Estimating Underlying Distribution using Incomplete Data: Trend of Income Inequality in Australia

by

Toan Vo Khanh Le

Submitted to the Research School of Economics, ANU
in partial fulfillment of the requirements for the degree of

Bachelor of Economics (Honours)

at the

AUSTRALIAN NATIONAL UNIVERSITY

October 2018

Author ....

Research School of Economics, ANU

October 15, 2018

Certified by ....

Fedor Iskhakov

Doctor

Thesis Supervisor

Accepted by ....

Timothy Kam

Chairman, Department Committee on Honours Theses

Estimating Underlying Distribution using Incomplete Data: Trend of Income Inequality in Australia

by

Toan Vo Khanh Le

Submitted to the Research School of Economics, ANU
on October 15, 2018, in partial fulfillment of the
requirements for the degree of
Bachelor of Economics (Honours)

Abstract

In this thesis, I developed a new econometrics framework to estimate the size distribution of income and obtain valid standard errors using grouped data. Data acquired are often available in grouped formats: class frequency data, quantile data, interquantile mean data. These three types of income data are common and more accessible, while microdata are difficult to obtain due to high costs or confidential reasons. When dealing with these data types, inequality literature often utilises imprecise statistical methods for parameter estimation. This is the challenge that I set out to address. Specifically in this thesis, I employ several distributions such as LogNormal, Gamma, Generalised Beta and some mixtures between them for modelling the underlying distribution of income. Under class frequency data, Expectation-Maximisation algorithm provides recursive formulas for maximum likelihood estimators under the assumption that the distribution is from, or can be mixed from exponential family. I then introduce the method of quantile matching, the method of interquantile mean matching, and method of proportion matching as a unified framework to solve the estimation problems involving three data types. Joint estimators are developed to make use of any combination of seemingly different data types for efficiency sake. These new estimators bear a resemblance to their GMM counterpart, from which various consistency and asymptotic normality properties are derived. I prove several efficiency results that demonstrate the power and potential of the new estimation framework. Finally, applications to Australian income data using new estimation techniques are considered.

Estimated Word Count: 19,500

Thesis Supervisor: Fedor Iskhakov
Title: Doctor

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my supervisor Doctor Fedor Iskhakov. His expertise and constructive criticism are an excellent source of ideas from which many improvements and advances are made in my thesis. Without his patience and encouragement, this work could not have been completed. Looking back, I am pleased to have someone so accomplished as the supervisor for my very first research project.

I would also like to acknowledge Doctor Thomas Tao Yang for his insightful discussion on various econometrics topics, Professor Timothy Kam for his support during the transition from my previous university to the Australian National University. I also want to thanks all the teachers from my past institution, the University of Queensland, who kindle my interest in economics. I am grateful to Thao Nguyen, Hien Trang Le, who spent time talking with me and read the draft for comment. Your company made my journey enjoyable and meaningful.

Finally, I want to dedicate this thesis to my parents, my sisters, and my newborn niece. Without their endless support and constant encouragement, I could not have made this far.

Contents

List of Figures

List of Tables

Abbreviations and Acronyms

CDF	Cumulative Distribution Function
CLT	Central Limit Theorem
EM	Expectation - Maximisation
GB	Generalised Beta
GMM	Generalised Method of Moments
iid	independently and identically distributed
IME	Interquantile Mean Matching (Estimation / Estimator)
LLN	Law of Large Number
MLE	Maximum Likelihood (Estimation / Estimator)
MM	Majorise - Minimisation / Minorise - Maximisation
PDF	Probability Density Function
PME	Proportion Matching (Estimation / Estimator)
PMF	Probability Mass Function
QME	Quantile Matching (Estimation / Estimator)
QIME	Joint Quantile, Interquantile Matching (Estimation / Estimator)
QIPME	Joint Quantile, Interquantile, Proportion Matching (Estimation / Estimator)

Mathematical Notation

Mathematical Symbols

$\stackrel{iid}{\sim}$	Are independent and identically distributed as
$\mathbf{x}$, $\boldsymbol{\theta}$	Vector $\mathbf{x}$, parameters vector $\boldsymbol{\theta}$ (always assume $\boldsymbol{\theta} \in \mathbb{R}^d$)
$\|\mathbf{x}\|$	Euclidean norm of vector $\mathbf{x}$
f_x, f_y	Partial derivatives of f with respect to x, y
∇f	Gradient of $f : \mathbb{R}^n \to \mathbb{R}$ (column vector of size n)
	Jacobian of $f : \mathbb{R}^n \to \mathbb{R}^m$ (matrix of size $m \times n$)
$\nabla^2 f$	Hessian of $f : \mathbb{R}^n \to \mathbb{R}$ (matrix of size $n \times n$)
A^{T}, $\mathbf{x}^{\mathrm{T}}$	Transpose of matrix A or vector $\mathbf{x}$
X, Y	Random variables
$\mathbf{X}$, $\mathbf{Y}$	Random vectors
$\hat{\boldsymbol{\theta}}$, $\tilde{\boldsymbol{\theta}}$, $\hat{\theta}$, $\tilde{\theta}$	Estimator / Estimate of parameters vector $\boldsymbol{\theta}$, or scalar θ
o_p, O_p	Little-o, Big-O notation in probability
Q, F^{-1}	Quantile Function
Q^A, Q^{A_h}	Anti-derivatives of quantile function, h^{th} power quantile function
ξ_p, $\hat{\xi}_p$, $\hat{\xi}_{p,n}$	p^{th} quantile, sample p^{th} quantile, or with explicit sample size
$X_{i:n}$	i^{th} order statistic, computed from iid sample of size n
$\mathring{f}$, $\mathring{p}$, $\mathring{F}$	Marginal PDF, PMF, CDF (often being used in MLE)
$\mathring{f}_k$, $\mathring{F}_k$	Conditional PDF, CDF given the random variable is in $[t_{k-1}, t_k)$
$C(x, z)$	Compatibility indicator function for $x = k$, $z \in [t_{k-1}, t_k)$
$\mu(\rho_1, \rho_2)$	Interquantile mean between ρ_1^{th} and ρ_2^{th} quantile
$\hat{\mu}(\rho_1, \rho_2)$	Sample interquantile mean between ρ_1^{th} and ρ_2^{th} quantile
$\mu^{(h)}(\rho_1, \rho_2)$	Interquantile h^{th} moment between ρ_1^{th} and ρ_2^{th} quantile
$\hat{\mu}^{(h)}(\rho_1, \rho_2)$	Sample interquantile h^{th} moment between ρ_1^{th} and ρ_2^{th} quantile
$\sigma^2(\rho_1, \rho_2)$	Interquantile variance between ρ_1^{th} and ρ_2^{th} quantile
L, l	Likelihood function, log-likelihood function
$\odot$	A product involving three-dimensional array
$\xrightarrow{p}$	Convergence in probability
$\xrightarrow{as}$	Almost sure convergence (convergence with probability 1)
$\xrightarrow{d}$	Convergence in distribution

Mathematical Functions

$x \wedge y$	$\min\{x, y\}$
$\lfloor x \rfloor$	Floor function
$\Gamma(a)$	Gamma function
$\mathbb{1}_A(x)$	Characteristic function for set A (1 if $x \in A$, 0 otherwise)
$\Gamma(a, z)$	Incomplete gamma function
$\mathrm{Erf}(x)$	Error function, $\mathrm{Erf}(x) = 1 - \mathrm{Erfc}(x)$
$\mathrm{Erfc}(x)$	Complementary error function
$Q(a, 0, z)$	Regularised incomplete gamma function
${}_pF_q(\mathbf{v}_p; \mathbf{v}_q; z)$	Hypergeometric function, $\mathbf{v}_p$, $\mathbf{v}_q$ are vectors of length p, q
$\psi(a)$	Euler digamma function, $\psi(a) = \Gamma'(a)/\Gamma(a)$
$\nu(a)$	$\ln a - \psi(a)$, decreasing bijective function from $(0, \infty)$ to $(0, \infty)$
$\nu^{-1}(b)$	Inverse of $\nu(\cdot)$, decreasing bijective function
$B(p, q)$	Beta function, $B(p, q) = \Gamma(p)\Gamma(q)/\Gamma(p + q)$
$I_z(p, q)$	Regularised incomplete beta function
$I_s^{-1}(p, q)$	Inverse of the regularised incomplete beta function
$F_1\left(a; b_1, b_2; c; x, y\right)$	Appell hypergeometric function of two variables

Statistical Distributions

$\mathrm{Bern}(p)$	Bernoulli Distribution
$\mathrm{Beta}(p, q)$	Beta Distribution
$\chi^2(n)$	Chi-squared distribution
$\mathrm{Expo}(\lambda)$	Exponential Distribution
$\mathrm{Gamma}(a, \lambda)$	Gamma Distribution
$\mathrm{GB1}(p, q, a, b)$	Generalised Beta of the First kind
$\mathrm{GB2}(p, q, a, b)$	Generalised Beta of the Second kind
$\mathrm{GB}(p, q, a, b, c)$	Generalised Beta
$\mathrm{LN}(\mu, \sigma^2)$	LogNormal Distribution
$\mathrm{Mult}(n, \mathbf{p})$	Multinomial Distribution
$\mathrm{N}(\mu, \sigma^2)$	Normal Distribution
$\mathrm{N}(\boldsymbol{\mu}, \Sigma)$	Multivariate Normal Distribution
$\mathrm{P}(a, z_m)$	Pareto Distribution
$\mathrm{Unif}(a, b)$	Uniform Distribution

Chapter 1

Introduction

1.1 Outline of Thesis

A wide range of economic and social problems are linked to inequality: gender inequality, income and wealth inequality, voting power inequality and so on. When handling these problems, economists and econometricians often want to quantify inequality systematically. This endeavour is a major part of a branch of economics called **distributional analysis** in which the question of how to measure inequality is of great importance. **Parametric solution** to this question generally boils down to three steps: finding a relevant economic variable that is measurable and comparable, discovering the underlying statistical distribution serving as a "blueprint" for the realised data, and computing the inequality indices based on the estimated distribution (Cowell, 2011).

As the objective is investigating income inequality, our interested economic variable throughout the paper is individual earnings. However, the insurmountable difficulty we face is the scarcity of income microdata. Instead of reporting exact individual earnings, income data are usually displayed in grouped forms: class frequency data, quantile data, and interquantile mean data. In **class frequency form**, earnings data are classified into an array of intervals, and only the counts of individuals whose income levels fall within each income range get reported. Each income range will have a *lower class limit*, and *upper class limit*, and *class width* is defined as the distance between lower limits of adjacent classes. **Quantile data** reports sample income quantiles (usually at 10^{th}, 20^{th}, ..., 90^{th} percentiles) based on the empirical distribution of income. Sample median or sample quartiles are notable examples of quantile values: sample 0.5-quantile is the median, sample $0.25,\ 0.5,\ 0.75$ - quantiles are the three quartiles. Standardised test results such as GMAT, GRE and ATAR[1] adopt the quantile framework to report scores, and the use of quantiles in summarising data are growing in popularity. **Interquantile mean data** reports sample truncated means restricted to a number of non-overlapping interquantile ranges (usually divided into quintiles or deciles). As the simplest example, income data are partitioned into two groups equal in size: lower than sample median, higher than sample median. The average income for each group is then computed and displayed. Interquantile mean (also known as *trimmed mean*) report is typical in many income and wealth surveys. The usage of income data in summary forms can be found various times in Piketty (2014, 2015); Alvaredo, Chancel, Piketty, Saez and Zucman (2018).

Motivated by the challenge of distribution recovery, I devised new estimation procedures to handle grouped data with proper theoretical foundation. The new estimation techniques are applications of the **Generalised Method of Moments** and constructed upon asymptotic behaviours of sample quantiles, interquantile means and sample proportions. The estimated income distribution then enables us to perform descriptive analysis, extract inequality measures, and even compute confidence intervals and perform hypothesis testings. In the inequality literature, goodness of fit is traditionally valued more than uncertainty quantification. This paradigm, however, goes against the tradition of econometrics, and speaks to a quote by MIT physicist Walter Lewin: "Any measurement that you make without the knowledge of its uncertainty is completely meaningless". Even though he makes this statement within the context of measurement in physics, I think the general idea should also be extended to any statistics and econometrics estimation framework.

One should be aware that the use of proposed estimation techniques is independent of the variables chosen. The variables could as well be wealth, consumption,

[1]GMAT and GRE are graduate admission tests; ATAR is admission score for tertiary education.

test scores, to name a few, and are not restricted to non-negative value. Therefore, successfully developing these methods is significant for many other reasons apart from investigating income inequality. Throughout the paper, I will define our variable as income but readers should keep in mind that, the variable might be very general.

There are 7 chapters in the paper. Chapter 1 reviews the literature surrounding this area of research, and presents the main structure of the thesis. My main references for this chapter are two textbooks: *Measuring Inequality* (Cowell, 2011) and *Modeling Income Distributions and Lorenz Curves* (Chotikapanich, 2008).

Chapter 2 lays out the required foundation of statistics and inequality measures, which then plays a vital role in developing new estimation techniques in later chapters. Important statistical concepts such as quantile function (possibly the most important), exponential family and generalised beta distribution, and Expectation-Maximisation algorithm are covered. Main references for this part are: *Maximum Likelihood from incomplete data via the EM algorithm* (Dempster, Laird and Rubin, 1977), *Statistical Modeling and Computation* (Kroese, Chan et al., 2014) for EM algorithm and exponential family treatment, (McDonald, 1984; McDonald and Xu, 1995) for the development of generalised beta family, (Cowell, 2011) for inequality charting and measure. Statistics notations and thinking in this paper are influenced by two textbooks written by Blitzstein and Hwang (2014) and Kroese, Chan et al. (2014).

Chapter 3 explores the idea of using EM algorithm to construct maximum likelihood estimators to solve the estimation problem of class frequency data. The requirement for the applicability of EM algorithm is that the underlying distribution is from exponential family or can be represented as a mixture from exponential family. Many special mathematical functions naturally arise in formulas derived from this chapter such as gamma function, beta function, generalised hypergeometric function, appell function and so on. These functions are well-known in mathematical physics and readers are encouraged to consult *Formula and theorems for the special functions of mathematical physics* (Magnus, Oberhettinger and Soni, 2013) for reference. Interactive collection of formulas and graphics to these functions can be explored at *The*

Mathematical Functions Site[2].

Chapter 4 is a major part of this paper, where the construction of new estimators and their corresponding statistical properties are presented. The new estimators, which solve the estimation problems of quantile data and interquantile mean data, repeat arguments from Generalised Method of Moments developed by Hansen (1982); Hansen and Singleton (1982); Hansen, Heaton and Yaron (1996). This approach is possible due to identical pattern between moment conditions and asymptotic behaviours of quantiles and interquantile means. I introduced the method of quantile matching, the method of interquantile mean matching, culminating in the development of two joint methods. The joint methods of parameter estimation considerably improves efficiency and accuracy by showing that one can exploit quantile data, interquantile mean data, and even class frequency data simultaneously. Many fundamental asymptotic results required for the development of new estimators are taken from two statistics textbooks: *A First Course in order statistics* (Arnold, Balakrishnan and Nagaraja, 1992), *Approximation theorems of mathematical statistics* (Serfling, 2009). These results are extremely important and used repeatedly.

Chapter 5 concerns the efficiency and stability issues of the newly developed estimation techniques. Theoretical questions such as how efficient the estimators perform when the number of groups increases or tends to infinity are addressed. The new estimators are also compared to the benchmark maximum likelihood estimation in terms of "limiting " asymptotic efficiency. By looking at the information matrix from a different view, I developed indices that are capable of measuring how "multicollinear" a distribution exhibits. Proof techniques in this chapter involve extensive use of Linear Algebra. All the unstated results employed can be found in *Linear Algebra and Its Applications* (Strang, 1980).

Chapter 6 demonstrates an application of new estimation framework to restore the income distribution in Australia for 14 years. Estimation routines are implemented in *Python* and executed on *Ariadne Number Cruncher*[3]. Earnings data for

---

[2] `http://functions.wolfram.com/`

[3] I extend my sincerest thanks to Dr Fedor Iskhakov who permitted me to use Ariadne.

Australian citizens are obtained from the Australian Bureau of Statistics (ABS). Outputs obtained include estimates of parameters and various inequality indices with valid standard errors. The trend of income inequality in Australia is then examined. Finally, chapter 7 concludes the paper with a summary and suggests potential ideas for future research.

Many formulas in this paper, especially in chapter 3, are derived using *Mathematica*. Its rapid prototyping capabilities and symbolic computation allow me to avoid tedious computations and focus on important ideas. Style of chapter 4 and 5 are mathematical, sometimes intuitive, non-rigorous. Indeed, I often take reasoning to the extreme, for example, letting the number of groups tend to infinity, or using the discrete-continuous analogy.

1.2 Literature Review

This section reviews **parametric approach** to inequality measure surrounding distributional analysis literature. Non-parametric estimation methods are less popular and will not be our concern. Each method is devised to address an estimation problem associated with a particular type of data, or a functional form. A review of statistical distributions relevant to the field of distributional analysis is presented first, followed by estimation techniques corresponding to the data types used.

The first economist who started the quest for a descriptive model of income is Pareto more than a century ago (Pareto and Politique, 1897). He observed that in many populations, the number of individuals whose earnings exceeded a given level is proportional to its reciprocal power, later on, known as Pareto power law (Mandelbrot, 1960). However, evidence has shown Pareto law is only applicable to the upper tail of the income distribution. R. Gibrat showed Log-normal more properly describes the income distribution in the middle-income range (Gibrat, 1931). Dagum (1977) in his seminal paper proposed a new model for the size distribution of income that satisfies a set of important assumptions. The new model, later become known as the Dagum distribution, is established by investigating properties of distribution function based on an empirical foundation. The three-parameter model which bears Dagum's name "still represents one of the most complete formalisations with regard to economic theory, stochastic derivation and possibilities for use in empirical analysis. Every parameter has a precise economic interpretation, and the model fits a complete series of formal-logic properties." (Lemmi and Betti, 2007). At around the same time, a new model for income distribution was proposed by Singhi and Maddala (1976). By generalising Pareto and Weibull distribution based on the concept of failure rate, they derived a new three-parameter income distribution. This distribution is shown to fit income data from various countries very well; thus, it becomes widely used in empirical studies. McDonald (1984) suggests the use of two generalised beta families as a model for the size distribution of incomes. These two families are shown to include the beta of the first and the second kind, Singh-Maddala, the Log-normal, Gamma, Weibull, Fisk as particular cases. Ultimately, McDonald and Xu (1995) unified generalised beta of the first and second kind into a single family named generalised beta. The generalised beta is shown to be extremely flexible, including more than 30 distributions as special or limiting cases. Therefore, the use of the generalised beta family is dominant in this paper for modelling the hypothetical distribution of income. Other suggestions are Gamma distribution (Salem and Mount, 1974; McDonald and Jensen, 1979), sech^2 distribution (Fisk, 1961), Yule distribution (Simon, 1955, 1957; Simon and Bonini, 1958), generalised normal-Laplace distribution (Reed and Wu, 2008). Comprehensive coverage of the use of functional forms in describing income distribution is discussed by Kleiber and Kotz (2003).

A recent trend in modelling income distribution is using mixture models. Chotikapanich and Griffiths (2008) suggested a Bayesian procedure to estimate a gamma mixture with two or three components using a sample of Canadian income microdata. The paper also obtained a posterior density of Gini coefficient and Lorenz curve. However, caution must be raised because the chosen prior densities are informative. Paapaa and Van Dijk (1998) studied the distributions at global scale of real GDP per capita for a combination of 120 countries from 1960 to 1989. These distributions appear to be bimodal, and a mixture of Weibull and truncated normal densities is adopted to describe the bimodal distribution. With microdata, income inequality in New Zealand was examined by Bakker and Creedy (1999) using a mixed distribution between Gamma and Exponential based on maximum likelihood estimation. Lubrano and Ndoye (2016) adapted the Bayesian approach using a mixture of log-normal distributions for UK income microdata. Their finding confirms that a mixture of Log-normal distributions is a powerful descriptive model for the UK income distribution. A concluding remark was made that a Pareto tail is potentially more satisfactory, suggesting the mixture of Log-normal and Pareto for future research.

McDonald and Xu (1995) introduce generalised beta family and maximise the multinomial generalised beta likelihood function using household income in class frequency format. Due to exceptional versatility, the use of generalised beta as a descriptive model of income achieves substantial goodness of fit. Nevertheless, the paper contains one unsatisfactory aspect: there is no standard error for parameter estimates. Chotikapanich, Griffiths and Rao (2007) investigated income inequality in China by adopting the same multinomial likelihood approach with class frequency data. Hinkley and Cox (1979) show the estimator obtained by maximising multinomial likelihood function is asymptotically efficient relative to other estimators based on grouped data; however, they are less efficient than maximum likelihood estimators based on individual observation. Using maximum likelihood estimation is standard in dealing with class frequency data; however, several other asymptotically equivalent procedures are viable such as Pearson minimum χ^2 and least squares estimator (Hinkley and Cox, 1979; McDonald and Ransom, 1979).

Various estimation procedures are proposed to deal with interquantile mean data. Chotikapanich and Griffiths (2002) converted interquantile mean data to Lorenz shares, and performed analysis on share data instead. Since the shares add up to unity, they assume these income proportions are Dirichlet distributed and use maximum likelihood for parameter estimation. The authors claim maximum likelihood estimates under Dirichlet assumption provide better Lorenz curve fit than nonlinear least square and other available estimation techniques. Kakwani and Podder (1976) introduced a new coordinate system for the Lorenz curve, viewing the egalitarian line as the x-axis. By imposing a multiplicative separable functional form of Lorenz curve on the new axes, the authors provide a simple and elegant technique for estimating parameters and extracting standard errors by linear regression. Another important paper that deserves attention is: "Estimating and combining national income distributions using limited data" written by Chotikapanich, Griffiths and Rao (2007). The authors came up with an estimation procedure (CGR), motivated by GMM, to minimise a quadratic objective function involving a list of moment conditions under interquantile mean data. The insight of using GMM is valuable and truly promising; however, I show their proposed data generating process is spurious. Chotikapanich, Rao and Tang (2007) then applied the CGR method to investigate income inequality in China using interquantile mean data. Parameters estimates are reported, but the knowledge of how accurate these estimates are cannot be extracted using their framework. Despite widespread use of quantiles in income surveys, there are few methods available in the distributional analysis to deal with this data type except for the method of quantile matching. Some preliminary results in quantile matching method are published in the papers by Dominicy and Veredas (2013) and Sgouropoulos, Yao and Yastremiz (2015). Apart from being equivalent in some respects, my development delve more deeply into the estimator's asymptotic efficiency and how the quantile matching estimation is connected with other matching methods.

Several other non-standard methods are available. The least square frequency estimator (LSFE) is introduced by Bakker and Creedy (1998) to obtain estimates of exponential family parameters using class frequency data. They use log-linear approximations to derive a standard linear regression equation with random noise associated with numerical integration errors. By making some assumptions on this "quadrature" errors, the authors can supply standard errors for their estimators. Kobayashi and Kakamu (2016) adopted a Bayesian approach based on sequential Markov Chain Monte Carlo to estimate parameters from the Lorenz curve using class frequency data.

Chapter 2

Theoretical Foundation

This chapter outlines and reviews the fundamental inequality and statistics concepts necessary for subsequent development. These fundamental concepts include inequality charting and measure, Lorenz curve, notions of quantile and interquantile mean, generalised beta family, exponential family and Expectation-Maximisation algorithm. The results presented in this chapter are used repeatedly in the subsequent chapters.

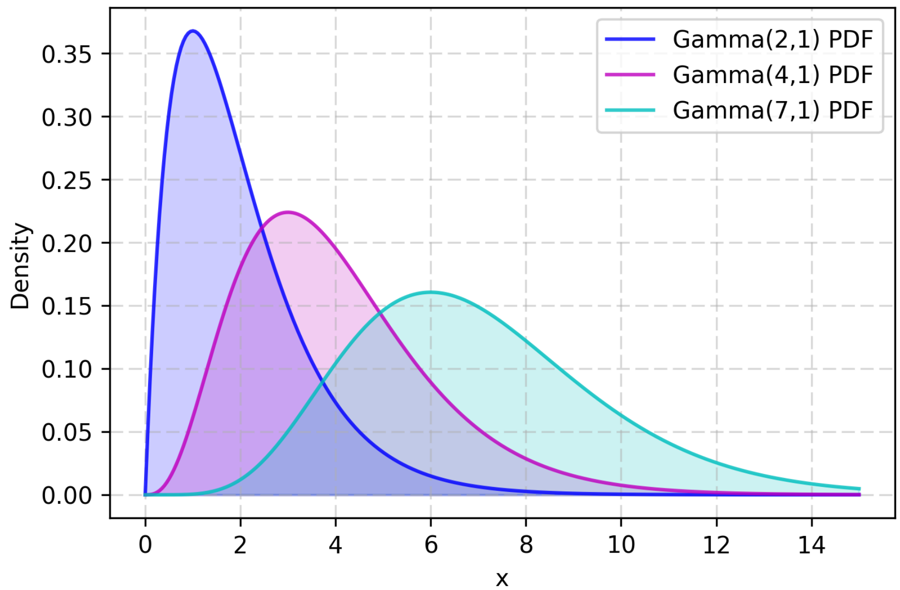
2.1 Inequality Charting and Measure

There are several ways to visualise the inequality of income, namely probability density function (PDF), cumulative distribution function (CDF), Parade of Dwarf, and Lorenz curve. Economists also desire to capture the inequality into one single measure or index, most notably the Gini coefficient, Theil index, and Pietra index.

