, V. T. (1979). Some results on the estimation of logistic models based on retrospective data. *Biometrika*, 66(1):27–32.

Fokianos, K. (2004). Merging information for semiparametric density estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(4):941–958.

Fokianos, K., Kedem, B., Jing, and Short, D. A. (2001). A semiparametric approach to the one-way layout. *Technometrics*, 43(1):56–65.

Fortin, M. and Glowinski, R. (2000). *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems*. ISSN. Elsevier Science.

Fukushima, M. (1993). Application of the alternating direction method of multipliers to separable convex programming problems. *Comput Optim Applic*, 1:93–111.

Gabay, D. (1983). Augmented lagrangian methods: Applications to the numerical solution of boundary-value problems, vol. 15 of studies in mathematics and its applications.

Gabay, D. and Mercier, B. (1976). A dual algorithm for the solution of nonlinear variational problems via finite element approximation.

Glowinski, R. and Marroco, A. (1975). Sur l'approximation, par éléments finis d'ordre un, et la résolution, par pénalisation-dualité d'une classe de problèmes de dirichlet non linéaires. *ESAIM: Mathematical Modelling and Numerical Analysis - Modélisation Mathématique et Analyse Numérique*, 9(R2):41–76.

He, B. S., Yang, H., and Wang, S. L. (2000). Alternating Direction Method with Self-Adaptive Penalty Parameters for Monotone Variational Inequalities. *Journal of Optimization Theory and Applications*, 106(2):337–356.

Keziou, A. and Leoni-Aubin, S. (2008). On empirical likelihood for semiparametric two-sample density ratio models. *Journal of Statistical Planning and Inference*, 138(4):915 – 928.

Kiwiel, K. C. (2001). Convergence and efficiency of subgradient methods for quasi-convex minimization. *Mathematical programming*, 90(1):1–25.

Mantel, N. (1973). Synthetic retrospective studies and related topics. *Biometrics*, 29(3):479–486.

Prentice, R. L. and Pyke, R. (1979). Logistic disease incidence models and case-control studies. *Biometrika*, 66(3):403–411.

Qin, J. (1993). Empirical likelihood in biased sample problems. *The Annals of Statistics*, 21(3):1182–1196.

Qin, J. (1998). Inferences for case-control and semiparametric two-sample density ratio models. *Biometrika*, 85(3):619–630.

Qin, J. and Zhang, B. (1997). A goodness-of-fit test for logistic regression models based on case-control data. *Biometrika*, 84(3):609–618.

Tseng, P. (1991). Applications of a splitting algorithm to decomposition in convex programming and variational inequalities. *SIAM journal on control and optimization*.

Yuan, X. (2009). Alternating direction methods for sparse covariance selection.

Zhang, B. (2000). Quantile estimation under a two-sample semi-parametric model. *Bernoulli*, 6(3):491–511.

Zhang, Y., Duchi, J., and Wainwright, M. (2013). Divide and conquer kernel ridge regression. In Shalev-Shwartz, S. and Steinwart, I., editors, *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 592–617, Princeton, NJ, USA. PMLR.

Zou, F. and Fine, J. P. (2002). A note on a partial empirical likelihood. *Biometrika*, 89(4):958–961.