Probability Density Function

Probability density function (PDF) describes how likely a random variable takes on a given value, denoted by $f_Y(y)$ (see Figure 2-1 for examples of PDF graph). PDF characterises the behaviour of a continuous random variable completely. All possible values that Y can take are the support of Y. If Y has the unit of dollar, then the density has unit probability per dollar. This simple fact proved very useful in dimensional analysis when one wants a simple check of some complicated probability-

Figure 2-1: GAMMA PROBABILITY DENSITY FUNCTIONS



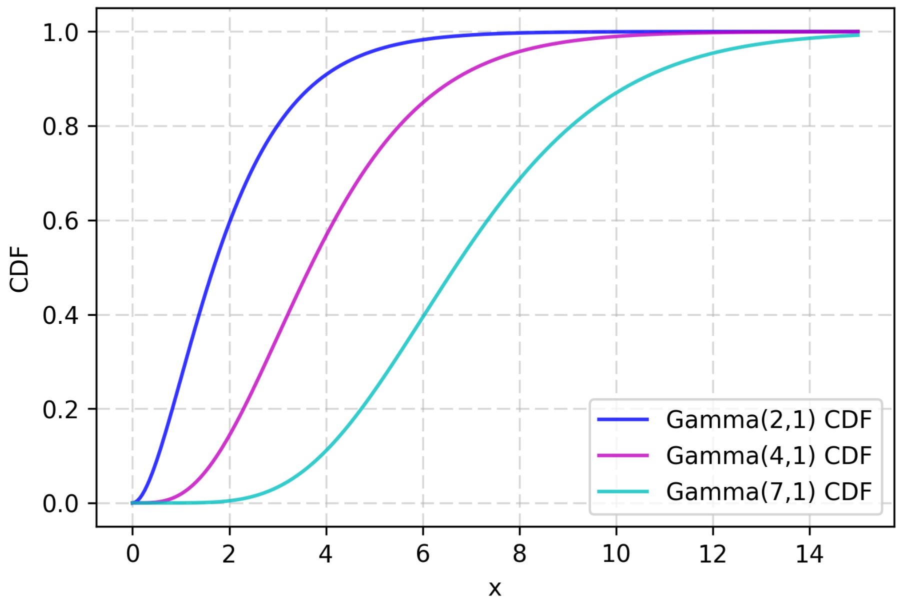
Notes: The leftmost curve represents Gamma$(2,1)$ density, The middle curve represents Gamma$(4,1)$ density, and the rightmost curve represents Gamma$(7,1)$ density. The shaded area under each curve is always 1.

related formula. In this thesis, we will focus exclusively on continuous distribution with nonnegative support $[0,\infty)$ to describe the hypothetical income distribution. Income density function also has another interpretation: suppose at income level y, the corresponding density is $f(y)$, then the proportion of the population having income ranging from $y-\epsilon$ to $y+\epsilon$ is roughly $2\epsilon f(y)$, provided that ϵ is a small number. The proportion of people whose income range from a to b is the area under the density curve and above the income range. If we employ density graphs to visualise the income distributions, a rightward shift in the density curves over time corresponds to a growth of the overall income level.

Cumulative Distribution Function

The cumulative distribution function (CDF) describes the probability that a random variable Y will be realised at a value less than y (see Figure 2-2 for examples of CDF). A random variable has a continuous distribution if its CDF is differentiable. Strictly

Figure 2-2: GAMMA CUMULATIVE DISTRIBUTION FUNCTIONS



Notes: The leftmost curve represents Gamma$(2, 1)$ CDF, The middle curve represents Gamma$(4, 1)$ CDF, and the rightmost curve represents Gamma$(7, 1)$ CDF. The CDF for a continuous random variable is always a smooth curve, strictly increasing from 0 to 1.

speaking, finitely many points can be allowed to be continuous and not differentiable, and density function for a continuous random variable is defined as the derivative of its CDF. The fundamental theorem of calculus says the CDF can be derived by integrating the PDF:

$$F_Y(y) = \mathrm{P}(Y \leq y) = \int_{-\infty}^{y} f_Y(t)\, dt \tag{2.1}$$

Expectation

The expected value of random variable Y (also known as mean, first moment), denoted by μ, is a measure of location of a distribution and is the most commonly used measure of central tendency. The expected value of a continuous random variable with PDF

$F_Y(y)$ is defined as:

$$\mu = E(Y) = \int_{-\infty}^{\infty} y f_Y(y)\, dy \tag{2.2}$$

Variance

In order to quantify the dispersion, statisticians often use variance that is defined as the average value of squared deviation from the mean. The variance of a continuous random variable with PDF $F_Y(y)$ is:

$$\sigma^2 = \mathrm{Var}(Y) = \int_{-\infty}^{\infty} (y-\mu)^2 f_Y(y)\, dy \tag{2.3}$$

Parade of Dwarfs

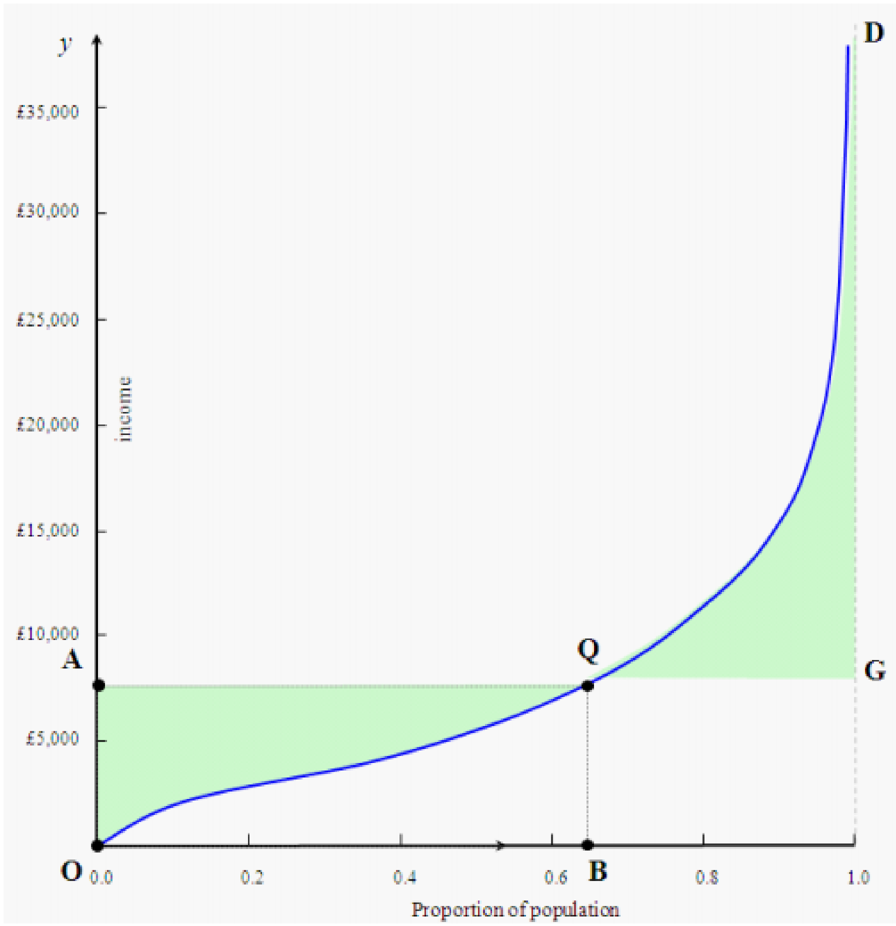
An instructive diagram capable of representing inequality is Jan Pen's *Parade of Dwarfs* Pen (1974). In fact, it is one of the most persuasive visual devices for investigating income distribution. Imagine that each person in the population is assigned a height proportional to his or her income. The person with the highest income is the tallest whereas the person with the lowest income is the shortest. Each person is standing at a point between 0 and 1, and they must line up in ascending order of height. Parade of Dwarfs' diagram 2-3 depicted an example of the imaginary march specified above.

Quantile Function

Suppose continuous random variable Y admits CDF $F_Y(y)$. The CDF is continuous and strictly increasing; therefore it exhibits an inverse function called F^{-1}, known as quantile function. Denote $Q = F^{-1}$ as an alternative notation for quantile function. With new notation, the median can be alternatively written as $Q(0.5)$, the first quartile is $Q(0.25)$. Quantile and quantile function are fundamental concepts throughout the thesis, which are shown in the following important theorem.

Theorem 2.1.1 (**Universality of the Uniform**). *X is a random variable with strictly increasing, continuous CDF F, so that the quantile function Q is well-defined.*

Figure 2-3: PARADE OF DWARF



Notes: Parade of Dwarf diagram also emphasises the presence of abnormally large income, which is not easy to spot when looking at a corresponding density function. Source: Cowell (2011)

Let ρ_1, ρ_2 be any number between $[0, 1]$ with $\rho_1 < \rho_2$ and put $x_1 = Q(\rho_1), x_2 = Q(\rho_2)$. $F_{(x_1,x_2)}$ represents the truncated distribution of F on (x_1, x_2). Then the following statements hold.

i. $F(X) \sim \text{Unif}(0, 1)$

ii. $Q(U) \sim F$ *where* $U \sim \text{Unif}(0, 1)$

iii. $Q(U) \sim F_{(x_1,x_2)}$ *where* $U \sim \text{Unif}(\rho_1, \rho_2)$*, giving*

$$\mathrm{E}(X^h|\rho_1 \leq F(X) \leq \rho_2) = \frac{1}{\rho_2 - \rho_1} \int_{\rho_1}^{\rho_2} Q(u)^h \, du \tag{2.4}$$

Proof. For any given u between $(0, 1)$:

$$P(F(X) \leq u) = P(X \leq Q(u)) = F(Q(u)) = u \tag{2.5}$$

which coincides with the CDF of standard uniformly distributed random variable. Therefore $F(X) \sim \text{Unif}[0, 1]$.
For any given x, we also have:

$$P(Q(U) \leq x) = P(U \leq F_X(x)) = F_X(x) \tag{2.6}$$

which agrees with the CDF of X. The third statement can be proved similarly. □

Equation 2.4 provides formula for population h^{th} interquantile moments of distribution F from ρ_1-quantile to ρ_2-quantile (also denote as $\mu^{(h)}(\rho_1, \rho_2)$ with the suppression of the super-index when $h = 1$). As an important special cases, we can obtain interquantile mean by setting $h = 1$ and interquantile second moment $h = 2$, and interquantile variance is given by:

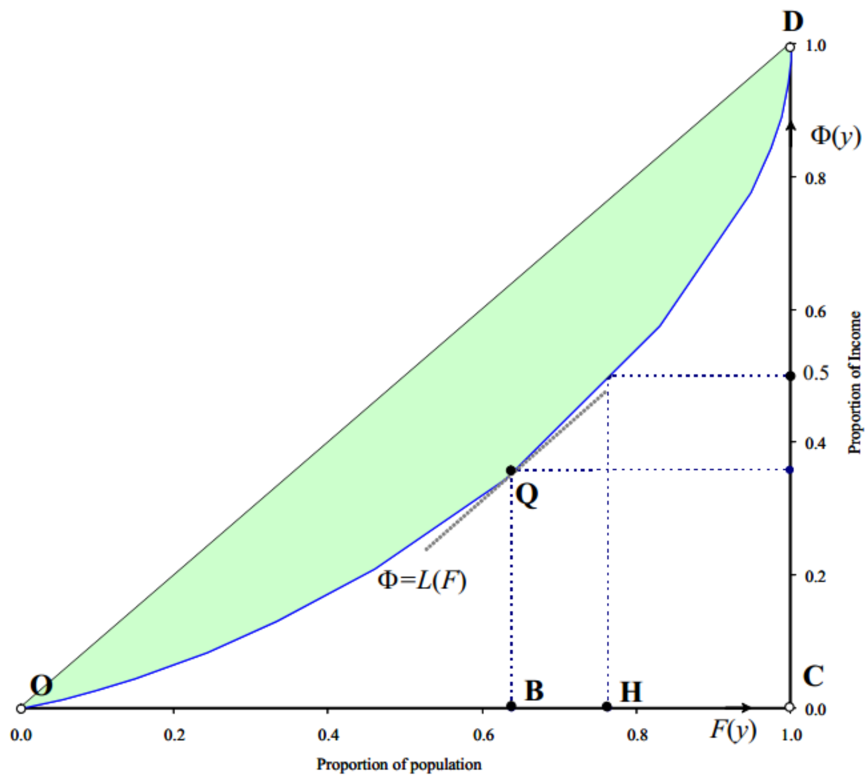
$$\sigma^2(\rho_1, \rho_2) = \mu^{(2)}(\rho_1, \rho_2) - \mu(\rho_1, \rho_2)^2 \tag{2.7}$$

Lorenz Curve

Lorenz (1905) introduced the Lorenz curve as a powerful method for illustrating the concentration of income and wealth distribution. The curve shows the proportion of overall income π_2 for wealth possessed by the bottom $x\% = \pi_1$ of the people. A society where all people have the same income is represented by a straight line $\pi_2 = \pi_1$, also known as the *egalitarian line*. Figure 2-4 illustrates a typical graph of the Lorenz curve. There are two ways to handle the Lorenz curve analytically: Direct function, Parametric Equation. The results are summarised by the following two theorems.

Theorem 2.1.2 (Parametric Equation of Lorenz Curve). *Assuming that an underlying income distribution has PDF $f_Y(y)$ and CDF $F_Y(y)$. Lorenz curve is*

Figure 2-4: LORENZ CURVE



Notes: The Lorenz curve always starts at $(0,0)$ and ends at $(1,1)$ and exhibits convexity. Point Q on the curve says the bottom 64% of the population have 36% of total income. Source: Cowell (2011)

traced out by the parametric equation:

$$\begin{aligned} \pi_1(y) &= F_Y(y) \\ \pi_2(y) &= \frac{1}{E(Y)} \int_0^y t f(t)\, dt \end{aligned} \tag{2.8}$$

Proof. At income level y, the fraction of population earning income less than y is $\pi_1(y) = F_Y(y)$ while the income share owned by this fraction of population is:

$$\pi_2(y) = \frac{P(Y \le y)E(Y|Y \le y)}{1 \times E(Y)} = \frac{1}{E(Y)} \int_0^y t f(t)\, dt \tag{2.9}$$

as desired. □

Theorem 2.1.3 (**Direct Lorenz Function**). *Direct formula for Lorenz curve is given by:*

$$L(\pi) = \frac{1}{EY} \int_0^\pi F_Y^{-1}(q)\, dq = \frac{1}{EY} \int_0^\pi Q_Y(q)\, dq, \quad \pi \in [0, 1] \tag{2.10}$$

Proof. Put $y = Q_Y(\pi)$, we immediately have $\pi = F_Y(y)$. Then using part 3 of Theorem 2.1.1:

$$E(Y|Y \leq y) = \frac{1}{\pi} \int_0^\pi Q_Y(q)\, dq \tag{2.11}$$

The income share of the group having income less than y is:

$$\pi_2(y) = \frac{P(Y \leq y)E(Y|Y \leq y)}{1 \times E(Y)} = \frac{1}{EY} \int_0^\pi Q_Y(q)\, dq \tag{2.12}$$

as desired. □

Gini coefficient

Gini coefficient is a measure of variability introduced by Gini (1912). The Gini coefficient is defined as twice the area trapped between the Lorenz curve and the egalitarian line (twice the green-shaded area in Figure 2-4). The Gini coefficient, denoted by G, provides a simple quantification of income inequality between 0 and 1. Indeed, $G = 0$ represents the situation where every individual receives the same income level, while $G = 1$ represents the situation where all persons receive 0 income except for one. Mathematically, the Gini coefficient for a continuous distribution is:

$$G = 1 - 2 \int_0^1 L(\pi)\, d\pi \tag{2.13}$$

The computation of Gini coefficient is easier due to the following theorem, proved by Dorfman (1979).

Theorem 2.1.4 (Dorfman). *Gini coefficient can be alternatively expressed as*

$$G = \frac{1}{\mathrm{EY}} \int_0^\infty (1 - F_Y(y))F_Y(y)\, dy = 1 - \frac{1}{\mathrm{EY}} \int_0^\infty (1 - F_Y(y))^2\, dy$$

Proof. The alternative and shorter proof is deferred to Appendix A.1 □

Theil Index

Theil (1967) discovered that the entropy concept from information theory provides a useful way of measuring inequality. It can be viewed as a measure of redundancy, lack of diversity, isolation, segregation, inequality, non-randomness, and compressibility [1]. The Theil-T index is defined as:

$$T_T = \mathrm{E}\left(\frac{Y}{\mathrm{E}Y}\ln\left(\frac{Y}{\mathrm{E}Y}\right)\right) = \int_0^\infty \frac{y}{\mathrm{E}Y}\ln\left(\frac{y}{\mathrm{E}Y}\right) f_Y(y)\,dy \tag{2.14}$$

Pietra Index

Another common inequality index is Pietra, which is defined as the maximal vertical distance between the Lorenz curve and egalitarian line:

$$P = \max_y\{\pi_1(y) - \pi_2(y)\} = \max_\pi\{\pi - L(\pi)\} \tag{2.15}$$

Two other representations for Pietra index, which facilitate numerical computation later on, are provided by Sarabia and Jordá (2014).

$$P = \frac{\mathrm{E}|X-\mu|}{2\mu} \tag{2.16}$$

$$P = F_Y(\mu) - L(F_Y(\mu)) \tag{2.17}$$

Theil and Pietra indices are alternative inequality measures that will later on be estimated using Australian earnings panel data. A estimated series of different inequality indices allow us to investigate the trend of income inequality in Australia.

[1] `https://en.wikipedia.org/wiki/Theil_index`

2.2 Generalised Beta Family

Statistical distributions are the key to analysing empirical data and many statistical processes. In this section, we will review the development of Generalised beta of the first kind, generalised beta of the second kind, culminating in generalised beta family. The final class of distributions is extremely flexible and capable of fitting a wide range of distributional shapes, which make the use of generalised beta family as an income distribution model useful.

Generalised Beta of the first kind

The generalised beta of the first kind was first presented in a paper by McDonald (1984). In his work, he did not specify the process of deriving the new density function. However, it is worth understanding the transformation that relates beta distribution to the generalised beta of the first kind. The process is as follows: suppose random variable $X \sim \text{Beta}(p, q)$ has PDF:

$$f_X(x) = \frac{1}{B(p,q)} x^{p-1}(1-x)^{q-1}, \quad x \in (0,1) \tag{2.18}$$

This well known standard beta distribution with two parameters is undoubtedly one of the most frequently used distribution in statistics.

Despite its flexibility, the support of Beta distribution is restricted to the interval $(0, 1)$. To extend the range of values the distribution can take, we make a transformation:

$$Y_1 = X^{1/a} b, \quad b > 0 \tag{2.19}$$

Then say random variable $Y_1 \sim \text{GB1}(p, q, a, b)$. The PDF and CDF associated with random variable Y_1 are:

$$f_1(y) = \frac{|a| y^{ap-1}(1-(y/b)^a)^{q-1}}{b^{ap} B(p,q)}, \quad y \in (0,b) \tag{2.20}$$

$$F_1(y) = I_{y^a/b^a}(p,q), \quad y \in (0,b) \tag{2.21}$$

This four-parameter PDF is very versatile, in the sense that it includes beta of the first kind (set $a = 1$), Pareto distribution (set $a = -1, q = 1$), generalised gamma GG (set $b = q^{1/a}$ and let q go to infinity). It also nests some obvious distributions such as standard Beta (put $a = 1, b = 1$, cancelling the effect of the above transformation), or uniform (which is a special case of standard beta), power distribution P and Unit Gamma (McDonald, 1984; McDonald and Xu, 1995)

Statistical Properties of GB1 family

With the density and distribution functions of GB1 family, Statistical properties such as moment, quantile function, Lorenz function can be derived. First, the h^{th} moment associated with GB1 family is provided by the formula:

$$\mathrm{E}(Y_1^h) = \frac{b^h B(p + 1/a, q)}{B(p, q)} \tag{2.22}$$

Using Equation 2.22, one can acquire formulas for the mean, variance, skewness, kurtosis of GB family. By solving for the inverse CDF, the quantile function is obtained:

$$Q_1(\pi) = b I_\pi^{-1}(p, q)^{1/a}, \quad \pi \in (0, 1) \tag{2.23}$$

Denote Q_1^A as the antiderivative for quantile function: $\frac{d}{d\pi} Q_1^A(\pi) = Q_1(\pi)$. Q_1^A is chosen so that $Q_1^A(0) = 0$ and its analytic expression is given by:

$$\begin{aligned} Q_1^A(\pi) = {} & \frac{ab I_\pi^{-1}(p, q)^{\frac{1}{a}+p}}{(ap+1)(ap+a+1)B(p, q)} \times \\ & \left[(ap+a+1)\, {}_2F_1\left(p + \frac{1}{a}, -q; p + \frac{1}{a} + 1; I_\pi^{-1}(p, q)\right) + \right. \\ & \left. (ap+1) I_\pi^{-1}(p, q)\, {}_2F_1\left(p + \frac{1}{a} + 1, 1 - q; p + \frac{1}{a} + 2; I_\pi^{-1}(p, q)\right) \right] \end{aligned} \tag{2.24}$$

Using theorem 2.1.3 regarding the direct Lorenz function, the closed-form expression for Lorenz function is attainable from the expression 2.24 for antiderivative of

GB1 quantile function :

$$L_1(\pi) = \frac{1}{\mathrm{E}(Y_1)} Q_1^A(\pi), \quad \pi \in (0,1) \tag{2.25}$$

One can compute Gini coefficient numerically from equation 2.25 or consult analytic formula for Gini coefficient provided in (McDonald and Ransom, 1979).

From equation 2.4, to compute the interquantile moments, we need to integrate the quantile function raising to some power. Luckily, there is an analytic expression for such integral, and so the antiderivative of $Q_1^n(\pi)$ is:

$$\begin{aligned} Q^{A_n}(\pi) = & \frac{ab^n I_\pi^{-1}(p,q)^{p+\frac{n}{a}}}{(ap+n)(ap+a+n)B(p,q)} \times \\ & \Big[(ap+a+n)\,{}_2F_1\left(p+\frac{n}{a}, -q; p+\frac{n}{a}+1; I_\pi^{-1}(p,q)\right) + \\ & (ap+n)I_\pi^{-1}(p,q)\,{}_2F_1\left(p+\frac{n}{a}+1, 1-q; p+\frac{n}{a}+2; I_\pi^{-1}(p,q)\right)\Big] \end{aligned} \tag{2.26}$$

Generalised Beta of the second kind

The generalised beta of the second kind was first introduced by McDonald (1984). He did not specify the process of deriving the density function for tor GB2 family. However, it is worth understanding how beta distribution is related to the generalised beta of the second kind. Suppose random variable $X \sim \mathrm{Beta}(p,q)$, and we make a transformation:

$$Y_2 = \left(\frac{X}{1-X}\right)^{1/a} b, \quad a > 0, b > 0 \tag{2.27}$$

The support for random variable Y_2 is no longer as limited as $(0,1)$. In fact, Y_2 can take values on $(0,\infty)$, and its associated density function and distribution function are:

$$f_2(y) = \frac{ay^{ap-1}}{B(p,q)b^{ap}\left(1+\frac{y^a}{b^a}\right)^{p+q}}, \quad y > 0 \tag{2.28}$$

$$F_2(y) = BR\left(\frac{y^a}{b^a+y^a}; p, q\right), \quad y > 0, \quad (1 \text{ if } y^a > b^a + y^a) \tag{2.29}$$

Y_2 is then said to follow Generalised Beta distribution of the second kind, with core shape parameters p,q, additional shape parameter a, and scale parameter b:

$$Y_2 \sim \text{GB2}(p, q, a, b) \tag{2.30}$$

McDonald (1984); McDonald and Xu (1995) showed the GB2 family nests many important distributions as special or limiting cases, namely generalised gamma GG, Burr types 3 and 12 BR3 and BR12, gamma distribution, Fisher F, Lomax L, Weibull W, Log-normal LN, Rayleigh R, Chi Square χ^2, half student t^+ and Exponential distribution.

Statistical Properties of GB2 family

With the density and distribution functions of GB2 family, Statistical properties such as moment, quantile function, Lorenz function can be derived. First, the h^{th} moment associated with GB1 family is provided by the formula:

$$E(Y_2^h) = \frac{b^h \Gamma(p + h/a)\Gamma(q - h/a)}{\Gamma(p)\Gamma(q)} \tag{2.31}$$

Using Equation 2.31, one can acquire formulas for the mean, variance, skewness, kurtosis of GB1 family. Being able to obtain analytic expressions is significant in speeding up estimation procedures. By solving for the inverse CDF, the quantile function is given as:

$$Q_2(\pi) = b\left(\frac{I_\pi^{-1}(p,q)}{1 - I_\pi^{-1}(p,q)}\right)^{1/a}, \quad \pi \in (0,1) \tag{2.32}$$

Denote Q_2^A as the antiderivative for quantile function: $\frac{d}{d\pi}Q_2^A(\pi) = Q_2(\pi)$. Q_2^A is chosen so that $Q_2^A(0) = 0$ and its analytic expression is given by:

$$Q_2^A(\pi) = \frac{abI_\pi^{-1}(p,q)^{p+1/a}}{(ap+1)(ap+a+1)B(p,q)} \times \\ \Bigg[(ap+a+1)\,{}_2F_1\left(p+\frac{1}{a}, \frac{1}{a}-q; p+\frac{1}{a}+1; I_\pi^{-1}(p,q)\right) + \\ (ap+1)I_\pi^{-1}(p,q)\,{}_2F_1\left(p+\frac{1}{a}+1, -q+\frac{1}{a}+1; p+\frac{1}{a}+2; I_\pi^{-1}(p,q)\right)\Bigg] \quad (2.33)$$

Using theorem 2.1.3 regarding the direct Lorenz function, the closed-form expression for Lorenz function is attainable from the expression 2.33 for antiderivative of GB2 quantile function :

$$L_2(\pi) = \frac{1}{E(Y_2)} Q_2^A(\pi) \quad (2.34)$$

Gini coefficient can be computed numerically from equation 2.34 or from analytic formula provided in (McDonald and Ransom, 1979).

Using equation 2.4, one can derive is an analytic expression for the antiderivative of $Q_2^n(\pi)$:

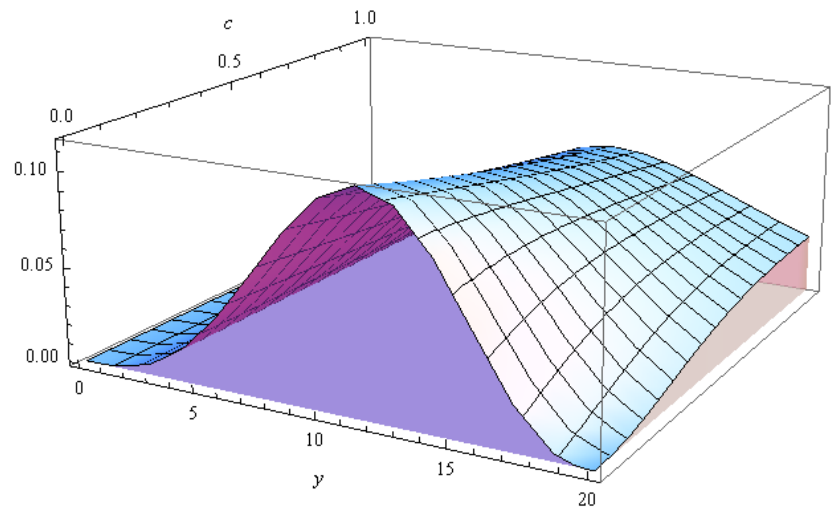
$$Q_2^{A_n}(\pi) = \frac{ab^n I_\pi^{-1}(p,q)^{p+n/a}}{(ap+n)(ap+a+n)B(p,q)} \times \\ \Big[(ap+a+n)\,{}_2F_1\left(p+\frac{n}{a}, \frac{n}{a}-q; p+\frac{n}{a}+1; I_\pi^{-1}(p,q)\right) + \\ (ap+n)I_\pi^{-1}(p,q)\,{}_2F_1\left(p+\frac{n}{a}+1, -q+\frac{n}{a}+1; p+\frac{n}{a}+2; I_\pi^{-1}(p,q)\right)\Big] \quad (2.35)$$

Generalised Beta Unification

In their influential paper, McDonald and Xu (1995) synthesised GB1 and GB2 families into a larger class of distributions, namely Generalised beta. The unification process is as follows: suppose random variable $X \sim \text{Beta}(p,q)$, we make a transformation that nests the transformations used in GB1 and GB2:

$$Y = \left(\frac{X}{1-cX}\right)^{1/a} b, \quad b > 0, c \in [0,1] \quad (2.36)$$

Figure 2-5: PDF OF GENERALISED BETA AT VARYING $c \in (0, 1)$



Notes: The outermost vertical cross-section corresponds to GB1 density with finite support, whereas the innermost vertical cross-section corresponds to GB2 density with infinite support. Intermediate vertical cross-sections depict the whole spectrum of the unified GB family densities corresponding to c running from 0 to 1.

When $c = 0$, Y is GB1 distributed while when $c = 1$, Y is GB2 distributed. Moreover, there is a whole spectrum of intermediate distributions associated with c running from 0 to 1. Figure 2-5 depicts the distributional shape of GB family at varying c between $(0, 1)$. The outermost cut corresponds to GB1 density with finite support, whereas the innermost cut corresponds to GB2 density with infinite support. Random variable Y has PDF and CDF:

$$f_Y(y) = \frac{|a|y^{ap-1}(1-(1-c)y^a/b^a)^{q-1}}{B(p,q)b^{ap}\Big(1+c\,\dfrac{y^a}{b^a}\Big)^{p+q}}, \quad y > 0 \tag{2.37}$$

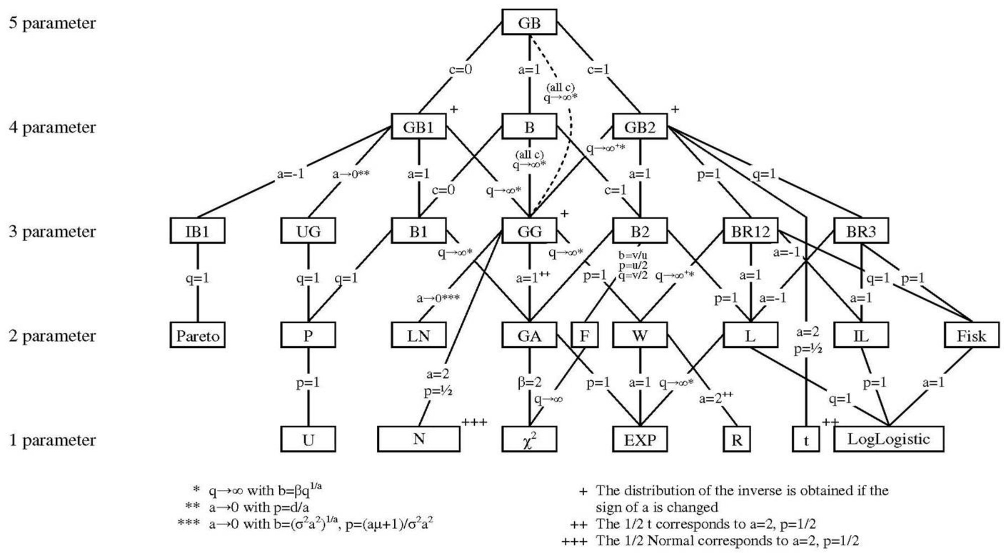
$$F_Y(y) = BR\Big(\frac{y^a}{b^a+cy^a}; p, q\Big), \quad y > 0 \tag{2.38}$$

Y is then said to follow Generalised Beta distribution, with parameters p, q, a, b, c

$$Y \sim \text{GB}(p, q, a, b, c) \tag{2.39}$$

The GB2 family nests many important distributions as special or limiting cases,

Figure 2-6: GENERALISED BETA FAMILY TREE



Source: (McDonald and Xu, 1995)

namely generalised gamma GG, Burr types 3 and 12 BR3 and BR12, gamma distribution, Fisher F, Lomax L, Weibull W, Log-normal LN, Rayleigh R, Chi Square χ^2, half student t^+,half Normal N^+, and Exponential distribution. Figure 2-6 show a remarkable family tree subsumed by the unified generalised beta family.

Statistical Properties of GB family

With the density and distribution functions of GB family, statistical properties such as moments, quantile function, Lorenz function can be derived. First, the h^{th} moment associated with GB family is provided by the formula:

$$E(Y^h) = \frac{b^h \Gamma(q) \Gamma\left(\frac{h}{a} + p\right) {}_2F_1\left(\frac{h}{a}, \frac{h}{a} + p; \frac{h}{a} + p + q; c\right)}{B(p,q)\Gamma\left(\frac{h}{a} + p + q\right)} \tag{2.40}$$

Using Equation 2.40, one can acquire formulas for the mean, variance, skewness, kurtosis of GB1 family. Being able to obtain analytic expressions is significant in speeding up estimation procedures. By solving for the inverse CDF, the quantile function is given as:

$$Q(\pi) = b\left(\frac{I_\pi^{-1}(p,q)}{1 - cI_\pi^{-1}(p,q)}\right)^{1/a} \tag{2.41}$$

Denote Q^A as the antiderivative for the GB quantile function: $\frac{d}{d\pi}Q^A(\pi) = Q(\pi)$. Q^A is chosen so that $Q^A(0) = 0$ and its analytic expression is given by:

$$Q^A(\pi) = \frac{abI_\pi^{-1}(p,q)^{p+1/a}}{(ap+1)(ap+a+1)B(p,q)} \times \tag{2.42}$$

$$\Bigg[(ap+a+1)F_1\left(p+\frac{1}{a}; -q, \frac{1}{a}; p+\frac{1}{a}+1; I_\pi^{-1}(p,q), cI_\pi^{-1}(p,q)\right) +$$

$$(ap+1)I_\pi^{-1}(p,q)F_1\left(p+\frac{1}{a}+1; 1-q, \frac{1}{a}; p+\frac{1}{a}+2; I_\pi^{-1}(p,q), cI_\pi^{-1}(p,q)\right)\Bigg] \tag{2.43}$$

Using theorem 2.1.3 regarding the direct Lorenz function, the closed-form expression for Lorenz function is attainable from the expression 2.42 for antiderivative of GB2 quantile function :

$$L(\pi) = \frac{1}{E(Y)}Q^A(\pi) \tag{2.44}$$

Using equation 2.4, an analytic expression for the antiderivative of $Q_2^n(\pi)$ is possible:

$$Q^{A_n}(\pi) = \frac{ab^nI_\pi^{-1}(p,q)^{p+n/a}}{(ap+n)(ap+a+n)B(p,q)} \times$$

$$\Big[(ap+a+n)F_1\left(p+\frac{n}{a}; -q, \frac{n}{a}; p+\frac{n}{a}+1; I_\pi^{-1}(p,q), cI_\pi^{-1}(p,q)\right) +$$

$$(ap+n)I_\pi^{-1}(p,q)F_1\left(p+\frac{n}{a}+1; 1-q, \frac{n}{a}; p+\frac{n}{a}+2; I_\pi^{-1}(p,q), cI_\pi^{-1}(p,q)\right)\Big] \tag{2.45}$$

2.3 Exponential Family

Most of commonly used distributions in statistics and econometrics are from exponential family: Binomial, Geometric, Normal, Exponential, Gamma, Beta and so on. Its representation is chosen for mathematical convenience and generality, and based on useful algebraic properties. Reviewing the representation and its properties is the main goal of this section. For a good treatment of exponential family, see Kroese, Chan et al. (2014).

Definition 2.3.1 (Exponential Family). . Let $\mathbf{X} = (X_1, X_2, ..., X_n)$ be a random vector with joint PDF $f(\mathbf{x}; \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, ...\theta_d)$ is a parameter vector. $\mathbf{X}$ is said to belong to an **m-dimensional exponential family** if there exist real-valued functions $T_j(\mathbf{x})$, $\eta_j(\boldsymbol{\theta})$, $j = 1, 2, ..., m$ and $h(\mathbf{x}) > 0$, and a normalising function $c(\boldsymbol{\theta}) > 0$, such that

$$f(\mathbf{x}; \boldsymbol{\theta}) = c(\boldsymbol{\theta}) \exp\left(\sum_{j=1}^{m} \eta_j(\boldsymbol{\theta}) T_j(\mathbf{x})\right) h(\mathbf{x}) \tag{2.46}$$

An important subclass of exponential family is natural exponential family, which is obtained by restricting $m = d$ and choosing $\eta_j(\boldsymbol{\theta}) = \theta_j$, $j = 1, 2, ..., d$.

Definition 2.3.2 (Natural Exponential Family). . Let $\mathbf{X} = (X_1, X_2, ..., X_n)$ be a random vector with joint PDF $f(\mathbf{x}; \boldsymbol{\theta})$, where $\boldsymbol{\theta} = (\theta_1, \theta_2, ...\theta_d)$ is a parameter vector. X is said to belong to an **natural exponential family** if there exist real-valued functions $T_j(\mathbf{x})$, $j = 1, 2, ..., d$ and $h(\mathbf{x}) > 0$, and a normalising function $c(\boldsymbol{\theta}) > 0$, such that

$$\begin{aligned} f(\mathbf{x}; \boldsymbol{\theta}) &= c(\boldsymbol{\theta}) \exp\left(\sum_{j=1}^{d} \theta_j T_j(\mathbf{x})\right) h(\mathbf{x}) \\ &= c(\boldsymbol{\theta}) e^{\boldsymbol{\theta}^{\mathrm{T}} \mathbf{T}(\mathbf{x})} h(\mathbf{x}) \end{aligned} \tag{2.47}$$

where $\mathbf{T}(\mathbf{x}) = (T_1(\mathbf{x}), T_2(\mathbf{x}), ..., T_d(\mathbf{x}))$ is a vector of sufficient statistics associated with the natural exponential family.

Lemma 2.3.3 (Truncated Exponential Family is Exponential Family). *The truncated distribution on interval $[a, b)$ for a univariate natural exponential family with density $f(x; \boldsymbol{\theta}) = c(\boldsymbol{\theta}) e^{\boldsymbol{\theta}^{\mathrm{T}} \boldsymbol{T}(x)} h(x)$ still belongs to natural exponential family.*

Proof. Let $F(x; \boldsymbol{\theta})$ be the corresponding CDF, then the truncated density is given by:

$$f_{[a,b)}(x;\boldsymbol{\theta}) = \frac{c(\boldsymbol{\theta})e^{\boldsymbol{\theta}^{\mathrm{T}}\mathbf{T}(x)}h(x)\mathbb{1}_{[a,b)}(x)}{F(b;\boldsymbol{\theta}) - F(a;\boldsymbol{\theta})} \tag{2.48}$$

By putting $\tilde{c}(\boldsymbol{\theta}) = c(\boldsymbol{\theta})/(F(b;\boldsymbol{\theta}) - F(a;\boldsymbol{\theta}))$ and $\tilde{h}(x) = h(x)\mathbb{1}_{[a,b)}(x)$, the truncated density $f_{[a,b)}$ has the desired representation of natural exponential family specified in definition 2.3.2. □

The result remains valid if using the general exponential family instead of its natural version with similar proof. It is a simple and striking fact that a truncated exponential family also belongs to exponential family.

Lemma 2.3.4 (Efficient Score for an Exponential Family). *The efficient Score for a natural exponential family with PDF $f(\boldsymbol{x};\boldsymbol{\theta}) = c(\boldsymbol{\theta})e^{\boldsymbol{\theta}^{\mathrm{T}}\boldsymbol{T}(\boldsymbol{x})}h(\boldsymbol{x})$ is given by*

$$S(\boldsymbol{\theta};\boldsymbol{X}) = \frac{\nabla c(\boldsymbol{\theta})}{c(\boldsymbol{\theta})} + \boldsymbol{T}(\boldsymbol{X}) \tag{2.49}$$

then deduce that

$$\mathrm{E}_{\boldsymbol{\theta}}\,\boldsymbol{T}(\boldsymbol{X}) = -\frac{\nabla c(\boldsymbol{\theta})}{c(\boldsymbol{\theta})} \tag{2.50}$$

Proof. The log-likelihood function is

$$\ln f(\mathbf{X};\boldsymbol{\theta}) = \ln c(\boldsymbol{\theta}) + \boldsymbol{\theta}^{\mathrm{T}}\mathbf{T}(\mathbf{X}) + \ln h(\mathbf{X}) \tag{2.51}$$

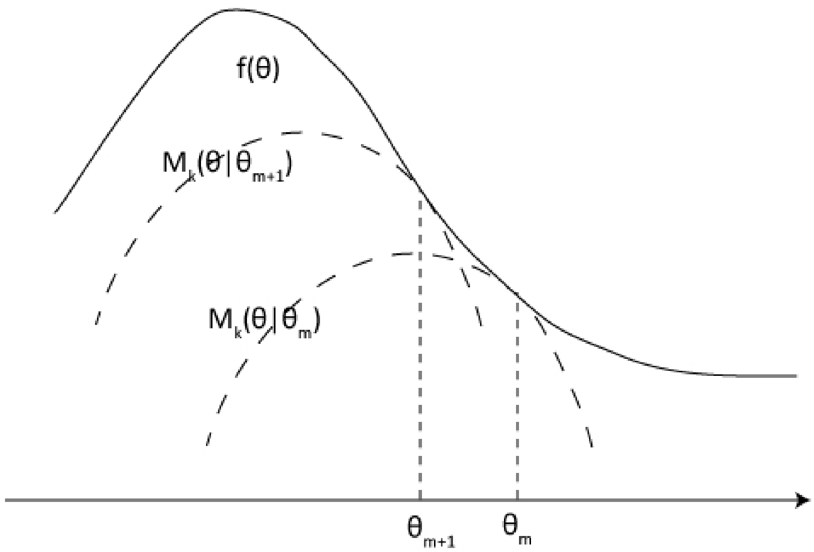
Now take the gradient with respect to θ:

$$S(\boldsymbol{\theta};\mathbf{X}) = \frac{\nabla c(\boldsymbol{\theta})}{c(\boldsymbol{\theta})} + \mathbf{T}(\mathbf{X}) \tag{2.52}$$

and obtain desired result by using the fact $\mathrm{E}_{\boldsymbol{\theta}}S(\boldsymbol{\theta};\mathbf{X}) = \mathbf{0}$. □

This result provides the analytical formula for the expectation of sufficient statistics. By invoking Lemma 2.3.3 and Lemma 2.3.4, if a univariate distribution is from natural exponential family and its CDF is expressible, then the expectation of sufficient statistics for truncated distribution also have closed-form expression.

Figure 2-7: MM ALGORITHM: MINORISE-MAXIMISATION



Source: `https://en.wikipedia.org/wiki/MM_algorithm`

2.4 Expectation Maximisation Algorithm

Expectation Maximisation (EM) algorithm is a practical numerical method for maximising likelihood function. EM algorithm is first introduced in the seminal paper written by Dempster, Laird and Rubin (1977). The algorithm can be considered as a subclass of **MM algorithm** (Majorise-Minimisation or Minorise-Maximisation). Indeed, EM algorithm is a special case of Minorise-Maximisation algorithm. The use of the word "algorithm" here is a misnomer because the algorithm itself is not an algorithm, but a description of how to construct an optimisation algorithm. This section adapts the coverage of EM algorithm in Kroese, Chan et al. (2014) with several changes in notation and explanation.

Figure 2-7 depicts a typical iteration of doing MM algorithm. Suppose we wish to find argument θ to maximise objective function $f(\theta)$. One approach is to find a sequence of guess $(\theta_m)_{m\geq 0}$ such that function f shows improvement for every consecutive terms in the sequence, i.e. $f(\theta_m) \leq f(\theta_{m+1})$. At each m, if we are able to find a minor $M(\theta|\theta_m)$ such that the following conditions are satisfied: $(i)\, M(\theta_m|\theta_m) =$

$f(\theta_m)$, $(ii)\, M(\theta|\theta_m) \leq f(\theta)$. Typically, we aim to find the minor that is easier to maximise than the objective function itself. Suppose θ_{m+1} maximises the minor, then:

$$f(\theta_m) = M(\theta_m|\theta_m) \leq M(\theta_{m+1}|\theta_m) \leq f(\theta_{m+1}) \tag{2.53}$$

Suppose that, for a given vector of observation $\mathbf{x} = (x_1, x_2, ..., x_n)$, we wish to compute the MLE:

$$\hat{\boldsymbol{\theta}} = \text{argmax}\, L(\boldsymbol{\theta}; \mathbf{x}) \tag{2.54}$$

where $L(\boldsymbol{\theta}, \mathbf{x}) = f(\mathbf{x}; \boldsymbol{\theta})$ is the likelihood function

A root-finding routine can be used for score function to obtain $\hat{\boldsymbol{\theta}}$. However, computing the score function and Hessian matrix analytically (required by Newton-Ralphson method) is potentially very complicated. Instead of maximising the likelihood function directly, the EM algorithm augments the data $\mathbf{x}$ with suitable vector of latent (or hidden) variable $\mathbf{z}$ such that

$$f(\mathbf{x}; \boldsymbol{\theta}) = \int f^c(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})\, d\mathbf{z} \tag{2.55}$$

The function of $\boldsymbol{\theta}$

$$L^c(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) = f^c(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) \tag{2.56}$$

is called **complete-data likelihood** function. The main merit of the data augmentation step is that it is often feasible to introduce latent variables $\mathbf{z}$ in such a way that the maximisation of the complete-data likelihood $L^c(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z})$ or its logarithm is much easier than maximising the original likelihood function $L(\boldsymbol{\theta}; \mathbf{x})$ or log-likelihood function $l(\boldsymbol{\theta}; \mathbf{x}) = \ln L(\boldsymbol{\theta}; \mathbf{x})$

Definition 2.4.1 (**KL Distance**)**.** A useful way to measure how far away a PDF g from another PDF h is the **Kullback-Leibler (KL) distance** (also known as KL divergence or cross-entropy distance), defined as:

$$\mathcal{D}(g, h) = \mathrm{E}_g \ln \frac{g(\mathbf{x})}{h(\mathbf{x})}$$

Theorem 2.4.2 (**Jensen's Inequality**). *$f : D \subset \mathbb{R}^n \to \mathbb{R}$ is a concave function over convex domain D. $\boldsymbol{x}$ is a random vector, taking values on set D, then*

$$\mathrm{E}f(\boldsymbol{x}) \leq f(\mathrm{E}\boldsymbol{x})$$

Theorem 2.4.3 (**Nonegativity of KL Distance**). *Given two density functions g and h, then the Kullback-Leibler distance between them is nonnegative.*

Proof. Apply theorem 2.4.2 to the natural logarithm function, which is concave

$$-\mathcal{D}(g,h) = \mathrm{E}_g\left(\ln \frac{h(\mathbf{x})}{g(\mathbf{x})}\right) \leq \ln\left(\mathrm{E}_g \frac{h(\mathbf{x})}{g(\mathbf{x})}\right) = \ln\left(\int h(\mathbf{x})\, d\mathbf{x}\right) = \ln 1 = 0$$

□

The principle behind EM algorithm rests on the following decomposition, where $g(\mathbf{z})$ is any valid density function

$$\begin{aligned}
\ln f(\mathbf{x};\boldsymbol{\theta}) &= \int g(\mathbf{z}) \ln f(\mathbf{x};\boldsymbol{\theta})\, d\mathbf{z} \\
&= \int g(\mathbf{z}) \ln\left(\frac{f^c(\mathbf{x},\mathbf{z};\boldsymbol{\theta})/g(\mathbf{z})}{f^c_{\mathbf{Z}|\mathbf{x}}(\mathbf{z}|\mathbf{x};\boldsymbol{\theta})/g(\mathbf{z})}\right) d\mathbf{z} \\
&= \int g(\mathbf{z}) \ln\left(\frac{f^c(\mathbf{x},\mathbf{z};\boldsymbol{\theta})}{g(\mathbf{z})}\right) d\mathbf{z} + \mathcal{D}(g, f^c_{\mathbf{Z}|\mathbf{x}}(\cdot;\boldsymbol{\theta})) \\
&= \mathrm{E}_g \ln f^c(\mathbf{x},\mathbf{Z};\boldsymbol{\theta}) - \mathrm{E}_g \ln g(\mathbf{Z}) + \mathcal{D}(g, f^c_{\mathbf{Z}|\mathbf{x}}(\cdot;\boldsymbol{\theta})) &\quad (2.57) \\
&\geq \mathrm{E}_g \ln f^c(\mathbf{x},\mathbf{Z};\boldsymbol{\theta}) - \mathrm{E}_g \ln g(\mathbf{Z}) &\quad (2.58)
\end{aligned}$$

The equality can be achieved in 2.58 if $g \equiv f^c_{\mathbf{Z}|\mathbf{X}}(\cdot;\boldsymbol{\theta})$. Now we are ready to present EM algorithm

Algorithm 2.4.4 (**EM Algorithm**). Suppose $\boldsymbol{\theta}_0$ is an initial guess for the maximiser. The EM algorithm consists of iterating the following steps:

1. **Expectation Steps (E-Step)**: Given the current vector θ_{t-1}, put:

$$g_{t-1}(\mathbf{z}) := f^c_{\mathbf{Z}|\mathbf{x}}(\mathbf{z}|\mathbf{x};\boldsymbol{\theta}_{t-1}) \qquad (2.59)$$

and compute the average complete-data log-likelihood function

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{t-1}) := \mathrm{E}_{g_{t-1}} \ln f^c(\mathbf{x}, \mathbf{Z}; \boldsymbol{\theta}) - \mathrm{E}_{g_{t-1}} \ln g_{t-1}(\mathbf{Z}) \tag{2.60}$$

The function $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{t-1})$ defined above is a minor for $\ln f(\mathbf{x};\boldsymbol{\theta})$ at $\boldsymbol{\theta}_{t-1}$. In practical context, we can ignore the constant part (depend on θ_{t-1} only) associated with minor. For example, $\mathrm{E}_{g_{t-1}} \ln g_{t-1}(\mathbf{Z})$ is a constant and very often, some parts of $\mathrm{E}_{g_{t-1}} \ln f^c(\mathbf{x}, \mathbf{Z}; \boldsymbol{\theta})$ after expansion can be ignored too if they depend on θ_{t-1} only. We abuse the notation by using the same function name $Q(\cdot|\boldsymbol{\theta}_{t-1})$ for a *minor up to a constant* from now on.

2. **Maximisation Step (M-Step)**: Maximise $Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{t-1})$ with respect to $\boldsymbol{\theta}$ to obtain better guess:

$$\boldsymbol{\theta}_t = \operatorname*{argmax}_{\boldsymbol{\theta}} Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{t-1})$$

3. **Stopping Condition**: If, for example, $|l(\boldsymbol{\theta}_T; \mathbf{x}) - l(\boldsymbol{\theta}_{T-1}; \mathbf{x})| < \epsilon$ for some small chosen tolerance ϵ, terminate the algorithm. An alternative terminating condition is using $\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_{T-1}\| < \epsilon$.

A direct consequence of the EM algorithm is that the sequence of log-likelihood values does not decrease with each iteration. In fact, using equality 2.57, and inequality 2.58, we have

$$\begin{aligned}
\ln f(\mathbf{x};\boldsymbol{\theta}_{t-1}) &= \mathrm{E}_{g_t} \ln f^c(\mathbf{x}, \mathbf{Z}; \boldsymbol{\theta}_{t-1}) - \mathrm{E}_{g_t}\ g_t(\mathbf{Z}) \quad (= Q(\boldsymbol{\theta}_{t-1}|\boldsymbol{\theta}_{t-1})) \\
&\leq \mathrm{E}_{g_t} \ln f^c(\mathbf{x}, \mathbf{Z}; \boldsymbol{\theta}_t) - \mathrm{E}_{g_t} g_t(\mathbf{Z}) \qquad (= Q(\boldsymbol{\theta}_t|\boldsymbol{\theta}_{t-1})) \\
&\leq \ln f(\mathbf{x};\boldsymbol{\theta}_t)
\end{aligned} \tag{2.61}$$

It remains an art to choose the convenient latent variables so that it is always easy, usually in the sense that first order conditions are solvable, to obtain the minor's maximiser for the M-step. Apart from analytical and graphical explanation of how EM algorithm works, excellent alternative statistical intuition can be found in Do and Batzoglou (2008)

Chapter 3

Multinomial Maximum Likelihood Estimation

In this chapter, I will propose a statistical model for generating class frequency data and construct the corresponding likelihood function. If the underlying distribution belongs to exponential family or can be mixed from exponential family then recursive formulas for MLE estimator are derivable. Otherwise, numerical optimisation is our last resort. Despite using simple and natural ideas, the involved computations are heavy and have to be performed on computer. The techniques developed in this chapter are separated from the rest of the thesis; therefore, readers are advised to skim the chapter when fully grasping the main line of development.

3.1 Statistical Model for Class Frequency Data

A government agency has the task of summarising individual income data by class frequency method. Suppose there are K income classes: $[t_0, t_1), [t_1, t_2), \dots, [t_{K-1}, t_K)$, where $t_0 = 0, t_K = \infty$ and $t_j < t_{j+1}$ for all j. Exact individual incomes $Z_1, Z_2, \dots, Z_n$ are independently and identically distributed with a continuous distribution $\mathring{F}(\cdot, \boldsymbol{\theta}_0)$ but true parameter vector $\boldsymbol{\theta}_0$ is unknown. Not only that, the variables $Z_1, Z_2, \dots Z_n$ are latent (hidden) variables[1] to an external observer, who can only realise $N_1, N_2, \dots, N_K$

[1] Latent variables is conventionally denoted by Z in EM literature, so I use Z instead of Y

of how many people whose income are within each interval. The probability of an individual whose earnings is in the j^{th} income class $[t_{j-1}, t_j)$ is $p_j = \mathring{F}(t_j; \boldsymbol{\theta}_0) - \mathring{F}(t_{j-1}; \boldsymbol{\theta}_0)$, and the vector of count data $\mathbf{N} = (N_1, N_2, ..., N_K)$ is randomly distributed according to Multinomial distribution with vector of probabilities $\mathbf{p} = (p_1, p_2, \ldots, p_K)$. The model can be summarised as follows:

$$
\begin{aligned}
&\textbf{K income intervals} \quad && [t_0, t_1),\ [t_1, t_2), \ldots, [t_{K-1}, t_K),\ t_0 = 0, t_K = \infty && (3.1)\\
&\textbf{Exact incomes} \quad && Z_1, Z_2, \ldots, Z_n \overset{iid}{\sim} F(\cdot; \boldsymbol{\theta}_0) && (3.2)\\
&\textbf{Probabilities Vector} \quad && \mathbf{p}(\boldsymbol{\theta}) = (p_1(\boldsymbol{\theta}), p_2(\boldsymbol{\theta}), ..., p_K(\boldsymbol{\theta})) && (3.3)\\
& && p_j(\boldsymbol{\theta}) = F(t_j; \boldsymbol{\theta}) - F(t_{j-1}; \boldsymbol{\theta}),\ j = 1, 2, \ldots K && (3.4)\\
&\textbf{Multinomial Counts} \quad && \mathbf{N} = (N_1, N_2, \ldots, N_K) \sim \text{Mult}(n, \mathbf{p}(\boldsymbol{\theta}_0)) && (3.5)
\end{aligned}
$$

The likelihood function and log-likelihood function are:

$$
L(\boldsymbol{\theta}; n_1, ..., n_K) = \frac{n!}{n_1! n_2! ... n_K!} p_1(\boldsymbol{\theta})^{n_1} p_2(\boldsymbol{\theta})^{n_2} ... p_K(\boldsymbol{\theta})^{n_K} \tag{3.6}
$$

$$
l(\boldsymbol{\theta}; n_1, ..., n_K) = \ln\left(\frac{n!}{n_1! n_2! ... n_K!}\right) + n_1 \ln p_1(\boldsymbol{\theta}) + ... + n_K \ln p_K(\boldsymbol{\theta}) \tag{3.7}
$$

Our objective is to find the maximiser $\hat{\boldsymbol{\theta}}$ for the log-likelihood function 3.7 given the realised count data $\mathbf{n} = (n_1, n_2, \ldots, n_K)$. $\hat{\boldsymbol{\theta}}$ is known as the maximum likelihood estimator for $\boldsymbol{\theta}_0$.

3.2 EM Algorithm: Exponential Family

Some well-known members of Exponential Family includes exponential distribution, Log-normal distribution, Gamma distribution. In this section, we apply EM algorithm to produce recursive procedures for estimating the underlying distribution using class frequency data. Then we formulate a general estimation approach for natural exponential family, and exponential family.

Suppose the PDF and CDF for the underlying income distribution are $\mathring{f}(\cdot; \boldsymbol{\theta}_0)$, $\mathring{F}(\cdot; \boldsymbol{\theta}_0)$, respectively. There are K income ranges, whose bounds in ascending order are

$$0 = t_0, t_1, t_2, ..., t_{K-1}, t_K = \infty$$

In particular, the j^{th} income range is $[t_{j-1}, t_j)$. We twist the data generating process in the previous section slightly, by introducing the label variable X that takes value $X = j$ whenever individual random income is in the j^{th} class interval (ie. $Z \in [t_{j-1}, t_j)$). The modified model is presented as follows:

$$\textbf{K income intervals} \quad [t_0, t_1),\ [t_1, t_2), \ldots, [t_{K-1}, t_K),\ t_0 = 0, t_K = \infty \tag{3.8}$$

$$\textbf{Exact incomes} \quad Z_1, Z_2, \ldots, Z_n \overset{iid}{\sim} F(\cdot; \boldsymbol{\theta}_0) \tag{3.9}$$

$$\textbf{Probabilities Vector} \quad \mathbf{p}(\boldsymbol{\theta}) = (p_1(\boldsymbol{\theta}), p_2(\boldsymbol{\theta}), ..., p_K(\boldsymbol{\theta})) \tag{3.10}$$

$$p_i(\boldsymbol{\theta}) = F(t_j; \boldsymbol{\theta}) - F(t_{j-1}; \boldsymbol{\theta}),\ j = 1, 2, \ldots K \tag{3.11}$$

$$\textbf{Deterministic Labelling} \quad X = j \text{ whenever } Z_i \in [t_{j-1}, t_j),\ j = 1, 2, \ldots K \tag{3.12}$$

$$\mathrm{P}(X = j) = p_j(\boldsymbol{\theta}_0) \tag{3.13}$$

$$\textbf{Multinomial Counts} \quad \mathbf{N} = (N_1, N_2, \ldots, N_K) \sim \mathrm{Mult}(n, \mathbf{p}(\boldsymbol{\theta}_0)) \tag{3.14}$$

$$N_i = \sum_{j=1}^{N} I_{\{X_j = j\}},\ j = 1, 2, \ldots K \tag{3.15}$$

The corresponding log-likelihood function for the adjusted model still resembles expression 3.7. Direct maximisation of this log-likelihood function is hard and time-consuming. And to simplify the computation, we augment the likelihood function with a vector of latent variables $\mathbf{Z} = (Z_1, Z_2, \ldots, Z_n)$. Let $c(x, z)$ be the compatibility function that takes 1 if x and z are *compatible*: $X = k$, $Z \in [t_{k-1}, t_k)$ for some k (0 otherwise). Denote the joint compatibility function as $C(\mathbf{x}, \mathbf{z}) = c(x_1, z_1) \ldots c(x_n, z_n)$ which takes 1 if $\mathbf{x}$ and $\mathbf{z}$ are pointwise compatible (and 0 otherwise). Complete-data log-likelihood function is:

$$\ln f^c(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta}) = \sum_{i=1}^{n} \ln \mathring{f}(z_i; \boldsymbol{\theta}) + \ln C(\mathbf{x}, \mathbf{z}) \tag{3.16}$$

To complete the E-step, the conditional distribution of $\mathbf{Z}|\mathbf{X}$ are required. By iid assumption, $Z_i|X_i,\ i = 1, 2, \ldots n$ are conditionally independent. Therefore, what we need is determining the conditional distribution of $X_i|Z_i$, which is easily shown to be a truncated distribution:

$$Z_i|X_i = k \sim \mathring{f}_{[t_{k-1}, t_k)}(\cdot; \boldsymbol{\theta}_0), \quad i = 1, 2, \ldots n \tag{3.17}$$

To implement the EM algorithm, suppose that $\boldsymbol{\theta}_{t-1}$ is the current guess for $\boldsymbol{\theta}_0$, and put

$$g_{t-1}(\mathbf{z}) = f^c_{\mathbf{Z}|\mathbf{X}}(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta}_{t-1}) = \prod_{i=1}^{n} \mathring{f}_{Z_i|X_i}(z_i|x_i; \boldsymbol{\theta}_{t-1}) \tag{3.18}$$

The expected complete-data log-likelihood (minor) in the E-step is then:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{t-1}) = \mathrm{E}_{g_{t-1}} \ln f^c(\mathbf{x}, \mathbf{Z}; \boldsymbol{\theta}) \tag{3.19}$$

The remaining task to be done (M-step) is to find the next guess:

$$\boldsymbol{\theta}_t = \underset{\boldsymbol{\theta}}{\operatorname{argmax}}\ Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{t-1}) \tag{3.20}$$

The striking fact is if the underlying distribution is from exponential family, then the E-step provides closed-form expression for the minor, and the M-step provides closed-form iterative equations for the next guess. Illustrations are made for Exponential, Pareto, Log-Normal, and Gamma distributions, finally culminating in the universal approach for natural exponential family and general exponential family.

Exponential Distribution

Exponential distribution is often used in queuing theory to model service time of agent. In the inequality context, despite poor descriptive power of income distribution, estimated rate λ serves as a simple check whether the income distribution shifts to the right overtime. Suppose the underlying income distribution is Expo(λ). For $z \in [0, \infty)$, the exponential distribution PDF and CDF are:

$$\overset{\circ}{f}(z;\lambda) = \lambda e^{-\lambda z} \tag{3.21}$$

$$\overset{\circ}{F}(z;\lambda) = 1 - e^{-\lambda z} \tag{3.22}$$

Conditional distribution of Z given $X = k$ has a truncated distribution on interval t_{k-1}, t_k with density function[2]:

$$\overset{\circ}{f}_k(z;\lambda) = \frac{\lambda e^{-\lambda z}}{e^{-\lambda t_{k-1}} - e^{-\lambda t_k}}, \quad z \in [t_{k-1}, t_k) \tag{3.23}$$

Complete data log-likelihood function is:

$$\ln f^c(\mathbf{x}, \mathbf{z}; \lambda) = n \ln \lambda - \lambda \sum_{i=1}^{n} z_i + \ln C(\mathbf{x}, \mathbf{z}) \tag{3.24}$$

Therefore, to accomplish E-step, we need:

$$\mathrm{E}(Z|X = k; \lambda) = \frac{1}{\lambda} + t_{k-1} + \frac{e^{\lambda t_{k-1}}(t_k - t_{k-1})}{e^{\lambda t_{k-1}} - e^{\lambda t_k}} \tag{3.25}$$

Algorithm 3.2.1 (EM algorithm for exponential distribution)**.**

1. **E-Step**: Given the current guess λ_{t-1}, propose $Z_i|X_i = k \sim \mathrm{Expo}_{[t_{k-1},t_k)}(\lambda_{t-1})$ where $i = 1, \ldots n$, for data completion, and the minor is given by:

$$Q(\lambda|\lambda_{t-1}) = n \ln \lambda - \lambda \mathrm{E}(\sum_{i=1}^{n} Z_i|\mathbf{X} = \mathbf{x}; \lambda_{t-1}) \tag{3.26}$$

$$= n\,(\ln \lambda - \lambda A_{t-1}) \tag{3.27}$$

where:

$$A_{t-1} = \frac{1}{n} \sum_{i=1}^{n} \mathrm{E}(Z_i|X_i = x_i; \lambda_{t-1}) \tag{3.28}$$

[2]When defining the density function restricted to an interval, the function implicitly takes value 0 outside.

2. **M-Step**: By solving $\frac{d}{d\lambda}Q(\lambda|\lambda_{t-1}) = 0$, the next guess is given by

$$\lambda_t = \frac{1}{A_{t-1}}$$

3. **Terminating condition**: Terminate the algorithm if $|\lambda_T - \lambda_{T-1}| < \epsilon$ (alternatively $|l(\lambda_T; \mathbf{x}) - l(\lambda_{T-1}; \mathbf{x})| < \epsilon$) for a chosen tolerance ϵ.

Pareto Distribution

Consider z_m as a fixed known parameter. Suppose the underlying income distribution is $\mathrm{P}(a, z_m)$ where the PDF and CDF are:

$$\mathring{f}(z; a) = \frac{a z_m^a}{z^{a+1}}, \quad z \geq z_m \tag{3.29}$$

$$\mathring{F}(z; a) = 1 - \left(\frac{z_m}{z}\right)^a, \quad z \geq z_m \tag{3.30}$$

Conditional distribution of Z given $X = k$ has a truncated distribution on interval $[t_{k-1}, t_k)$ with density function:

$$\mathring{f}_k(z; a) = \frac{a z_m^a / z^{a+1}}{(z_m/t_{k-1})^a - (z_m/t_k)^a}, \quad z \in [t_{k-1}, t_k] \tag{3.31}$$

Complete data log-likelihood function is:

$$\ln f^c(\mathbf{x}, \mathbf{z}; a) = n \ln a + na \ln z_m - (a+1) \sum_{i=1}^{n} \ln z_i + \ln C(\mathbf{x}, \mathbf{z}) \tag{3.32}$$

Therefore, to accomplish E-step, we need:

$$\mathrm{E}(\ln Z | X = k) = \frac{z_m^a \left(t_{k-1}^{-a} \left(a \log\left(t_{k-1}\right) + 1\right) - t_k^{-a} \left(a \log\left(t_k\right) + 1\right)\right)}{a((z_m/t_{k-1})^a - (z_m/t_k)^a)} \tag{3.33}$$

Algorithm 3.2.2 (EM algorithm for Pareto distribution)**.**

1. **E-Step**: Given the current guess a_{t-1}, propose $Z_i | X_i = k \sim \mathrm{P}_{[t_{k-1}, t_k)}(a_{t-1}, z_m)$

where $i = 1, \dots n$, for data completion, and the minor is given by:

$$Q(a|a_{t-1}) = n \ln a + na \ln z_m - (a+1)\mathrm{E}\left(\sum_{i=1}^{n} \ln Z_i | \mathbf{X} = \mathbf{x}; a_{t-1}\right) \quad (3.34)$$

$$= n\left[\ln a + a \ln z_m - (a+1)L_{t-1}\right] \quad (3.35)$$

where:

$$L_{t-1} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(\ln Z_i | X_i = x_i; a_{t-1}) \quad (3.36)$$

2. **M-Step**: Solving the First order condition, the next guess is given by:

$$a_t = \frac{1}{L_{t-1} - \ln z_m} \quad (3.37)$$

3. **Terminating condition**: Terminate the algorithm if $|a_T - a_{T-1}| < \epsilon$ (alternatively $|l(a_T; \mathbf{x}) - l(a_{T-1}; \mathbf{x})| < \epsilon$) for a chosen tolerance ϵ

Log-Normal Distribution

Suppose the underlying income distribution is $\mathrm{LN}(\mu, \sigma^2)$. For $z \in [0, \infty)$ the Log-normal distribution PDF and CDF are:

$$\mathring{f}(z; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(\frac{-1}{2\sigma^2}(\ln z - \mu)^2\right)\frac{1}{z} \quad (3.38)$$

$$\mathring{F}(z; \mu, \sigma^2) = \frac{1}{2}\mathrm{erfc}\left(\frac{\mu - \log(z)}{\sqrt{2}\sigma}\right) \quad (3.39)$$

Conditional distribution of Z given $X = k$ has a truncated distribution on interval $[t_{k-1}, t_k)$ with density function:

$$\mathring{f}_k(z; \mu, \sigma^2) = \frac{\frac{1}{\sqrt{2\pi}\sigma}\exp\left(\frac{-1}{2\sigma^2}(\ln z - \mu)^2\right)\frac{1}{z}}{\frac{1}{2}\mathrm{Erfc}\left(\frac{\mu - \ln t_k}{\sqrt{2}\sigma}\right) - \frac{1}{2}\mathrm{Erfc}\left(\frac{\mu - \ln t_{k-1}}{\sqrt{2}\sigma}\right)}, \quad z \in [t_{k-1}, t_k] \quad (3.40)$$

Complete data log-likelihood function is:

$$\ln f^c(\mathbf{x},\mathbf{z};\mu,\sigma^2) = -\frac{n}{2}\ln 2\pi - \frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(\ln z_i)^2 + (\frac{\mu}{\sigma^2}-1)\sum_{i=1}^{n}\ln z_i - \frac{n\mu^2}{2\sigma^2} + \ln C(\mathbf{x},\mathbf{z}) \tag{3.41}$$

Therefore, to accomplish E-step, we need:

$$\mathrm{E}(\ln Z|X=k;\mu,\sigma^2) = \frac{\sqrt{2/\pi}}{\mathrm{erfc}\left(\frac{\mu-\log(t_{k-1})}{\sqrt{2}\sigma}\right) - \mathrm{erfc}\left(\frac{\mu-\log(t_k)}{\sqrt{2}\sigma}\right)} \times \tag{3.42}$$

$$\left[\sqrt{\frac{\pi}{2}}\mu\mathrm{erf}\left(\frac{\mu-\log(t_k)}{\sqrt{2}\sigma}\right) + \sigma t_k^{\frac{\mu}{\sigma^2}} e^{-\frac{\log^2(t_k)+\mu^2}{2\sigma^2}} - \sqrt{\frac{\pi}{2}}\mu\mathrm{erf}\left(\frac{\mu-\log(t_{k-1})}{\sqrt{2}\sigma}\right) - \sigma e^{-\frac{(\mu-\log(t_{k-1}))^2}{2\sigma^2}}\right]$$

$$\mathrm{E}((\ln Z)^2|X=k;\mu,\sigma^2) = \frac{\sqrt{2/\pi}}{\left(\mathrm{erfc}\left(\frac{\mu-\log(t_{k-1})}{\sqrt{2}\sigma}\right) - \mathrm{erfc}\left(\frac{\mu-\log(t_k)}{\sqrt{2}\sigma}\right)\right)} \times \tag{3.43}$$

$$\left[\frac{1}{2}e^{-\frac{\log^2(t_k)+\mu^2}{2\sigma^2}}\left(\sqrt{2\pi}\left(\mu^2+\sigma^2\right)e^{\frac{\log^2(t_k)+\mu^2}{2\sigma^2}}\mathrm{erf}\left(\frac{\mu-\log(t_k)}{\sqrt{2}\sigma}\right) + 2\sigma t_k^{\frac{\mu}{\sigma^2}}\left(\log(t_k)+\mu\right)\right) - \right.$$
$$\left.\frac{1}{2}e^{-\frac{\log^2(t_{k-1})+\mu^2}{2\sigma^2}}\left(\sqrt{2\pi}\left(\mu^2+\sigma^2\right)e^{\frac{\log^2(t_{k-1})+\mu^2}{2\sigma^2}}\mathrm{erf}\left(\frac{\mu-\log(t_{k-1})}{\sqrt{2}\sigma}\right) + 2\sigma t_{k-1}^{\frac{\mu}{\sigma^2}}\left(\log(t_{k-1})+\mu\right)\right)\right]$$

Algorithm 3.2.3 (EM algorithm for Log-Normal Distribution)**.** .

1. **E-Step**: Given the current guess μ_{t-1},σ_{t-1}^2, propose conditional distribution $Z_i|X_i = k \sim \mathrm{LN}_{[t_{k-1},t_k)}(\mu_{t-1},\sigma_{t-1}^2)$ where $i = 1,\ldots n$, for data completion:

$$\begin{aligned} Q(\mu,\sigma^2|\boldsymbol{\theta}_{t-1}) =& -\frac{n}{2}\ln\sigma^2 - \frac{1}{2\sigma^2}\mathrm{E}(\sum_{i=1}^{n}(\ln Z_i)^2|\mathbf{X}=\mathbf{x};\boldsymbol{\theta}_{t-1}) \\ &+ (\frac{\mu}{\sigma^2}-1)\mathrm{E}(\sum_{i=1}^{n}\ln Z_i|\mathbf{X}=\mathbf{x};\boldsymbol{\theta}_{t-1}) - \frac{n\mu^2}{2\sigma^2} \\ =& n\left(\frac{-1}{2}\ln\sigma^2 - \frac{A_{t-1}}{2\sigma^2} + \frac{\mu B_{t-1}}{\sigma^2} - B_{t-1} - \frac{\mu^2}{2\sigma^2}\right) \end{aligned} \tag{3.44}$$

where:

$$A_{t-1} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}((\ln Z_i)^2|X_i = x_i;\boldsymbol{\theta}_{t-1}) \tag{3.45}$$

$$B_{t-1} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(\ln Z_i|X_i = x_i;\boldsymbol{\theta}_{t-1}) \tag{3.46}$$

2. **M-Step**: The next guess is given by

$$\begin{aligned}\mu_t &= B_{t-1}\\ \sigma_t^2 &= A_{t-1} - B_{t-1}^2\end{aligned}$$

3. **Terminating condition**: Terminate the algorithm if $\|(\mu_T, \sigma_T^2) - (\mu_{T-1}, \sigma_{T-1}^2)\| < \epsilon$ (alternatively $|l(\mu_T, \sigma_T^2; \mathbf{x}) - l(\mu_{T-1}, \sigma_{T-1}^2; \mathbf{x})| < \epsilon$) for a chosen tolerance ϵ

Gamma Distribution

Suppose the underlying income distribution is Gamma(a, λ). For $z \in [0, \infty)$, the Gamma distribution PDF and CDF are:

$$\overset{\circ}{f}(z; a, \lambda) = \tfrac{\lambda^a z^{a-1} e^{-\lambda z}}{\Gamma(a)} \tag{3.47}$$

$$\overset{\circ}{F}(z; a, \lambda) = Q(a, 0, \lambda z) \tag{3.48}$$

Conditional distribution of Z given $X = k$ has a truncated distribution on interval $[t_{k-1}, t_k)$ with density function:

$$\overset{\circ}{f}_k(z; a, \lambda) = \frac{\lambda^a z^{a-1} e^{-\lambda z}}{\Gamma(a)\left(Q(a, 0, \lambda t_k) - Q(a, 0, \lambda t_{k-1})\right)}, \quad z \in [t_{k-1}, t_k) \tag{3.49}$$

Complete data log-likelihood function is:

$$\ln f^c(\mathbf{x}, \mathbf{z}; a, \lambda) = na \ln \lambda - n \ln \Gamma(a) - \lambda \sum_{i=1}^{n} z_i + (a-1) \sum_{i=1}^{n} \ln z_i + \ln C(\mathbf{x}, \mathbf{z}) \tag{3.50}$$

Therefore, to accomplish E-step, we need:

$$\mathrm{E}(Z|X = k; a, \lambda) = \frac{\lambda^{a-1}\left[t_{k-1}^a \left(\lambda t_{k-1}\right)^{-a} \Gamma\left(a+1, \lambda t_{k-1}\right) - t_k^a \left(\lambda t_k\right)^{-a} \Gamma\left(a+1, \lambda t_k\right)\right]}{\Gamma(a)\left(Q\left(a, 0, \lambda t_k\right) - Q\left(a, 0, \lambda t_{k-1}\right)\right)} \tag{3.51}$$

$$\mathrm{E}(\ln Z|X=k;a,\lambda) = \Bigg[\frac{\lambda^a t_{k-1}^a\, {}_2F_2\left(a,a;a+1,a+1;-\lambda t_{k-1}\right)}{a^2\Gamma(a)} -$$
$$\frac{\lambda^a t_k^a\, {}_2F_2\left(a,a;a+1,a+1;-\lambda t_k\right)}{a^2\Gamma(a)} + \frac{\lambda^a t_{k-1}^a \log\left(t_{k-1}\right)\left(\lambda t_{k-1}\right)^{-a}\Gamma\left(a,\lambda t_{k-1}\right)}{\Gamma(a)} -$$
$$\frac{\lambda^a t_k^a \log\left(t_k\right)\left(\lambda t_k\right)^{-a}\Gamma\left(a,\lambda t_k\right)}{\Gamma(a)} + \lambda^a t_k^a \log\left(t_k\right)\left(\lambda t_k\right)^{-a} - \lambda^a t_{k-1}^a \log\left(t_{k-1}\right)\left(\lambda t_{k-1}\right)^{-a}\Bigg] \times$$
$$\frac{1}{Q\left(a,0,\lambda t_k\right) - Q\left(a,0,\lambda t_{k-1}\right)} \tag{3.52}$$

Algorithm 3.2.4 (EM algorithm for Gamma distribution)**.**

1. **E-Step**: Given the current guess a_{t-1}, λ_{t-1}, propose conditional distribution $Z_i|X_i = k \sim \text{Gamma}_{[t_{k-1},t_k)}(a_{t-1}, \lambda_{t-1})$ where $i = 1, \ldots n$, for data completion, and the minor is given by:

$$\begin{aligned} Q(a,\lambda;\boldsymbol{\theta}_{t-1}) =& na\ln\lambda - n\ln\Gamma(a) - \lambda\mathrm{E}(\sum_{i=1}^{n} Z_i|\mathbf{X}=\mathbf{x};\boldsymbol{\theta}_{t-1}) \\ &+ (a-1)\mathrm{E}(\sum_{i=1}^{n}\ln Z_i|\mathbf{X}=\mathbf{x};\boldsymbol{\theta}_{t-1}) \qquad (3.53)\\ =& n(a\ln\lambda - \ln\Gamma(a) - \lambda A_{t-1} + (a-1)B_{t-1}) \qquad (3.54)\end{aligned}$$

   where:

$$A_{t-1} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(Z_i|X_i = x_i;\lambda_{t-1}) \tag{3.55}$$
$$B_{t-1} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(\ln Z_i|X_i = x_i;\lambda_{t-1}) \tag{3.56}$$

2. **M-Step**: Solving the first order condition, the next guess is given implicitly by

$$\ln\lambda_t - \psi(a_t) + B_{t-1} = 0 \tag{3.57}$$
$$\frac{a_t}{\lambda_t} - A_{t-1} = 0 \tag{3.58}$$

Put $\nu(x) = \ln x - \psi(x)$, then the bijective function $\nu : (0, \infty) \to (0, \infty)$ is a strictly decreasing function, exhibiting the inverse function $\nu^{-1} : (0, \infty) \to (0, \infty)$. This simple looking function is well-bahaved and have many nice properties. However, I can not find any reference for it, so I name it ν for "no name". With the new function on hand, the direct solution for the system of equations (3.57 and 3.58) is:

$$a_t = \nu^{-1}(\ln A_{t-1} - B_{t-1}) \tag{3.59}$$

$$\lambda_t = \frac{\nu^{-1}(\ln A_{t-1} - B_{t-1})}{A_{t-1}} \tag{3.60}$$

3. **Terminating condition**: Terminate the algorithm if $\|(a_T, \lambda_T) - (a_{T-1}, \lambda_{T-1}\| < \epsilon$ (alternatively $|l(a_T, \lambda_T; x) - l(a_{T-1}, \lambda_{T-1}; \mathbf{x})| < \epsilon$) for a chosen tolerance ϵ

Natural Exponential Family

Suppose the underlying distribution is from Natural exponential family, its PDF and CDF:

$$\overset{\circ}{f}(z; \boldsymbol{\theta}) = c(\boldsymbol{\theta}) \exp\left(\sum_{j=1}^{d} T_j(z)\theta_j\right) h(z) \tag{3.61}$$

$$\overset{\circ}{F}(z; \boldsymbol{\theta}) = \text{have expression} \tag{3.62}$$

Conditional distribution of Z given $X = k$ has a truncated distribution on interval $[t_{k-1}, t_k)$ with density function:

$$\overset{\circ}{f}_k(z) = \frac{c(\boldsymbol{\theta}) \exp\left(\sum_{j=1}^{d} T_j(z)\theta_j\right) h(z)}{\overset{\circ}{F}(t_k; \boldsymbol{\theta}) - \overset{\circ}{F}(t_{k-1}; \boldsymbol{\theta})}, \quad z \in [t_{k-1}, t_k) \tag{3.63}$$

Complete data log-likelihood function is given by

$$\ln f^c(\mathbf{x}, \mathbf{z}; \lambda) = n \ln c(\boldsymbol{\theta}) + \sum_{j=1}^{d} \theta_j \left(\sum_{i=1}^{n} T_j(z_i)\right) + \sum_{i=1}^{n} \ln h(z_i) + \ln C(\mathbf{x}, \mathbf{z}) \tag{3.64}$$

To accomplish E-step, we use lemma 2.3.3 telling that a truncated exponential family is still exponential family, and lemma 2.3.4 telling how to compute expectation of sufficient statistics associated with exponential family:

$$\mathrm{E}(\mathbf{T}(Z)|X=k;\boldsymbol{\theta}) = -\frac{\nabla_{\boldsymbol{\theta}}\{c(\boldsymbol{\theta})/(\mathring{F}(t_k;\boldsymbol{\theta})-\mathring{F}(t_{k-1};\boldsymbol{\theta}))\}}{c(\boldsymbol{\theta})/(\mathring{F}(t_k;\boldsymbol{\theta})-\mathring{F}(t_{k-1};\boldsymbol{\theta}))} \tag{3.65}$$

Algorithm 3.2.5 (EM algorithm for natural exponential family)**.**

1. **E-Step**: Given the current guess $\boldsymbol{\theta}_{t-1}$, propose $Z_i|X_i = k \sim \mathrm{NE}_{[t_{k-1},t_k)}(\boldsymbol{\theta}_{t-1})$ where $i = 1, \ldots n$, for data completion, and the minor is given by:

$$Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{t-1}) = n\ln c(\boldsymbol{\theta}) + \sum_{j=1}^{d}\theta_j \mathrm{E}\left(\sum_{i=1}^{n} T_j(Z_i)|\mathbf{X}=\mathbf{x};\boldsymbol{\theta}_{t-1}\right) \tag{3.66}$$

$$= n\left(\ln c(\boldsymbol{\theta}) + \sum_{j=1}^{d}\theta_j A_{j,t-1}\right) \tag{3.67}$$

where:

$$A_{j,t-1} = \frac{1}{n}\sum_{i=1}^{n}\mathrm{E}(T_j(Z_i)|X_i = x_i;\boldsymbol{\theta}_{t-1}), \quad j = 1, 2, \ldots d \tag{3.68}$$

2. **M-Step**: The next guess $\boldsymbol{\theta}_t$ is given implicitly by

$$\frac{\nabla c(\boldsymbol{\theta}_t)}{c(\boldsymbol{\theta}_t)} = \begin{bmatrix} A_{1,t-1} \\ A_{2,t-1} \\ \vdots \\ A_{m-1,t-1} \\ A_{m,t-1} \end{bmatrix} \tag{3.69}$$

3. **Terminating condition**: Terminate the algorithm if $\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_{T-1}\| < \epsilon$ for a chosen tolerance ϵ. Alternative, terminate if $|l(\boldsymbol{\theta}_T;\mathbf{x}) - l(\boldsymbol{\theta}_{T-1};\mathbf{x})| < \epsilon$.

General Exponential Family

Suppose the underlying distribution is from exponential family. Its PDF and CDF are:

$$\overset{\circ}{f}(z;\boldsymbol{\theta}) = c(\boldsymbol{\theta}) \exp\left(\sum_{j=1}^{m} T_j(z)\eta_j(\boldsymbol{\theta})\right) h(z) \tag{3.70}$$

$$\overset{\circ}{F}(z;\boldsymbol{\theta}) = \text{either have expression or not} \tag{3.71}$$

Let $\propto$ denote proportionality. Conditional distribution of Z given $X = k$ has a truncated distribution on interval $[t_{k-1}, t_k)$ with density function proportional to:

$$\overset{\circ}{f}_k(z) \propto \exp\left(\sum_{j=1}^{m} T_j(z)\eta_j(\boldsymbol{\theta})\right) h(z), \quad z \in [t_{k-1}, t_k) \tag{3.72}$$

Complete data log-likelihood function is given by

$$\ln f^c(\mathbf{x}, \mathbf{z}; \lambda) = n \ln c(\boldsymbol{\theta}) + \sum_{j=1}^{m} \eta_j(\boldsymbol{\theta}) \left(\sum_{i=1}^{n} T_j(z_i)\right) + \sum_{i=1}^{n} \ln h(z_i) + \ln C(\mathbf{x}, \mathbf{z}) \tag{3.73}$$

To accomplish E-step, we use numerical integration to obtain the normalising constant for the truncated density in expression 3.72 and m integrals to compute expectation of m sufficient statistics:

$$\mathrm{E}(\mathbf{T}(Z)|X = k; \boldsymbol{\theta}) = \text{numerical integration} \tag{3.74}$$

Algorithm 3.2.6 (EM algorithm for natural exponential family)**.** .

1. **E-Step**: Given the current guess $\boldsymbol{\theta}_{t-1}$, propose $Z_i|X_i = k \sim \mathrm{NE}_{[t_{k-1},t_k)}(\boldsymbol{\theta}_{t-1})$ where $i = 1, \dots n$, for data completion, and the minor is given by:

$$\begin{aligned} Q(\boldsymbol{\theta}|\boldsymbol{\theta}_{t-1}; \boldsymbol{\theta}) &= n \ln c(\boldsymbol{\theta}) + \sum_{j=1}^{m} \eta_j(\boldsymbol{\theta}) \mathrm{E}\left(\sum_{i=1}^{n} T_j(Z_i)|\mathbf{X} = \mathbf{x}; \boldsymbol{\theta}_{t-1}\right) \\ &= n \left(\ln c(\boldsymbol{\theta}) + \sum_{j=1}^{m} \eta_j(\boldsymbol{\theta}) A_{j,t-1}\right) \end{aligned}$$

where:

$$A_{j,t-1} = \frac{1}{n}\sum_{i=1}^{n} \mathrm{E}(T_j(Z_i)|X_i = x_i; \boldsymbol{\theta}_{t-1}), \quad j = 1, 2, \ldots m \tag{3.75}$$

2. **M-Step**: The next guess $\boldsymbol{\theta}_t$ is given implicitly by the first order condition:

$$\frac{\nabla c(\boldsymbol{\theta}_t)}{c(\boldsymbol{\theta}_t)} = -\sum_{j=1}^{m} \nabla \eta_j(\boldsymbol{\theta}) A_{j,t-1} \tag{3.76}$$

3. **Terminating condition**: Terminate the algorithm if $\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_{T-1}\| < \epsilon$ for a chosen tolerance ϵ. Alternatively, terminate if $|l(\boldsymbol{\theta}_T; \mathbf{x}) - l(\boldsymbol{\theta}_{T-1}; \mathbf{x})| < \epsilon$.

3.3 EM Algorithm: Mixture Model from Exponential Family

A more sophisticated way in modelling income distribution is using mixture models, especially when the data indicate bimodal or multimodal behaviours. In this section, we will explore the use of EM algorithm in constructing recursive maximum likelihood estimator for the mixture of two distributions from exponential family using class frequency data. The extension to the mixture of more than two distributions is straightforward.

Log-Normal and Pareto

Lubrano and Ndoye (2016) argued neither Pareto nor Log-normal is adequate in describing wealth and income distribution. One of his finding is the distribution is better explained by a mixture between Pareto and Log-Normal. In the paper, the authors ran the model on income microdata, and estimated the size distribution of income by a mixture of two Log-normal distributions. With class frequency data, I will show the recursive procedure for MLE is feasible. First, the data generating process for mixture model is:

$$\textbf{Branching} \quad D_1, D_2, ..., D_n \overset{iid}{\sim} \text{Bern}(\rho) \tag{3.77}$$

$$\textbf{Pareto Branch} \quad Z_i|D_i = 0 \overset{iid}{\sim} \text{P}(a, z_m) \text{ fixed } \mathrm{z_m} \tag{3.78}$$

$$\textbf{Log-Normal Branch} \quad Z_i|D_i = 1 \overset{iid}{\sim} \text{LN}(\mu, \sigma^2) \tag{3.79}$$

$$\textbf{Deterministic Labelling} \quad X_i = m(Z_i) \tag{3.80}$$

It is easy to verify that Z_i is independently and identically distributed with a mixed density function:

$$\overset{\circ}{f}(z; \rho, \mu, \sigma^2, a) = (1-\rho)\overset{\circ}{f}_0(z; a) + \rho\overset{\circ}{f}_1(z; \mu, \sigma^2) \tag{3.81}$$

where $\overset{\circ}{f}_0$ and $\overset{\circ}{f}_1$ are density functions of Pareto and Log-normal distribution defined in equation 3.29 and equation 3.38. To shorten the notation, set $\boldsymbol{\theta} = (\rho, a, \mu, \sigma^2)$ as the set of parameters. Complete-data likelihood function is:

$$\begin{aligned} f^c(\mathbf{x}, \mathbf{z}, \mathbf{d}; \boldsymbol{\theta}) &= \prod_{i=1}^{n} \overset{\circ}{p}(d_i; \boldsymbol{\theta}) \overset{\circ}{f}(z_i|d_i; \boldsymbol{\theta}) c(x_i, z_i, d_i) \\ &= \prod_{i=1}^{n} \rho^{d_i}(1-\rho)^{1-d_i} \left(\frac{1}{\sqrt{2\pi}\sigma z_i} \exp\left(\frac{-1}{2\sigma^2}(\ln z_i - \mu)^2 \right) \right)^{d_i} \left(a \frac{z_{min}^a}{z_i^{a+1}} \right)^{1-d_i} c(x_i, z_i, d_i) \end{aligned} \tag{3.82}$$

The complete data log-likelihood function is:

$$\begin{aligned} \ln f^c(\mathbf{x}, \mathbf{z}, \mathbf{d}; \boldsymbol{\theta}) = &\left(\sum_{i=1}^{n} d_i \right) \ln \rho + \left(n - \sum_{i=1}^{n} d_i \right) \ln(1-\rho) - \frac{1}{2} \ln(2\pi) \left(\sum_{i=1}^{n} d_i \right) \\ &- \frac{1}{2} \ln \sigma^2 \left(\sum_{i=1}^{n} d_i \right) - \left(\sum_{i=1}^{n} d_i \ln z_i \right) - \frac{1}{2\sigma^2} \left(\sum_{i=1}^{n} d_i (\ln z_i - \mu)^2 \right) \\ &+ \left(n - \sum_{i=1}^{n} d_i \right) \ln a + \left(n - \sum_{i=1}^{n} d_i \right) a \ln z_{min} \\ &- (a+1) \left(\sum_{i=1}^{n} (1-d_i) \ln z_i \right) + C(\mathbf{x}, \mathbf{z}, \mathbf{d}) \end{aligned} \tag{3.83}$$

Conditional distribution of latent variables Z, D can be derived using Bayes Rule, and the result are summarised below:

$$D|X = k; \boldsymbol{\theta} \sim \text{Bern}\left(\frac{\rho(F_1(t_k; \boldsymbol{\theta}) - F_1(t_{k-1}; \boldsymbol{\theta}))}{\rho(F_1(t_k; \boldsymbol{\theta}) - F_1(t_{k-1}; \boldsymbol{\theta})) + (1-\rho)(F_0(t_k; \boldsymbol{\theta}) - F_0(t_{k-1}; \boldsymbol{\theta}))}\right) \quad (3.84)$$

$$Z|X = k, D = 0; \boldsymbol{\theta} \sim \text{Pareto}_{[t_{k-1}, t_k)}(a) \quad (3.85)$$

$$Z|X = k, D = 1; \boldsymbol{\theta} \sim \text{LN}_{[t_{k-1}, t_k)}(\mu, \sigma^2) \quad (3.86)$$

The conditional distributions specified above, together with formula 3.42, 3.43, 3.33 allow us to complete the E-step.

Algorithm 3.3.1 (EM algorithm for Lognormal-Pareto mixture model)**.** .

1. **E-Step**: Given the current guess $\boldsymbol{\theta}_{t-1} = (\rho_{t-1}, a_{t-1}, \mu_{t-1}, \sigma^2_{t-1})$, propose 3.84, 3.85, 3.86 for data completion. The minor is then given by:

$$\begin{aligned}\frac{1}{n}Q(\rho, a, \mu, \sigma^2|\boldsymbol{\theta}_{t-1}) = \frac{1}{n}\text{E}_{t-1}\{\ln f^c(\mathbf{x}, \mathbf{Z}, \mathbf{D}; \boldsymbol{\theta})\} \\ &= S_{t-1}\ln\rho + \ln(1-\rho) - S_{t-1}\ln(1-\rho) - \frac{1}{2}\ln(2\pi)S_{t-1} \\ &- \frac{1}{2}\ln\sigma^2 S_{t-1} - N_{t-1} - \frac{1}{2\sigma^2}M_{t-1} + \frac{\mu}{\sigma^2}N_{t-1} - \frac{\mu^2}{2\sigma^2}S_{t-1} \\ &+ \ln a - S_{t-1}\ln a + a\ln(z_{min}) - S_{t-1}a\ln(z_{min}) \\ &- (a+1)L_{t-1} + (a+1)N_{t-1} \end{aligned} \quad (3.87)$$

where

$$S_{t-1} = \frac{1}{n}\text{E}\left(\sum_{i=1}^{n} D_i|\mathbf{X} = \mathbf{x}; \boldsymbol{\theta}_{t-1}\right) \quad (3.88)$$

$$M_{t-1} = \frac{1}{n}\text{E}\left(\sum_{i=1}^{n} D_i(\ln Z_i)^2|\mathbf{X} = \mathbf{x}; \boldsymbol{\theta}_{t-1}\right) \quad (3.89)$$

$$N_{t-1} = \frac{1}{n}\text{E}\left(\sum_{i=1}^{n} D_i \ln Z_i|\mathbf{X} = \mathbf{x}; \boldsymbol{\theta}_{t-1}\right) \quad (3.90)$$

$$L_{t-1} = \frac{1}{n}\text{E}\left(\sum_{i=1}^{n} \ln Z_i|\mathbf{X} = \mathbf{x}; \boldsymbol{\theta}_{t-1}\right) \quad (3.91)$$

2. **M-Step**: Solving First order condition, the next improved guess is given by:

$$\rho_t = S_{t-1} \tag{3.92}$$
$$a_t = \frac{1 - S_{t-1}}{L_{t-1} - N_{t-1} - \ln(z_{min})(1 - S_{t-1})} \tag{3.93}$$
$$\mu_t = \frac{N_{t-1}}{S_{t-1}} \tag{3.94}$$
$$\sigma_t^2 = \frac{M_{t-1}}{S_{t-1}} - \frac{N_{t-1}^2}{S_{t-1}^2} \tag{3.95}$$

3. **Terminating condition**: Terminate the algorithm if $\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_{T-1}\| < \epsilon$ for a chosen tolerance ϵ. Alternative, terminate if $|l(\boldsymbol{\theta}_T; \mathbf{x}) - l(\boldsymbol{\theta}_{T-1}; \mathbf{x})| < \epsilon$.

Double Gamma

A gamma mixture with two or three components is proposed by Chotikapanich and Griffiths (2008) to estimate the size distribution of Canadian incomes using microdata. In their paper, the authors used a Bayesian procedure for estimation with slightly informative priors. The following procedure I lay out here is a MLE substitute for their proposed estimation technique using class frequency data instead. First, we specify the data generating process:

$$\begin{aligned}
&\textbf{Branching} \quad D_1, D_2, ..., D_n \overset{iid}{\sim} \text{Bern}(\rho) \\
&\textbf{Gamma Branch} \quad Z_i | D_i = 0 \overset{iid}{\sim} \text{Gamma}(a_0, \lambda_0) \\
&\textbf{Gamma Branch} \quad Z_i | D_i = 1 \overset{iid}{\sim} \text{Gamma}(a_1, \lambda_1) \\
&\textbf{Deterministic Labelling} \quad X_i = m(Z_i)
\end{aligned} \tag{3.96}$$

It is easy to verify that Z_i is independently and identically distributed with a mixed density function:

$$\mathring{f}(z; \rho, a_0, \lambda_0, a_1, \lambda_1) = (1 - \rho)\mathring{f}_0(z; a_0, \lambda_0) + \rho \mathring{f}_1(z; a_1, \lambda_1) \tag{3.97}$$

where $\mathring{f}_0$ and $\mathring{f}_1$ are density functions of Gamma distribution defined in 3.47. To shorten the notation, set $\boldsymbol{\theta} = (\rho, a_0, \lambda_0, a_1, \lambda_1)$ as the vector of parameters. The complete-data likelihood function is:

$$f^c(\mathbf{x},\mathbf{z},\mathbf{d};\boldsymbol{\theta}) = \prod_{i=1}^{n} \mathring{p}(d_i;\boldsymbol{\theta})\mathring{f}(z_i|d_i;\boldsymbol{\theta})c(x_i,z_i,d_i)$$

$$= \prod_{i=1}^{n} \rho^{d_i}(1-\rho)^{1-d_i} \left(\frac{\lambda_1^{a_1} z_i^{a_1-1} e^{-\lambda_1 z_1}}{\Gamma(a_1)}\right)^{d_i} \left(\frac{\lambda_0^{a_0} z_i^{a_0-1} e^{-\lambda_0 z_1}}{\Gamma(a_0)}\right)^{1-d_i} c(x_i,z_i,d_i) \tag{3.98}$$

The complete data log-likelihood function is:

$$\begin{aligned}
\ln f^c(\mathbf{x},\mathbf{z},\mathbf{d};\boldsymbol{\theta}) &= \left(\sum_{i=1}^{n} d_i\right)\ln\rho + \left(n-\sum_{i=1}^{n} d_i\right)\ln(1-\rho) - \left(\sum_{i=1}^{n} d_i\right)\ln\Gamma(a_1) \\
&+ \left(\sum_{i=1}^{n} d_i\right) a_1\ln\lambda_1 + (a_1-1)\left(\sum_{i=1}^{n} d_i\ln z_i\right) - \lambda_1\left(\sum_{i=1}^{n} d_i z_i\right) \\
&- \left(n-\sum_{i=1}^{n} d_i\right)\ln\Gamma(a_0) + \left(n-\sum_{i=1}^{n} d_i\right) a_0\ln\lambda_0 + (a_0-1)\left(\sum_{i=1}^{n}\ln z_i\right) \\
&- (a_0-1)\left(\sum_{i=1}^{n} d_i\ln z_i\right) - \lambda_0\left(\sum_{i=1}^{n} z_i\right) + \lambda_0\left(\sum_{i=1}^{n} d_i z_i\right) \\
&+ C(\mathbf{x},\mathbf{z},\mathbf{d})
\end{aligned} \tag{3.99}$$

Conditional distribution of latent variables Z, D can be derived using Bayes Rule, and the result are summarised below:

$$D|X=k;\boldsymbol{\theta} \sim \text{Bern}\left(\frac{\rho(F_1(t_k;\boldsymbol{\theta}) - F_1(t_{k-1};\boldsymbol{\theta}))}{\rho(F_1(t_k;\boldsymbol{\theta}) - F_1(t_{k-1};\boldsymbol{\theta})) + (1-\rho)(F_0(t_k;\boldsymbol{\theta}) - F_0(t_{k-1};\boldsymbol{\theta})}\right) \tag{3.100}$$

$$Z|X=k, D=0;\boldsymbol{\theta} \sim \text{Gamma}_{[t_{k-1},t_k)}(a_0,\lambda_0) \tag{3.101}$$

$$Z|X=k, D=1;\boldsymbol{\theta} \sim \text{Gamma}_{[t_{k-1},t_k)}(a_1,\lambda_1) \tag{3.102}$$

The conditional distributions specified above together with formula in 3.51, 3.52 allow us to complete the E-step

Algorithm 3.3.2 (EM algorithm for double Gamma mixture model)**.** .

1. **E-Step**: Given the current guess $\boldsymbol{\theta}_{t-1} = (\rho_{t-1}, a_{0,t-1}, \lambda_{0,t-1}, a_{1,t-1}, \lambda_{1,t-1})$, propose conditional distributions 3.100, 3.101, 3.102 for data completion, the minor

for log-likelihood function at $\boldsymbol{\theta}_{t-1}$ is:

$$\begin{aligned}\frac{1}{n}Q(\rho, a_0, \lambda_0, a_1, \lambda_1|\boldsymbol{\theta}_{t-1}) &= \frac{1}{n}\mathrm{E}_{t-1}\{\ln f^c(\mathbf{x}, \mathbf{Z}, \mathbf{D}; \boldsymbol{\theta})\} \\ &= S_{t-1}\ln\rho + (1-S_{t-1})\ln(1-\rho) - S_{t-1}\ln\Gamma(a_1) \\ &+ S_{t-1}a_1\ln\lambda_1 + (a_1-1)N_{t-1} - \lambda_1 P_{t-1} \\ &- (1-S_{t-1})\ln\Gamma(a_0) + (1-S_{t-1})a_0\ln\lambda_0 + (a_0-1)L_{t-1} \\ &- (a_0-1)N_{t-1} - \lambda_0 V_{t-1} + \lambda_0 P_{t-1} \end{aligned} \quad (3.103)$$

where

$$S_{t-1} = \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n} D_i|\mathbf{X}=\mathbf{x};\boldsymbol{\theta}_{t-1}\right) \quad (3.104)$$

$$N_{t-1} = \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n} D_i\ln Z_i|\mathbf{X}=\mathbf{x};\boldsymbol{\theta}_{t-1}\right) \quad (3.105)$$

$$L_{t-1} = \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n} \ln Z_i|\mathbf{X}=\mathbf{x};\boldsymbol{\theta}_{t-1}\right) \quad (3.106)$$

$$P_{t-1} = \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n} D_i Z_i|\mathbf{X}=\mathbf{x};\boldsymbol{\theta}_{t-1}\right) \quad (3.107)$$

$$V_{t-1} = \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n} Z_i|\mathbf{X}=\mathbf{x};\boldsymbol{\theta}_{t-1}\right) \quad (3.108)$$

2. **M-Step**: Solving First order condition, the next improved guess is given by:

$$\rho_t = S_{t-1} \quad (3.109)$$

$$a_{0,t} = \nu^{-1}\left(\ln\left(\frac{V_{t-1}-P_{t-1}}{1-S_{t-1}}\right) - \frac{L_{t-1}-N_{t-1}}{1-S_{t-1}}\right) \quad (3.110)$$

$$\lambda_{0,t} = \frac{a_{0,t}\,(1-S_{t-1})}{V_{t-1}-P_{t-1}} \quad (3.111)$$

$$a_{1,t} = \nu^{-1}\left(\ln\left(\frac{P_{t-1}}{S_{t-1}}\right) - \frac{N_{t-1}}{S_{t-1}}\right) \quad (3.112)$$

$$\lambda_{1,t} = \frac{a_{1,t}\,S_{t-1}}{P_{t-1}} \quad (3.113)$$

3. **Terminating condition**: Terminate the algorithm if $\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_{T-1}\| < \epsilon$ for a chosen tolerance ϵ. Alternatively, terminate if $|l(\boldsymbol{\theta}_T;\mathbf{x}) - l(\boldsymbol{\theta}_{T-1};\mathbf{x})| < \epsilon$.

Double Log-Normal

Lubrano and Ndoye (2016) employed the Bayesian approach using a mixture of Log-Normal distributions for UK income microdata. The author argued the mixture of Log-Normal distributions is a powerful descriptive model for the UK income distribution. The following procedure I lay out here is a MLE substitute for their proposed estimation technique using class frequency data instead. First, we specify the data generating process for the mixture model of income:

$$\begin{aligned}
&\textbf{Branching} \quad D_1, D_2, ..., D_n \overset{iid}{\sim} \text{Bern}(\rho) \\
&\textbf{Log-Normal Branch} \quad Z_i|D_i = 0 \overset{iid}{\sim} \text{LN}(\mu_0, \sigma_0^2) \\
&\textbf{Log-Normal Branch} \quad Z_i|D_i = 1 \overset{iid}{\sim} \text{LN}(\mu_1, \sigma_1^2) \\
&\textbf{Deterministic Labelling} \quad X_i = m(Z_i)
\end{aligned} \tag{3.114}$$

It is easy to verify that Z_i is independently and identically distributed with a mixed density function:

$$\overset{\circ}{f}(z; \mu, \sigma^2, a) = (1-\rho)\overset{\circ}{f}_0(z; \mu_0, \sigma_0^2) + \rho \overset{\circ}{f}_1(z; \mu_1, \sigma_1^2) \tag{3.115}$$

where $\overset{\circ}{f}_0$ and $\overset{\circ}{f}_1$ are density functions of Log-Normal distribution defined in 3.38.

To shorten the notation, set $\boldsymbol{\theta} = (\rho, a, \mu, \sigma^2)$ as the set of parameters. The complete-data likelihood function is:

$$\begin{aligned}
f^c(\mathbf{x}, \mathbf{z}, \mathbf{d}; \boldsymbol{\theta}) &= \prod_{i=1}^{n} \overset{\circ}{p}(d_i; \boldsymbol{\theta}) \overset{\circ}{f}(z_i | d_i; \boldsymbol{\theta}) c(x_i, z_i, d_i) \\
&= \prod_{i=1}^{n} \rho^{d_i} (1-\rho)^{1-d_i} \left(\frac{1}{\sqrt{2\pi}\sigma_1 z_i} \exp\left(\frac{-1}{2\sigma_1^2} (\ln z_i - \mu_1)^2 \right) \right)^{d_i} \times \\
&\left(\frac{1}{\sqrt{2\pi}\sigma_0 z_i} \exp\left(\frac{-1}{2\sigma_0^2} (\ln z_i - \mu_0)^2 \right) \right)^{1-d_i} c(x_i, z_i, d_i)
\end{aligned} \tag{3.116}$$

By taking the logarithm of the above expression, we obtain the complete-data log-likelihood function as follows:

$$
\begin{aligned}
\ln f^c(\mathbf{x}, \mathbf{z}, \mathbf{d}; \boldsymbol{\theta}) = & \left(\sum_{i=1}^{n} d_i\right) \ln \rho + \left(n - \sum_{i=1}^{n} d_i\right) \ln(1-\rho) - \frac{1}{2}\ln(2\pi)\left(\sum_{i=1}^{n} d_i\right) \\
& - \frac{1}{2}\ln \sigma_1^2 \left(\sum_{i=1}^{n} d_i\right) - \left(\sum_{i=1}^{n} d_i \ln z_i\right) - \frac{1}{2\sigma_1^2}\left(\sum_{i=1}^{n} d_i (\ln z_i - \mu_1)^2\right) \\
& - \frac{1}{2}\left(n - \sum_{i=1}^{n} d_i\right)\ln(2\pi) - \frac{1}{2}\left(n - \sum_{i=1}^{n} d_i\right)\ln \sigma_0^2 \\
& - \sum_{i=1}^{n}(1-d_i)\ln z_i - \frac{1}{2\sigma_0^2}\left(\sum_{i=1}^{n}(1-d_i)(\ln z_i - \mu_0)^2\right) + C(\mathbf{x}, \mathbf{z}, \mathbf{d}) \quad (3.117)
\end{aligned}
$$

Conditional distribution of latent variables Z, D can be derived using Bayes Rule, and the result are summarised below:

$$
D|X = k; \boldsymbol{\theta} \sim \text{Bern}\left(\frac{\rho(F_1(t_k; \boldsymbol{\theta}) - F_1(t_{k-1}; \boldsymbol{\theta}))}{\rho(F_1(t_k; \boldsymbol{\theta}) - F_1(t_{k-1}; \boldsymbol{\theta})) + (1-\rho)(F_0(t_k; \boldsymbol{\theta}) - F_0(t_{k-1}; \boldsymbol{\theta}))}\right) \quad (3.118)
$$

$$
Z|X = k, D = 0; \boldsymbol{\theta} \sim \text{LN}_{[t_{k-1}, t_k)}(\mu_0, \sigma_0^2) \quad (3.119)
$$

$$
Z|X = k, D = 1; \boldsymbol{\theta} \sim \text{LN}_{[t_{k-1}, t_k)}(\mu_1, \sigma_1^2) \quad (3.120)
$$

The conditional distributions specified above together with formulas in 3.42 and 3.43 allow us to complete the E-step

Algorithm 3.3.3 (EM algorithm for double log-normal mixture model)**.** .

1. **E-Step**: Given the current guess $\boldsymbol{\theta}_{t-1} = (\rho_{t-1}, \mu_{0,t-1}, \sigma^2_{0,t-1}, \mu_{1,t-1}, \sigma^2_{1,t-1})$, propose 3.118, 3.119, 3.120 for data completion. The minor is given by:

$$
\begin{aligned}
\frac{1}{n} Q(\rho, \mu_0, \sigma_0^2, \mu_1, \sigma_1^2 | \boldsymbol{\theta}_{t-1}) = & \frac{1}{n} \text{E}_{t-1}\{\ln f^c(\mathbf{x}, \mathbf{Z}, \mathbf{D}; \boldsymbol{\theta})\} \\
= & S_{t-1} \ln \rho + (1 - S_{t-1}) \ln(1-\rho) - \frac{1}{2}\ln(2\pi) \\
& - \frac{1}{2}\ln \sigma_1^2 S_{t-1} - N_{t-1} - \frac{1}{2\sigma_1^2} M_{t-1} + \frac{\mu_1}{\sigma_1^2} N_{t-1} - \frac{\mu_1^2}{2\sigma_1^2} S_{t-1} \\
& - \frac{1}{2}\ln \sigma_1^2 (1 - S_{t-1}) - L_{t-1} + N_{t-1} - \frac{1}{2\sigma_0^2} W_{t-1} \\
& + \frac{1}{2\sigma_0^2} M_{t-1} + \frac{\mu_0}{\sigma_0^2} L_{t-1} - \frac{\mu_0}{\sigma_0^2} N_{t-1} - \frac{\mu_0^2}{2\sigma_0^2}(1 - S_{t-1}) \quad (3.121)
\end{aligned}
$$

where:

$$S_{t-1} = \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n} D_i|\mathbf{X} = \mathbf{x}; \boldsymbol{\theta}_{t-1}\right) \tag{3.122}$$

$$M_{t-1} = \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n} D_i(\ln Z_i)^2|\mathbf{X} = \mathbf{x}; \boldsymbol{\theta}_{t-1}\right) \tag{3.123}$$

$$N_{t-1} = \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n} D_i \ln Z_i|\mathbf{X} = \mathbf{x}; \boldsymbol{\theta}_{t-1}\right) \tag{3.124}$$

$$L_{t-1} = \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n} \ln Z_i|\mathbf{X} = \mathbf{x}; \boldsymbol{\theta}_{t-1}\right) \tag{3.125}$$

$$W_{t-1} = \frac{1}{n}\mathrm{E}\left(\sum_{i=1}^{n} (\ln Z_i)^2|\mathbf{X} = \mathbf{x}; \boldsymbol{\theta}_{t-1}\right) \tag{3.126}$$

2. **M-Step**: Solving First order condition, the next improved guess is given by:

$$\rho_t = S_{t-1} \tag{3.127}$$

$$\mu_{0,t} = \frac{L_{t-1} - N_{t-1}}{1 - S_{t-1}} \tag{3.128}$$

$$\sigma^2_{0,t} = \frac{W_{t-1} - M_{t-1}}{1 - S_{t-1}} - \frac{(L_{t-1} - N_{t-1})^2}{(1 - S_{t-1})^2} \tag{3.129}$$

$$\mu_{1,t} = \frac{N_{t-1}}{S_{t-1}} \tag{3.130}$$

$$\sigma^2_{1,t} = \frac{M_{t-1}}{S_{t-1}} - \frac{N^2_{t-1}}{S^2_{t-1}} \tag{3.131}$$

3. **Terminating condition**: Terminate the algorithm if $\|\boldsymbol{\theta}_T - \boldsymbol{\theta}_{T-1}\| < \epsilon$ for a chosen tolerance ϵ. Alternatively, terminate if $|l(\boldsymbol{\theta}_T; \mathbf{x}) - l(\boldsymbol{\theta}_{T-1}; \mathbf{x})| < \epsilon$.

The EM algorithm can be employed to find the maximum likelihood parameters of a statistical model when it is difficult to solve these equation directly. Typically latent variables are involved in these models apart from unknown parameters and known data. If the underlying distribution used belongs to exponential family, then convergence to global maximum is established by (Dempster, Laird and Rubin, 1977). Wu (1983) confirms EM method's convergence outside the realm of Exponential Family. There is still room for improvement by using α-EM algorithm introduced by Matsuyama (2003), which shows faster convergence rate when the right α is chosen.

Chapter 4

Quadratic Matching Estimation

In this chapter, I will construct a sequence of matching estimators, starting from quantile matching, interquantile mean matching estimators to quantile interquantile mean matching and proportion matching estimators, culminating in the development of a unified estimator for three data types. The effort to exploit all data sources is significant in practice when earnings data are summarised in different ways, and using all the information can potentially lead to efficiency gain. The development for each estimator is presented as follows: (1) fundamental asymptotic results, (2) construction and statistical properties, (3) two-step estimator (4) hypothesis testing.

4.1 Method of Quantile Matching

One popular method of summarising income data is to report quantile values. For instance, income levels at 10^{th}, 20^{th}, ..., 90^{th} percentiles are available in income surveys from many countries. In this section, we developed a parametric approach to estimate parameters of a income distribution from quantile data. The method repeats arguments from Generalised Method of Moments (GMM) developed in Hansen (1982); Hansen and Singleton (1982); Hansen, Heaton and Yaron (1996) by recognising similarity of moment conditions and asymptotic behaviour of sample quantiles. Similar development of quantile matching method can be found in Dominicy and Veredas (2013); Sgouropoulos, Yao and Yastremiz (2015).

4.1.1 Statistical Model for Quantile data

A government agency has a job of collecting income data and summarising by quantiles. Let $Y_1, Y_2, \ldots, Y_n$ be the random income values collected by the agency, independently and identically distributed from a distribution with CDF $F(\cdot; \boldsymbol{\theta})$ with d-dimensional vector of parameters. For $\pi \in (0, 1)$, the π^{th} quantile of F is defined as $\xi_\pi = \inf\{y : F(y) \geq \pi\}$, which is alternatively denoted by $F^{-1}(\pi)$. Corresponding to a income sample $Y_1, Y_2, \ldots, Y_n$ of $F(\cdot; \boldsymbol{\theta})$, the sample π^{th} quantile is defined as the p^{th} quantile of the empirical distribution $\hat{F}_n$, denoted by $\hat{\xi}_{\pi,n}$ or $\hat{\xi}_\pi$ when convenient. Increasing sequence of $\pi_1, \pi_2, \ldots, \pi_K \in (0, 1)$ are fixed percentages points chosen by the government before realising the random variables. An external observer can only realise a vector of sample quantiles $(\hat{\xi}_{\pi_1}, \hat{\xi}_{\pi_2}, \ldots, \hat{\xi}_{\pi_K})$ reported by the agency.

We impose an assumption that the distribution function $F(\cdot; \boldsymbol{\theta})$ is strictly increasing and differentiable with density function $f(\cdot; \boldsymbol{\theta})$. Despite telling a story about income, the distribution support is not necessarily between $[0, \infty)$. Reader should keep in mind that the support can be as general as $(-\infty, \infty)$. By making this assumption, we are able to use $F^{-1}(\pi)$, $Q(\pi)$, ξ_π interchangeably for π^{th} quantile. The second assumption concerning a specific formula for sample quantile should be clarified. In practice, there are various rules for obtaining sample quantile; however, when sample size gets larger, the discrepancy between these rules become negligible. Asymptotic results on quantiles remain valid regardless of different rules of computing sample quantiles. It is thus fine to assume the agency use the rule defined previously. The final assumption that is worth mentioning philosophically is iid assumption. In practice, there is often a finite population that the agency can do random sampling on. Therefore, the sample is iid with respect to the discrete distribution associated with finite population. In order to make the analysis tractable, we have to posit there is an underlying “nice, smooth” statistical distribution generating the income, and our sample are iid with respect to that meta-distribution. This assumption is in tune with the concept of **population model** in econometrics (Wooldridge, 2010, p. 5). Even if there are micro-data for the whole population, we still consider them as random sample from some hypothetical statistical distributions.

It can be shown that $\hat{\xi}_\pi$ is a strongly consistent estimator of ξ_π, under mild conditions. Moreover, under mild smoothness requirements on F in the neighbourhoods of points $\xi_{\pi_1}, \xi_{\pi_2}, \ldots, \xi_{\pi_K}$, the vector of sample quantiles $(\hat{\xi}_{\pi_1}, \hat{\xi}_{\pi_2}, \ldots, \hat{\xi}_{\pi_K})$ is asymptotically normal. Intuitively, the matching approach require us to find $\hat{\boldsymbol{\theta}} \in \mathbb{R}^d$, so that the sample quantiles vector $(\hat{\xi}_{\pi_1}, \hat{\xi}_{\pi_2}, \ldots, \hat{\xi}_{\pi_K})$ is as close as possible to $(Q(\pi_1; \hat{\boldsymbol{\theta}}), Q(\pi_2; \hat{\boldsymbol{\theta}}), \ldots, Q(\pi_K; \hat{\boldsymbol{\theta}}))$. If $d = K$, we attempt to solve the system of equation with d unknowns and d equations, known as **exact-identification**. If $d > K$, the problem of **under-identification** shows up, and it is impossible to estimate the parameters. Finally and most importantly, in the case of **over-identification** $K > d$, we have more equations than unknowns to solve, what estimation procedure should we employ to make the most of the abundance of quantile conditions?

4.1.2 Quantile Matching Estimator

Fundamental Asymptotic Results of Sample Quantiles

Before introducing the method of quantile matching, we recall four important asymptotic results on sample quantiles in Serfling (2009).

Theorem 4.1.1 (Consistency of $\hat{\xi}_\pi$)**.** *Let $\pi \in (0,1)$. If true quantile ξ_π is the unique solution x to $F(x-) \leq \pi \leq F(x)$, then $\hat{\xi}_{\pi,n} \xrightarrow{as} \xi_p$, which implies $\hat{\xi}_{\pi,n} \xrightarrow{p} \xi_p$.*

Proof. See Serfling (2009, p. 75). □

Theorem 4.1.2 (Strong Consistency of $\hat{\xi}_\pi$)**.** *Let $\pi \in (0,1)$. If true quantile ξ_π is the unique solution x to $F(x-) \leq \pi \leq F(x)$, then for every $\epsilon > 0$:*

$$P\left(\sup_{m \geq n} |\hat{\xi}_{\pi,m} - \xi_\pi| > \epsilon\right) \leq \frac{2}{1-\rho_\epsilon}\rho_\epsilon^n, \quad \text{for all}\, n$$

where $\rho_\epsilon = \exp(-2\delta_\epsilon^2)$ and $\delta_\epsilon = \min\{F(\xi_\pi + \epsilon) - \pi, \pi - F(\xi_\pi - \epsilon)\}$

Proof. See Serfling (2009, p. 76). □

Immediate consequence from either theorem 4.1.1 or theorem 4.1.2 is the sample quantile converges to true quantile in probability. This consistency result inspires the matching strategy of sample quantiles and true quantiles for parameters estimation.

Theorem 4.1.3 (Asymptotic Normality of $\hat{\xi}_\pi$)**.** *Let* $\pi \in (0,1)$*. Assume that there is a density* f *around the neighbourhood of* ξ_π *and* $f(\xi_\pi) > 0$*, then:*

$$\sqrt{n}(\hat{\xi}_{\pi,n} - \xi_\pi) \xrightarrow{d} \mathrm{N}\left(0, \frac{\pi(1-\pi)}{f(\xi_\pi)^2}\right) \tag{4.1}$$

Proof. See Serfling (2009, p. 77). □

Theorem 4.1.4 (Asymptotic Multivariate Normality of Quantiles)**.** *Let* $0 < \pi_1 < \pi_2 < ... < \pi_K < 1$*. Suppose that* F *has a density* f *in neighbourhoods of* $\xi_{\pi_1}, \xi_{\pi_2}, ..., \xi_{\pi_K}$ *and that* f *is positive and continuous at these quantiles. Then* $(\hat{\xi}_{\pi_1}, \hat{\xi}_{\pi_2}, ..., \hat{\xi}_{\pi_K})$ *is asymptotically normal with mean vector* $(\xi_{\pi_1}, \xi_{\pi_2}, ..., \xi_{\pi_K})$ *and covariance* σ_{ij}/n *where:*

$$\sigma_{ij} = \frac{\pi_i(1-\pi_j)}{f(\xi_{\pi_i})f(\xi_{\pi_j})}; \quad i \leq j$$

Proof. See Serfling (2009, p. 80). □

Construction & Properties of Quantile Matching Estimator

The previous four theorems are the essential ingredients for the development of Quantile Matching Estimator (QME), which repeats the construction of GMM estimator. Define quantiles discrepancy vector as analogy for moment condition vector:

$$q_n(\boldsymbol{\theta}) = \begin{bmatrix} \hat{\xi}_{\pi_1,n} - Q(\pi_1;\boldsymbol{\theta}) \\ \hat{\xi}_{\pi_2,n} - Q(\pi_2;\boldsymbol{\theta}) \\ \vdots \\ \hat{\xi}_{\pi_K,n} - Q(\pi_K;\boldsymbol{\theta}) \end{bmatrix} \tag{4.2}$$

If $K = d$, the quantile matching estimator is defined as the parameter values $\hat{\boldsymbol{\theta}}$ that sets $q_n(\hat{\boldsymbol{\theta}}) = \mathbf{0}$ (quantile conditions). This is generally not possible when there are more quantile conditions than parameters. To deal with excessive quantile conditions, for some $K \times K$ positive definite weighting matrix $\hat{W}$, define objective function:

$$J_n(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} q_n(\boldsymbol{\theta})^{\mathrm{T}} \hat{W} q_n(\boldsymbol{\theta}) \tag{4.3}$$

This is a nonegative objective function generalising the notion of Euclidean length of quantile discrepancy vector. For example, if $\hat{W} = I_K$, then $J_n(\boldsymbol{\theta}) = \|q_n(\boldsymbol{\theta})\|^2$ coincide with Euclidean length. Typically, the objective function can assign different weights to different square discrepancy entries, and there are interaction terms between entries as well. QME estimator is the minimiser of $J_n(\boldsymbol{\theta})$:

$$\hat{\boldsymbol{\theta}}_{QME} = \arg\min_{\boldsymbol{\theta}} \; J_n(\boldsymbol{\theta}) \tag{4.4}$$

If $K = d$, then there is $\hat{\boldsymbol{\theta}}$ such that the quantile conditions $q_n(\hat{\boldsymbol{\theta}}) = 0$. The QME does not depend on the weighting matrix $\hat{W}$. Except for quantile functions associated with uniform family, the objective function defined in equation 4.3 is highly nonlinear and an optimisation routine is needed to solve for the maximiser.

I want to clarify the notation before use to avoid readers' conflicts in convention. Function $J_n : \mathbb{R}^d \to \mathbb{R}_+$ has gradient $\nabla J_n : \mathbb{R}^d \to \mathbb{R}^d$ and Hessian $\nabla^2 J_n : \mathbb{R}^d \to \mathbb{R}^{d\times d}$. Moreover, function $q_n : \mathbb{R}^d \to \mathbb{R}^K$ has Jacobian $\nabla q_n : \mathbb{R}^d \to \mathbb{R}^{K\times d}$ and Jacobian tranpose $\nabla^{\mathrm{T}} q_n : \mathbb{R}^d \to \mathbb{R}^{d\times K}$. The gradient and Hessian of J_n are:

$$\nabla J_n(\boldsymbol{\theta}) = 2\nabla^{\mathrm{T}} q_n(\boldsymbol{\theta}) \, \hat{W} \, q_n(\boldsymbol{\theta}) \tag{4.5}$$

$$\nabla^2 J_n(\boldsymbol{\theta}) = 2\nabla^{\mathrm{T}} q_n(\boldsymbol{\theta}) \, \hat{W} \, \nabla q_n(\boldsymbol{\theta}) + 2\mathcal{M} \odot \hat{W} \, q_n(\boldsymbol{\theta}) \tag{4.6}$$

where $\mathcal{M}$ is a mysterious term, representing the total derivative of matrix-valued function $\nabla q_n : \mathbb{R}^d \to \mathbb{R}^{K\times d}$ and the multiplication rule for this term is also not proper. Again, the Jacobian of discrepancy function $q_n(\boldsymbol{\theta})$ is non-stochastic:

$$\nabla q_n(\boldsymbol{\theta}) = -\begin{bmatrix} \nabla^{\mathrm{T}} Q(\pi_1; \boldsymbol{\theta}) \\ \nabla^{\mathrm{T}} Q(\pi_2; \boldsymbol{\theta}) \\ \vdots \\ \nabla^{\mathrm{T}} Q(\pi_K; \boldsymbol{\theta}) \end{bmatrix} \tag{4.7}$$

and so is the $\mathcal{M}$ term. Denote $A(\boldsymbol{\theta}) = -\nabla q_n(\boldsymbol{\theta})$ and called **discrete A-matrix** from now on. In proving consistency and asymptotic normality of QME, there is no need to invoke central limit theorem as in the case of GMM to show the $\mathcal{M}$ term and ∇q_n stabilises. The next two theorems shows consistency and asymptotic normality of QME using Taylor expansion with a regularity condition.

Assumption 4.1.5 (No Multicollinearity)**.** $A = -\nabla q_n(\boldsymbol{\theta}_0)$ has independent columns at true parameters $\boldsymbol{\theta}_0$. Alternatively, A has no vector in the nullspace except 0.

This is a mild regularity condition since every well-known distributions satisfy this assumption. We will implicitly assume this condition from now on. Readers are encouraged to pause and think about the connection between this assumption and the problem of multicolinearity in linear regression.

Theorem 4.1.6 (Consistency of QME)**.** *For any weighting matrix $\hat{W}$ with $\hat{W} \xrightarrow{p} W$, the quantile matching estimator $\hat{\boldsymbol{\theta}}_n = \arg\min J_n(\boldsymbol{\theta})$ is consistent estimator of the true parameters $\boldsymbol{\theta}_0$.*

Proof. Fix $\epsilon > 0$, we wish to show $\mathrm{P}(\|\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0\| > \epsilon) \to 0$ as $n \to \infty$. For all $\boldsymbol{\theta} \in C_\epsilon(\boldsymbol{\theta}_0)$ (hypersphere surface of radius ϵ about $\boldsymbol{\theta}_0$), Taylor expansion of J_n around the true parameter $\boldsymbol{\theta}_0$ gives:

$$J_n(\boldsymbol{\theta}) - J_n(\boldsymbol{\theta}_0) = \nabla J_n(\boldsymbol{\theta}_0)^{\mathrm{T}}(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + \frac{1}{2}(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\mathrm{T}} \nabla^2 J_n(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + R_n \tag{4.8}$$

where R_n is random remainder term.

One should be aware that the function J_n is a random function as it depends on iid sample $Y_1, Y_2, ..., Y_n$ withdrawn from $F(\cdot; \boldsymbol{\theta}_0)$. It is better to keep in mind $J_n(\boldsymbol{\theta}) =$

$J(\boldsymbol{\theta}, \mathbf{Y})$ but I will drop the $\mathbf{Y}$ for simplicity. Equation 4.8 can be rewritten as:

$$\begin{aligned} J_n(\boldsymbol{\theta}) - J_n(\boldsymbol{\theta}_0) = & \, 2q_n(\boldsymbol{\theta}_0)^{\mathrm{T}} \hat{W} \nabla q_n(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ & + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\mathrm{T}} \nabla^{\mathrm{T}} q_n(\boldsymbol{\theta}_0) \, \hat{W} \, \nabla q_n(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \\ & + (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\mathrm{T}} \mathcal{M} \odot \hat{W} \, q_n(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) + R_n \end{aligned} \tag{4.9}$$

Since $q_n(\boldsymbol{\theta}_0) \xrightarrow{p} 0$ by consistency of sample quantiles (theorem 4.1.1) and $\hat{W} \xrightarrow{p} W$ by assumption and $\nabla q_n(\boldsymbol{\theta}_0)$ is a constant $K \times d$ matrix, Slutky's theorem tells:

$$2q_n(\boldsymbol{\theta}_0)^{\mathrm{T}} \hat{W} \nabla q_n(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \xrightarrow{p} 0 \tag{4.10}$$

$$(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\mathrm{T}} \nabla q_n(\boldsymbol{\theta}_0)^{\mathrm{T}} \hat{W} \nabla q_n(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \xrightarrow{p} (\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\mathrm{T}} A^{\mathrm{T}} W A (\boldsymbol{\theta} - \boldsymbol{\theta}_0) \tag{4.11}$$

$$(\boldsymbol{\theta} - \boldsymbol{\theta}_0)^{\mathrm{T}} \mathcal{M} \odot \hat{W} \, q_n(\boldsymbol{\theta}_0)(\boldsymbol{\theta} - \boldsymbol{\theta}_0) \xrightarrow{p} 0 \tag{4.12}$$

$$R_n \sim O(\epsilon^3) \tag{4.13}$$

A has independent columns (assumption 4.1.5) and W is positive definite, so $A^{\mathrm{T}} W A$ is guaranteed to be positive definite. Suppose ϵ is fixed but small enough, then R_n is negligible. Therefore, on the surface of $C_\epsilon(\boldsymbol{\theta}_0)$, the probability tend to 1 that $J_n(\boldsymbol{\theta}) > J_n(\boldsymbol{\theta}_0)$. It is equivalent to $\mathrm{P}(\|\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\| > \epsilon) \to 0$ as $n \to \infty$ as desired. □

By theorem 4.1.4, we have $\sqrt{n} q_n(\boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}(0, \Sigma)$ where:

$$\Sigma = \Sigma(\boldsymbol{\theta}_0) = \begin{bmatrix} \sigma_1^2 & \sigma_{21} & \cdots & \sigma_{1K} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{K1} & \sigma_{K2} & \cdots & \sigma_K^2 \end{bmatrix} \tag{4.14}$$

and

$$\sigma_{ij} = \frac{\pi_i \wedge \pi_j - \pi_i \pi_j}{f(Q(\pi_i; \boldsymbol{\theta}_0); \boldsymbol{\theta}_0) f(Q(\pi_j; \boldsymbol{\theta}_0); \boldsymbol{\theta}_0)} \tag{4.15}$$

Theorem 4.1.7 (Asymptotic Normality of QME). *For any weighting matrix $\hat{W}$ with $\hat{W} \xrightarrow{p} W$, the quantile matching estimator $\hat{\boldsymbol{\theta}}_n = \arg\min J_n(\boldsymbol{\theta})$ is asymptotically* *normal:*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N\left(0, (A^{\mathrm{T}}WA)^{-1}A^{\mathrm{T}}W\Sigma WA(A^{\mathrm{T}}WA)^{-1}\right)$$

where $A = -\nabla q_n(\boldsymbol{\theta}_0)$ is discrete A-matrix at true parameters vector.

Proof. Using Taylor expansion at $\boldsymbol{\theta}_0$ we have:

$$\nabla J_n(\hat{\boldsymbol{\theta}}_n) = \nabla J_n(\boldsymbol{\theta}_0) + \nabla^2 J_n(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + O(\|\boldsymbol{\theta}_n - \boldsymbol{\theta}_0\|^2) \tag{4.16}$$

At quantile matching estimator $\hat{\boldsymbol{\theta}}_n$, $\nabla J_n(\hat{\boldsymbol{\theta}}_n) = 0$ by first order condition:

$$\begin{aligned} 0 &= 2\nabla^{\mathrm{T}} q_n(\boldsymbol{\theta}_0)\hat{W} q_n(\boldsymbol{\theta}_0) + 2\nabla^{\mathrm{T}} q_n(\boldsymbol{\theta}_0)\,\hat{W}\,\nabla q_n(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \\ &\quad + 2\mathcal{M} \odot \hat{W}\, q_n(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \\ &= -2A\hat{W} q_n(\boldsymbol{\theta}_0) + 2A^{\mathrm{T}}\hat{W}A(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) + 2\mathcal{M} \odot \hat{W}\, q_n(\boldsymbol{\theta}_0)(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \end{aligned} \tag{4.17}$$

On the right hand side of the above equation, the first are of order $O_p\left(\frac{1}{\sqrt{n}}\right)$ and the second term are of order $o_p\,(1)$ whereas the third term is of order $o_p\left(\frac{1}{\sqrt{n}}\right)$ due to $q_n(\boldsymbol{\theta}_0) \sim O_p\left(\frac{1}{\sqrt{n}}\right)$ and $\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0 \sim o_p\,(1)$ (theorem 4.1.3 and theorem 4.1.6). Therefore:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) = (A^{\mathrm{T}}\hat{W}A)^{-1}A\hat{W}\sqrt{n}q_n(\boldsymbol{\theta}_0) + o_p\,(1) \tag{4.18}$$

By assumption $\hat{W} \xrightarrow{p} W$, and by invoking $\sqrt{n}q_n(\boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}(0, \Sigma)$ again (theorem 4.1.4) we eventually obtain the desired result. □

The next theorem concerns about the most efficient weighting matrix

Theorem 4.1.8 (Optimal Weight of QME)**.** *If weighting matrix $\hat{W} \xrightarrow{p} \Sigma^{-1}$, then*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}(0, (A^{\mathrm{T}}\Sigma^{-1}A)^{-1})$$

and the covariance matrix is the lowest among all choices of weighting matrix. In other words, the estimator corresponding to this weighting matrix is the most efficient.

Proof. The first statement of normality is clear by using theorem 4.1.7 with special weighting matrix $W = \Sigma^{-1}$. In order to show the covariance matrix is the smallest, we show it is less than the covariance matrix associated with any weighting matrix W in matrix sense. In fact:

$$\begin{aligned}
&V_{\boldsymbol{\theta}}(W) - V_{\boldsymbol{\theta}}(\Sigma^{-1}) = (A^{\mathrm{T}}WA)^{-1}A^{\mathrm{T}}W\Sigma WA(A^{\mathrm{T}}WA)^{-1} - (A^{\mathrm{T}}\Sigma^{-1}A)^{-1} \\
=&(A^{\mathrm{T}}WA)^{-1}\Big(A^{\mathrm{T}}W\Sigma WA - A^{\mathrm{T}}WA(A^{\mathrm{T}}\Sigma^{-1}A)^{-1}A^{\mathrm{T}}WA\Big)(A^{\mathrm{T}}WA)^{-1} \\
=&(A^{\mathrm{T}}WA)^{-1}A^{\mathrm{T}}W\Sigma^{1/2}\Big(I - \Sigma^{-1/2}A(A^{\mathrm{T}}\Sigma^{-1}A)^{-1}A^{\mathrm{T}}\Sigma^{-1/2}\Big)\Sigma^{1/2}WA(A^{\mathrm{T}}WA)^{-1}
\end{aligned} \tag{4.19}$$

where $M = I - \Sigma^{-1/2}A(A^{\mathrm{T}}\Sigma^{-1}A)^{-1}A^{\mathrm{T}}\Sigma^{-1/2}$ is projection matrix onto the left nullspace of $\Sigma^{-1/2}A$. M is symmetric and $M^2 = M$, therefore the covariance difference matrix can be expressed as $L^{\mathrm{T}}L$, proving the fact that it is at least positive semidefinite as desired. □

Two Step Quantile Matching Estimator

To construct asymptotically efficient QME estimator, we need a weighting matrix $\hat{W}$ that converges in probability to Σ^{-1}. To achieve this consistency for weighting matrix, we construct $\hat{W}$ as an estimate of Σ^{-1}. Given a preliminary weighting matrix $\tilde{W}$, we construct QME estimator corresponding to $\tilde{W}$:

$$J_n(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} q_n(\boldsymbol{\theta})^{\mathrm{T}}\tilde{W}q_n(\boldsymbol{\theta}) \tag{4.20}$$

$$\tilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \; J_n(\boldsymbol{\theta}) \tag{4.21}$$

For any weighting matrix W_1, theorem 4.1.6 implies $\tilde{\boldsymbol{\theta}}$ is consistent. We compute estimate for covariance matrix by equation 4.14: $\tilde{\Sigma} = \Sigma(\tilde{\boldsymbol{\theta}})$ and put $\hat{W} = \tilde{\Sigma}^{-1}$. Then it is clear that $\hat{W} \xrightarrow{p} \Sigma^{-1}$. Now we are ready to conduct the second step:

$$J_n(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} q_n(\boldsymbol{\theta})^{\mathrm{T}}\hat{W}q_n(\boldsymbol{\theta}) \tag{4.22}$$

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \; J_n(\boldsymbol{\theta}) \tag{4.23}$$

Then one can show $\hat{\boldsymbol{\theta}}$ constructed by two step estimation is consistent and asymptotically most efficient.

Theorem 4.1.9 (Efficient Two-step QME). *The two-step QME estimator* $\hat{\boldsymbol{\theta}}$ *constructed above is consistent estimator of* $\boldsymbol{\theta}_0$ *and asymptotically normal:*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}(0, (A^{\mathrm{T}}\Sigma^{-1}A)^{-1})$$

Proof. A direct application of theorem 4.1.7 with weighting matrix $\hat{W} \xrightarrow{p} \Sigma^{-1}$. □

The result also show that the asymptotic behaviour of two-step estimator $\hat{\boldsymbol{\theta}}$ does not depend on the choice of preliminary weighting matrix $\hat{W}$. In finite samples, $\hat{\boldsymbol{\theta}}$, will be affected by the choice of $\hat{W}$. This undesirable dependency can be eliminated by iterating the estimation sequence until convergence. This is called **iterated QME estimator**. Another strategy called **continuously-updated QME estimator** can be repeated in the same way as its counterpart in GMM, introduced in Hansen, Heaton and Yaron (1996). However, we will not adopt these advanced strategies in this paper due to high computational cost. Instead we will use two-step estimator with default preliminary weighting matrix $\tilde{W} = I_K$

After obtaining the two step estimator $\hat{\boldsymbol{\theta}}$, the estimate for asymptotic covariance matrix of $\sqrt{n}\,\hat{\boldsymbol{\theta}}$ is $(\hat{A}^{\mathrm{T}}\hat{\Sigma}^{-1}\hat{A})^{-1}$ converging in probability to $(A^{\mathrm{T}}\Sigma^{-1}A)^{-1}$ where:

$$\hat{A} = -\nabla q_n(\hat{\boldsymbol{\theta}}) \xrightarrow{p} A = -\nabla q_n(\boldsymbol{\theta}_0) \tag{4.24}$$

$$\hat{\Sigma} = \Sigma(\hat{\boldsymbol{\theta}}) \xrightarrow{p} \Sigma = \Sigma(\boldsymbol{\theta}_0) \tag{4.25}$$

Hypothesis Testing

The estimate of covariance matrix enables us to construct confidence intervals, confidence ellipsoid, hypothesis testing of parameters, and even functions of parameters (see Delta Method 4.1.12 for how to construct asymptotic distribution of a function of parameters). Because GMM and QME are parallel in mathematical structure, the two theorems provided next are immediate.

Theorem 4.1.10 (Sargan-Hansen Over-identification test)**.** *Under the hypothesis of correct specification, and if the weighting matrix is asymptotically efficient (valid for two step estimator) then:*

$$J_n(\hat{\boldsymbol{\theta}}) \xrightarrow{p} \chi^2_{K-d}$$

Proof. Adapted from Sargan (1958) and Hansen (1982). □

Theorem 4.1.11 (Distance Test)**.** *Suppose function $h : \mathbb{R} \to \mathbb{R}^l$ with $l \leq d$ is continuously differentiable. We want to test null hypothesis $h(\boldsymbol{\theta}_0) = 0$. Using the same weighting matrix, denote $\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\arg\min},\, J_n(\boldsymbol{\theta})$ and $\tilde{\boldsymbol{\theta}} = \underset{h(\boldsymbol{\theta})=0}{\arg\min} J_n(\boldsymbol{\theta})$ and put $D = J_n(\tilde{\boldsymbol{\theta}}) - J_n(\hat{\boldsymbol{\theta}})$. Then the following statements hold:*

1. *$D \geq 0$*
2. *$D \xrightarrow{d} \chi^2_l$*
3. *if h is linear in $\boldsymbol{\theta}$, then D coincides with Wald statistics.*

Proof. Adapted from Hayashi (2000, p. 220). □

Theorem 4.1.12 (Delta Method)**.** *$\boldsymbol{g} : \mathbb{R}^m \to \mathbb{R}^k$ is a continuously differentiable function. Suppose the sequence of random vectors X_n in $\mathbb{R}^m$ is asymptotically multivariate normal:*

$$\sqrt{n}(X_n - \mu) \xrightarrow{p} \mathrm{N}(0, \Sigma)$$

Then the sequence of transformed vectors $\boldsymbol{g}(X_n)$ in $\mathbb{R}^k$ is also asymptotically multivariate normal:

$$\sqrt{n}(\boldsymbol{g}(X_n) - \boldsymbol{g}(\mu)) \xrightarrow{p} \mathrm{N}(0, J\Sigma J^{\mathrm{T}})$$

where J is the Jacobian of function $\boldsymbol{g}$ (size $k \times m$), evaluated at μ.

Proof. See Kroese, Chan et al. (2014, p. 92). □

4.2 Method of Interquantile Mean Matching

Another popular method of summarising income data is reporting interquantile means. For instance, income data are partitioned into 5 or 10 equal groups (called quintiles, or deciles) and the mean for each group is computed. In this section, I developed a parametric approach to estimate parameters of income distribution from interquantile mean data. The strategy again repeats arguments from Generalised Method of Moment (GMM) by recognising similarity of moment conditions and asymptotic behaviour of interquantile means. This section will be considerably shorter since many proofs are omitted due to identical structure with previous section.

4.2.1 Statistical Model for Interquantile Mean Data

The story goes again: a government agency has a task of collecting income data and summarising the information by several interquantile means. Let $Y_1, Y_2, \ldots, Y_n$ be the random income values collected by the agency, independently and identically distributed from a distribution with strictly increasing, differentiable CDF $F(\cdot; \boldsymbol{\theta}_0)$. Quantile and density functions are well defined, and depend on d-dimensional vector of parameters $\boldsymbol{\theta}$. For $\rho_1, \rho_2 \in (0,1)$ with $\rho_1 < \rho_2$, the (ρ_1, ρ_2)-interquantile mean of F is defined as (refer to equation 2.4):

$$\mu(\rho_1, \rho_2; \boldsymbol{\theta}) = \frac{1}{\rho_2 - \rho_1} \int_{\rho_1}^{\rho_2} Q(u; \boldsymbol{\theta})\, du = \frac{1}{\rho_2 - \rho_1} (Q^A(\rho_2; \boldsymbol{\theta}) - Q^A(\rho_1; \boldsymbol{\theta})) \tag{4.26}$$

which is sometimes denoted $\mu(\rho_1, \rho_2)$ if there is no ambiguity in notation. Followed by the concept of interquantile mean, the notions of interquantile h-th moment, and interquantile variance are apparent and self-explainatory. The (ρ_1, ρ_2)-interquantile h-th moment is:

$$\mu^{(h)}(\rho_1, \rho_2; \boldsymbol{\theta}) = \frac{1}{\rho_2 - \rho_1} \int_{\rho_1}^{\rho_2} Q(u; \boldsymbol{\theta})^h\, du = \frac{1}{\rho_2 - \rho_1} (Q^{A_h}(\rho_2; \boldsymbol{\theta}) - Q^{A_h}(\rho_1; \boldsymbol{\theta})) \tag{4.27}$$

and for simplicity, we suppress the superscript when denoting interquantile mean. The (ρ_1, ρ_2)-interquantile variance is:

$$\sigma^2(\rho_1, \rho_2; \boldsymbol{\theta}) = \mu^{(2)}(\rho_1, \rho_2; \boldsymbol{\theta}) - (\mu(\rho_1, \rho_2; \boldsymbol{\theta}))^2 \tag{4.28}$$

Corresponding to a income sample $Y_1, Y_2, \ldots, Y_n$ of $F(\cdot; \boldsymbol{\theta})$, the sample (ρ_1, ρ_2)- interquantile mean is defined as:

$$\hat{\mu}(\rho_1, \rho_2) = \frac{1}{\lfloor n\rho_2 \rfloor - \lfloor n\rho_1 \rfloor} \sum_{i=\lfloor n\rho_1 \rfloor + 1}^{\lfloor n\rho_2 \rfloor} Y_{i:n} \tag{4.29}$$

where $Y_{1:n} \leq Y_{2:n} \leq \ldots \leq Y_{n-1:n} \leq Y_{n:n}$ are orders statistics obtained by ranking the sample $Y_1, Y_2, \ldots, Y_n$ in increasing order. Increasing sequence of $\rho_0, \rho_1, \ldots, \rho_{K-1}, \rho_K \in (0,1)$ are fixed percentages points chosen by the government before realising the income sample. As an external observer outside the agency, he or she can only realise a vector of sample interquantile means $(\hat{\mu}(\rho_0, \rho_1), \hat{\mu}(\rho_1, \rho_2), \ldots, \hat{\mu}(\rho_{K-1}, \rho_K))$.
It can be shown that, under mild regularity condition, $\hat{\mu}(\rho_1, \rho_2)$ is a consistent estimator of $\mu(\rho_1, \rho_2)$ and the vector of sample interquantile means

$$(\hat{\mu}(\rho_0, \rho_1),\ \hat{\mu}(\rho_1, \rho_2),\ \ldots,\ \hat{\mu}(\rho_{K-1}, \rho_K))$$

is asymptotically normal. Intuitively, the matching approach require us to find $\hat{\boldsymbol{\theta}} \in \mathbb{R}^d$, so that the sample interquantile means vector $\big(\hat{\mu}(\rho_0, \rho_1), \hat{\mu}(\rho_1, \rho_2), \ldots, \hat{\mu}(\rho_{K-1}, \rho_K)\big)$ is as close as possible to $\big(\mu(\rho_0, \rho_1; \boldsymbol{\theta}), \mu(\rho_1, \rho_2; \boldsymbol{\theta}), \ldots, \mu(\rho_{K-1}, \rho_K; \boldsymbol{\theta})\big)$. If $d = K$, we attempt to solve the system of equation with d unknowns and d equations, known as **exact-identification**. If $d > K$, the problem of **under-identification** shows up, and it is impossible to estimate the parameters. Finally and most importantly, in the case of **over-identification** $K > d$, we have more equations than unknowns to solve, what estimation procedure should we employ to make the most of the abundance of interquantile mean conditions?

4.2.2 Interquantile Mean Matching Estimator

The matching strategy I present next is for distributions whose support is $(0, \infty)$ and for non-overlapping interquantile ranges. However, the extensions to general support and overlapping interquantile ranges are straightforward. We again begin the development of new estimator by introducing necessary asymptotic theorems related to sample interquantile mean.

Fundamental Asymptotic Results of Sample Interquantile Means

An important statistic often seen in income report is sample interquantile mean. To obtain this statistic, either the upper or lower portion of the sample are deleted and the mean for the remaining sample is computed. Stigler (1973) has obtained a general asymptotic result for this statistic using advanced technique in asymptotic theory. I will instead present an alternative elementary proof whose technique enables us to prove several new asymptotic results essential for the development of a new estimator.

Theorem 4.2.1 (Stigler). *Suppose* $0 < p_1 < p_2 < 1$, *and* $M_n = \hat{\mu}(p_1, p_2)$ *is sample interquantile mean associated with random iid sample of size n from distribution* $F(\cdot)$. F *is strictly increasing, differentiable so quantile function* Q *and density function* f *is well-defined. Let* $\mu_H = \mu(p_1, p_2)$ *and* $\sigma_H^2 = \sigma^2(p_1, p_2)$ *be the true interquantile mean and interquantile variance defined previously. Then* M_n *is a consistent estimator of* μ_H *and the asymptotic behaviour of* M_n *is given by:*

$$
\begin{aligned}
&\sqrt{n}(M_n - \mu_H) \xrightarrow{d} \mathrm{N}(0, \sigma^2) \quad \text{where} \\
&(p_2 - p_1)^2 \sigma^2 = (p_2 - p_1)\sigma_H^2 + (Q(p_1) - \mu_H)^2 p_1(1 - p_1) \\
&\ + (Q(p_2) - \mu_H)^2 p_2(1 - p_2) - 2(Q(p_1) - \mu_H)(Q(p_2) - \mu_H) p_1(1 - p_2)
\end{aligned}
$$

Proof. (Alternative) The proof is based on the idea that as the sample size n goes to infinity, the random continuum vector $(\hat{\xi}_{p,n})_{p\in(0,1)}$ is asymptotically continuum-variate normal (take theorem 4.1.4 to the extreme) with mean vector $(\xi_p)_{p\in(0,1)}$ and covariance continuum matrix $(\sigma_{p,q}/n)_{p,q\in(0,1)}$ where:

$$
\sigma_{p,q} = \frac{p \wedge q - pq}{f(Q(p))f(Q(q))} \tag{4.30}
$$

An alternative view to look at sample interquantile mean instead of using order statistics (equation 4.29) is using quantiles. In fact, $Y_{i:n} = \hat{\xi}_{i/n}$ and:

$$
M_n = \frac{1}{\lfloor np_2 \rfloor - \lfloor np_1 \rfloor} \sum_{i=\lfloor np_1 \rfloor + 1}^{\lfloor np_2 \rfloor} \hat{\xi}_{i/n} = \frac{1}{p_2 - p_1} \int_{p_1}^{p_2} \hat{\xi}_{p,n} \, dp \tag{4.31}
$$

The use of equal sign in equation 4.31 is not entirely correct, and the full justification is deferred to Appendix A.2. In fact, there is a random "noise" that is inherent in the equation. Because the noise is extremely small even with a magnification by a factor of $\sqrt{n}$, it can be safely regarded as 0.

Since $(\hat{\xi}_{p,n})_{p\in(0,1)}$ is asymptotically continuum-variate normal and integration is a linear operator, M_n must be asymptotically normal with mean:

$$\frac{1}{p_2 - p_1}\int_{p_1}^{p_2} \xi_p \, dp = \mu_H \tag{4.32}$$

and variance:

$$\frac{1}{(p_1 - p_2)^2}\int_{p_1}^{p_2}\int_{p_1}^{p_2} \frac{p \wedge q - pq}{f(Q(p))f(Q(q))}\, dp\, dq \tag{4.33}$$

By symmetry, the variance can be expressed as $(2A - B)/(p_1 - p_2)^2$ where:

$$B = \int_{p_1}^{p_2}\int_{p_1}^{p_2} \frac{pq}{f(Q(p))f(Q(q))}\, dp\, dq \tag{4.34}$$

$$A = \int_{p_1}^{p_2}\int_{p_1}^{q} \frac{p}{f(Q(p))f(Q(q))}\, dp\, dq \tag{4.35}$$

The remaining task to be done is showing:

$$\begin{aligned} 2A - B = {} & (p_2 - p_1)\sigma_H^2 + (Q(p_1) - \mu_H)^2 p_1(1 - p_1) \\ & + (Q(p_2) - \mu_H)^2 p_2(1 - p_2) - 2(Q(p_1) - \mu_H)(Q(p_2) - \mu_H)p_1(1 - p_2) \end{aligned} \tag{4.36}$$

How to show equation 4.36 is deferred to Appendix A.3, completing the proof. □

Theorem 4.2.2 (Asymptotic Normality of Interquantile Means). *Suppose* $0 < p_1 < p_2 \le p_3 < p_4 < 1$, *and* $M_n = \hat{\mu}(p_1, p_2)$, $Q_n = \hat{\mu}(p_3, p_4)$ *are sample interquantile means associated with iid sample of size* n *from continuously differentiable distribution function* F. *Let* f *and* Q *be the density function and quantile function. Then* $\sqrt{n}\big(M_n - \mu(p_1, p_2),\, P_n - \mu(p_3, p_4)\big)$ *converges in distribution to a bivariate normal random vectors, whose means are 0, variances are given in theorem 4.2.1, and co-*

variance ω*:*

$$\omega = \frac{1}{(p_2 - p_1)(p_4 - p_3)} \big[(1 - p_4)Q(p_4) - (1 - p_3)Q(p_3) + (p_4 - p_3)\mu(p_3, p_4)\big] \times$$
$$\big[p_2 Q(p_2) - p_1 Q(p_1) - (p_2 - p_1)\mu(p_1, p_2)\big]$$

Proof. Recall that sample interquantile mean can be expressed as integral of sample quantiles:

$$M_n = \frac{1}{p_2 - p_1} \int_{p_1}^{p_2} \hat{\xi}_{p,n}\, dp, \quad Q_n = \frac{1}{p_4 - p_3} \int_{p_3}^{p_4} \hat{\xi}_{q,n}\, dq \tag{4.37}$$

Using the fact that random vector $(\hat{\xi}_{p,n})_{p \in (0,1)}$ is asymptotically continuum-variate normal (take theorem 4.1.4 to the extreme) with mean vector $(\xi_p)_{p \in (0,1)}$ and covariance continuum matrix $(\sigma_{p,q}/n)_{p,q \in (0,1)}$ where:

$$\sigma_{p,q} = \frac{p \wedge q - pq}{f(Q(p)) f(Q(q))}, \tag{4.38}$$

it must be the case that (M_n, P_n) is asymptotically bivariate normal with mean $\big(\mu(p_1, p_2), \mu(p_3, p_4)\big)$ and covariance ω/n where:

$$\omega = \frac{1}{(p_2 - p_1)(p_4 - p_3)} \int_{p_1}^{p_2} \int_{p_3}^{p_4} \frac{p(1 - q)}{f(Q(p)) f(Q(q))}\, dp\, dq \tag{4.39}$$

It is not hard to show that this double integral yields the desired result. □

In the case of multiple sample interquantile means, the extension is straightforward. These statistics form a random vector that is asymptotically multivariate normal whose means, variances and covariances are specified above. In the case of overlapping interquantile ranges, the asymptotic covariance can be derived from two previous theorems by breaking into non-overlapping ranges. The final remark is: if the limits to 0 and 1 make sense, theorem 4.2.1 can be extended to the case $p_1 = 0, p_2 = 1$, and theorem 4.2.2 can be extended to the case $p_1 = 0, p_4 = 1$. I hope your intuition says these extensions are correct.

Construction & Properties of Interquantile Mean Matching Estimator

The previous two theorems are elements necessary for the development of Interquantile Mean Matching Estimator (IME), which repeats the construction of GMM estimator. Here we develop an estimator specifically for non-overlapping interquantile range; however, the extension to overlapping interquantile ranges are straightforward. As an analogy for moment condition vector, we define interquantile means discrepancy vector :

$$m_n(\boldsymbol{\theta}) = \begin{bmatrix} \hat{\mu}_n(\rho_0,\rho_1) - \mu(\rho_0,\rho_1;\boldsymbol{\theta}) \\ \hat{\mu}_n(\rho_1,\rho_2) - \mu(\rho_1,\rho_2;\boldsymbol{\theta}) \\ \vdots \\ \hat{\mu}_n(\rho_{K-1},\rho_K) - \mu(\rho_{K-1},\rho_K;\boldsymbol{\theta}) \end{bmatrix} \tag{4.40}$$

If $K = d$, the IME estimator is defined as the parameter values $\hat{\boldsymbol{\theta}}$ that satisfies $m_n(\boldsymbol{\theta}) = 0$ (interquantile mean conditions). This is generally not possible when there are more quantile conditions than parameters. For some $K \times K$ positive definite weighting matrix $\hat{W}$, define objective function:

$$J_n(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} m_n(\boldsymbol{\theta})^{\mathrm{T}} \hat{W} m_n(\boldsymbol{\theta}) \tag{4.41}$$

This is a nonegative objective function generalising the notion of Euclidean length of quantile discrepancy vector. For example, if $\hat{W} = I_K$, then $J_n(\boldsymbol{\theta}) = \|m_n(\boldsymbol{\theta})\|^2$ coincide with Euclidean length. Typically, the objective function can assign different weights to different square discrepancy entries, and there are interaction terms between entries as well. QME estimator is the minimiser of $J_n(\boldsymbol{\theta})$:

$$\hat{\boldsymbol{\theta}}_{\mathrm{IME}} = \arg\min_{\boldsymbol{\theta}} \; J_n(\boldsymbol{\theta}) \tag{4.42}$$

Function $J_n : \mathbb{R}^d \to \mathbb{R}_+$ has gradient $\nabla J_n : \mathbb{R}^d \to \mathbb{R}^d$ and Hessian $\nabla^2 J_n : \mathbb{R}^d \to \mathbb{R}^{d\times d}$. Moreover, function $m_n : \mathbb{R}^d \to \mathbb{R}^K$ has Jacobian $\nabla m_n : \mathbb{R}^d \to \mathbb{R}^{K\times d}$ and

Jacobian tranpose $\nabla^{\mathrm{T}} m_n : \mathbb{R}^d \to \mathbb{R}^{d \times K}$. The gradient and Hessian of J_n are:

$$\nabla J_n(\boldsymbol{\theta}) = 2\nabla^{\mathrm{T}} m_n(\boldsymbol{\theta})\, \hat{W}\, m_n(\boldsymbol{\theta}) \tag{4.43}$$

$$\nabla^2 J_n(\boldsymbol{\theta}) = 2\nabla^{\mathrm{T}} m_n(\boldsymbol{\theta})\, \hat{W}\, \nabla m_n(\boldsymbol{\theta}) + 2\mathcal{M} \odot \hat{W}\, m_n(\boldsymbol{\theta}) \tag{4.44}$$

where $\mathcal{M}$ is a mysterious term, representing the total derivative of matrix-valued function $\nabla m_n : \mathbb{R}^d \to \mathbb{R}^{K \times d}$ and the multiplication rule for this term is also not proper (denoted harmlessly by a circle with dot). We only need to know that all entries of $2\mathcal{M} \odot \hat{W}\, m_n(\boldsymbol{\theta})$ contain at least an entry of $m_n(\boldsymbol{\theta})$ as a factor. One feature in IME that I find easier than GMM is that the Jacobian of discrepancy function $m_n(\boldsymbol{\theta})$ is no longer a random function:

$$\nabla m_n(\boldsymbol{\theta}) = -\begin{bmatrix} \nabla^{\mathrm{T}}_{\boldsymbol{\theta}} \mu(\rho_0, \rho_1; \boldsymbol{\theta}) \\ \nabla^{\mathrm{T}}_{\boldsymbol{\theta}} \mu(\rho_1, \rho_2; \boldsymbol{\theta}) \\ \vdots \\ \nabla^{\mathrm{T}}_{\boldsymbol{\theta}} \mu(\rho_{K-1}, \rho_K; \boldsymbol{\theta}) \end{bmatrix} \tag{4.45}$$

and so is the $\mathcal{M}$ term. Denote $B(\boldsymbol{\theta}) = -\nabla m_n(\boldsymbol{\theta})$ and I shall call B as **discrete B-matrix** from now on.

Assumption 4.2.3 (No Multicollinearity)**.** $B = -\nabla m_n(\boldsymbol{\theta}_0)$ has independent columns at true parameters $\boldsymbol{\theta}_0$. Alternatively, B has no vector in the nullspace except 0.

This is a mild regularity condition because all well-known distributions satisfy this assumption. We will implicitly assume this condition from now on without explicitly state where it was used. With exactly the same arguments as QME theory, the next several theorems in this section are immediate.

Theorem 4.2.4 (Consistency of IME)**.** *For any weighting matrix $\hat{W}$ with $\hat{W} \xrightarrow{p} W$, the interquantile mean matching estimator $\hat{\boldsymbol{\theta}}_n = \arg\min\, J_n(\boldsymbol{\theta})$ is consistent estimator of the true parameters $\boldsymbol{\theta}_0$.*

Proof. Adapted from theorem 4.1.6. □

By theorem 4.2.1 and theorem 4.2.2, we have $\sqrt{n}\, m_n(\boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}(0, \Omega)$ where:

$$\Omega = \Omega(\boldsymbol{\theta}_0) = \begin{bmatrix} \omega_1^2 & \omega_{21} & \cdots & \omega_{1K} \\ \omega_{21} & \omega_2^2 & \cdots & \omega_{2K} \\ \vdots & \vdots & \ddots & \vdots \\ \omega_{K1} & \omega_{K2} & \cdots & \omega_K^2 \end{bmatrix} \tag{4.46}$$

and the entries ω_i^2, ω_{ij} are supplied by these two theorems.

Theorem 4.2.5 (Asymptotic Normality of IME)**.** *For any weighting matrix $\hat{W}$ with $\hat{W} \xrightarrow{p} W$, the quantile matching estimator $\hat{\boldsymbol{\theta}}_n = \arg\min J_n(\boldsymbol{\theta})$ is asymptotically normal:*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N\left(0, (B^{\mathrm{T}}WB)^{-1}B^{\mathrm{T}}W\Omega WB(B^{\mathrm{T}}WB)^{-1}\right)$$

where $B = -\nabla m_n(\boldsymbol{\theta}_0)$ is discrete B-matrix at true parameters vector.

Proof. Adapted from theorem 4.1.7. □

Theorem 4.2.6 (Optimal Weight of IME)**.** *If weighting matrix $\hat{W} \xrightarrow{p} \Omega^{-1}$, then*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}(0, (B^{\mathrm{T}}\Omega^{-1}B)^{-1})$$

and the covariance matrix is the lowest among all choices of weighting matrix. In other words, the estimator corresponding to this weighting matrix is the most efficient.

Proof. Adapted from theorem 4.1.8 □

Two Step Interquantile Mean Matching Estimator

To construct asymptotically efficient IME estimator, we need a weighting matrix $\hat{W}$ that converges in probability to Ω^{-1}. To achieve this consistency for weighting matrix, we construct $\hat{W}$ as an estimate of Ω^{-1}. Given a preliminary weighting matrix

$\tilde{W}$, we construct IME estimator corresponding to $\tilde{W}$:

$$J_n(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} m_n(\boldsymbol{\theta})^{\mathrm{T}} \tilde{W} m_n(\boldsymbol{\theta}) \tag{4.47}$$

$$\tilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \ J_n(\boldsymbol{\theta}) \tag{4.48}$$

For any weighting matrix $\tilde{W}$, theorem 4.1.6 implies $\tilde{\boldsymbol{\theta}}$ is consistent. We compute estimate for covariance matrix by equation 4.46: $\tilde{\Omega} = \Omega(\tilde{\boldsymbol{\theta}})$ and put $\hat{W} = \tilde{\Omega}^{-1}$. Then it is clear that $\hat{W} \xrightarrow{p} \Omega^{-1}$. Now we are ready to conduct the second step:

$$J_n(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} m_n(\boldsymbol{\theta})^{\mathrm{T}} \hat{W} m_n(\boldsymbol{\theta}) \tag{4.49}$$

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \ J_n(\boldsymbol{\theta}) \tag{4.50}$$

Then one can show $\hat{\boldsymbol{\theta}}$ constructed by two step estimation is consistent and asymptotically most efficient.

Theorem 4.2.7 (Efficient Two-step IME)**.** *The two step estimator $\hat{\boldsymbol{\theta}}$ constructed above is consistent estimator of $\boldsymbol{\theta}_0$ and asymptotically normal:*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}(0, (B^{\mathrm{T}}\Omega^{-1}B)^{-1})$$

Proof. A direct application of theorem 4.2.5 with weighting matrix $\hat{W} \xrightarrow{p} \Omega^{-1}$. □

The result also show that the asymptotic behaviour of two-step estimator $\hat{\boldsymbol{\theta}}$ does not depend on the choice of preliminary weighting matrix $\hat{W}$. In finite samples, $\hat{\boldsymbol{\theta}}$, will be affected by the choice of $\hat{W}$. This undesirable dependency can be eliminated by interating the estimation sequence until convergence. This is called **iterated IME estimator**. Another strategy called **continuously-updated IME estimator** can be repeated in the same way as its counterpart in GMM.

After obtaining the two step estimator $\hat{\boldsymbol{\theta}}$, the estimate for asymptotic covariance

matrix of $\sqrt{n}\,\hat{\boldsymbol{\theta}}$ is $(\hat{B}^{\mathrm{T}}\hat{\Omega}^{-1}\hat{B})^{-1}$ converging in probability to $(B^{\mathrm{T}}\Omega^{-1}B)^{-1}$ where:

$$\hat{B} = -\nabla m_n(\hat{\boldsymbol{\theta}}) \xrightarrow{p} B = -\nabla m_n(\boldsymbol{\theta}_0) \tag{4.51}$$

$$\hat{\Omega} = \Omega(\hat{\boldsymbol{\theta}}) \xrightarrow{p} \Omega = \Omega(\boldsymbol{\theta}_0) \tag{4.52}$$

Hypothesis testing

The estimate of covariance matrix enables us to construct confidence intervals, confidence ellipsoid, hypothesis testing of parameters, and even functions of parameters (using Delta Method). Not only that, because GMM and IME are parallel in mathematical structure, Hansen-Sargan Test and Distance test are also applicable.

4.3 Joint Method of Quantile and Interquantile Mean Matching

If we are endowed with both quantile data and interquantile mean data, what is the best estimation strategy to utilise all the given information? The answer to this question is foreshadowed in terminologies used: quantile and interquantile. Successfully discover the statistical relationship between sample quantile and sample interquantile mean is the key to estimate parameters of income distribution with highest efficiency. The strategy again repeats arguments from Generalised Method of Moment (GMM) by recognising similarity of moment conditions and joint asymptotic behaviours of quantile and interquantile mean.

4.3.1 Statistical Model

A government agency is collecting income data and summarising the result by both quantiles and interquantile means. Let $Y_1, Y_2, \ldots, Y_n$ be the random income values collected by the agency, independently and identically distributed from a continuous distribution with CDF $F(\cdot;\boldsymbol{\theta})$ with d-dimensional vector of parameters. Suppose the CDF is well-bahaved: differentiable, exhibiting density function $f(\cdot;\boldsymbol{\theta})$, and quantile function $Q(\cdot;\boldsymbol{\theta})$. Two increasing sequences of $\pi_1, \pi_2, \ldots, \pi_{K_1} \in (0,1)$ and $\rho_0, \rho_1, \ldots, \rho_{K_2-1}, \rho_{K_2} \in [0,1]$ are fixed percentages points chosen by the government before performing data collection. An external observer then has access to a vector of sample quantiles $(\hat{\xi}_{\pi_1}, \hat{\xi}_{\pi_2}, \ldots, \hat{\xi}_{\pi_K})$, and a vector of sample interquantile means $\big(\hat{\mu}(\rho_0, \rho_1), \hat{\mu}(\rho_1, \rho_2), \ldots, \hat{\mu}(\rho_{K_2-1}, \rho_{K_2})\big)$ reported by the agency.

It can be shown that $\hat{\xi}_\pi$ is a strongly consistent estimator of ξ_π, under mild conditions. Under mild smoothness requirements on F in the neighbourhoods of points $\xi_{\pi_1}, \xi_{\pi_2}, \ldots, \xi_{\pi_{K_1}}$, the vector of sample quantiles $(\hat{\xi}_{\pi_1}, \hat{\xi}_{\pi_2}, \ldots, \hat{\xi}_{\pi_{K_1}})$ is asymptotically normal. Moreover, under mild regularity condition, $\hat{\mu}(\rho_j, \rho_{j+1})$ is a consistent estimator of $\mu(\rho_j, \rho_{j+1})$ and the vector of sample interquantile means

$$(\hat{\mu}(\rho_0, \rho_1), \hat{\mu}(\rho_1, \rho_2), \ldots, \hat{\mu}(\rho_{K_2-1}, \rho_{K_2}))$$

is asymptotically normal. Intuitively, the matching approach require us to find $\hat{\boldsymbol{\theta}} \in \mathbb{R}^d$, so that the vector of sample quantiles and the vector of sample interquantile means are as close as possible to their population counterparts.. If $d = K_1 + K_2\,(= K)$, we solve the system of equation with d unknowns and d equations, known as **exact-identification**. If $d > K$, we encounter the problem of **under-identification**, and it is impossible to estimate the parameters. Finally and most importantly, in the case of **over-identification** $K > d$, we have more equations than unknowns to solve, what estimation procedure should we employ to make the most of the abundance of quantile conditions and interquantile mean conditions?

4.3.2 Joint Estimator of Type *QI*

Fundamental Result of Quantile and Interquantile Mean

The following theorem shows asymptotic bivariate normality between a sample quantile and a sample interquantile mean. The intuition for this result comes from the fact that sample quantiles $(\hat{\xi})_{p\in(0,1)}$ are multivariate normal in limit, and sample interquantile mean can be expressed as linear combination of sample quantiles. The extension to asymptotic multivariate normality of multiple such statistics are straightforward.

Theorem 4.3.1 (Asymptotic Normality of Quantile and Interquantile Mean). *Suppose $0 < p_1 < p_2 < 1$, $q \in (0,1)$ and there are n iid random observations withdrawn from a distribution with distribution function F, density function f and quantile function Q. Let M_n be the sample (p_1, p_2)-interquantile mean , and $\hat{\xi}_{q,n}$ be the sample q-th quantile corresponding to the random sample, then $\sqrt{n}\big(M_n - \mu(p_1,p_2),\, \hat{\xi}_{q,n} - \xi_q\big)$ converges in distribution to a bivariate normal random vector, whose means are 0, variances are given in theorem 4.1.3 and theorem 4.2.1, and covariance is ψ/n where ψ is provided by formula:*

$$
\begin{aligned}
(1)\quad \psi &= \frac{q}{(p_2-p_1)f(Q(q))}\Big[(1-p_2)Q(p_2) - (1-p_1)Q(p_1) + (p_2-p_1)\mu(p_1,p_2)\Big] \\
&\quad \text{if } q \in (0, p_1) \\
(2)\quad \psi &= \frac{1-q}{(p_2-p_1)f(Q(q))}\Big[qQ(q) - p_1Q(p_1) - (q-p_1)\mu(p_1,q)\Big] \\
&\quad + \frac{q}{(p_2-p_1)f(Q(q))}\Big[(1-p_2)Q(p_2) - (1-q)Q(q) + (p_2-q)\mu(q,p_2)\Big] \\
&\quad \text{if } q \in [p_1, p_2] \\
(3)\quad \psi &= \frac{1-q}{(p_2-p_1)f(Q(q))}\Big[p_2Q(p_2) - p_1Q(p_1) - (p_2-p_1)\mu(p_1,p_2)\Big] \\
&\quad \text{if } q \in (p_2, 1)
\end{aligned}
$$

Proof. Using equation 4.31, the sample interquantile mean can be expressed as integral of sample quantiles:

$$M_n = \frac{1}{p_2 - p_1}\int_{p_1}^{p_2} \hat{\xi}_{p,n}\, dp \tag{4.53}$$

Using the fact that random vector $(\hat{\xi}_{p,n})_{p\in(0,1)}$ is asymptotically continuum-variate normal (take theorem 4.1.4 to the extreme) with mean vector $(\xi_p)_{p\in(0,1)}$ and covariance continuum matrix $(\sigma_{p,q}/n)_{p,q\in(0,1)}$ where:

$$\sigma_{p,q} = \frac{p \wedge q - pq}{f(Q(p))f(Q(q))}, \tag{4.54}$$

It must be the case that $(M_n, \hat{\xi}_{q,n})$ is asymptotically bivariate normal with mean

$\big(\mu(p_1, p_2), \xi_q\big)$ and covariance ψ/n where:

$$\psi = \frac{1}{p_2 - p_1} \int_{p_1}^{p_2} \frac{p \wedge q - pq}{f(Q(p))f(Q(q))}\, dp \tag{4.55}$$

To compute the intergal, there are 3 cases to consider:

$$\text{(i)}\, q \in (0, p_1) \qquad \text{(ii)}\, q \in [p_1, p_2] \qquad \text{(iii)}\, q \in (p_2, 1) \tag{4.56}$$

It is straightforward to compute simple integral for each case (using a change of variable $y = Q(p)$ and integration by part) and obtain the desired result. □

Construction & Properties of the Joint Estimator

The previous theorem are the necessary ingredient we need for the development of the Joint Quantile Interquantile Mean Estimator (QIME), which repeats the construction of GMM estimator. Here we develop an estimator that involves non-overlapping interquantile range only; however, the extension to overlapping interquantile ranges are straightforward. As a variation for moment condition vector, we define the joint discrepancy vector :

$$l_n(\boldsymbol{\theta}) = \begin{bmatrix} q_n(\boldsymbol{\theta}) \\ m_n(\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} \hat{\xi}_{\pi_1,n} - Q(\pi_1;\boldsymbol{\theta}) \\ \vdots \\ \hat{\xi}_{\pi_{K_1},n} - Q(\pi_{K_1};\boldsymbol{\theta}) \\ \hat{\mu}_n(\rho_0,\rho_1) - \mu(\rho_0,\rho_1;\boldsymbol{\theta}) \\ \vdots \\ \hat{\mu}_n(\rho_{K_2-1},\rho_{K_2}) - \mu(\rho_{K_2-1},\rho_{K_2};\boldsymbol{\theta}) \end{bmatrix} \tag{4.57}$$

where $q_n(\boldsymbol{\theta})$ and $m_n(\boldsymbol{\theta})$ are discrepancy vectors associated with quantile data and interquantile mean data. The size for $q_n(\boldsymbol{\theta})$ and $m_n(\boldsymbol{\theta})$ are $K_1 \times 1$ and $K_2 \times 1$, respectively.

I have to make up a new terminology at this point by introducing the notion of

asmoment condition[1]. By asmoment condition, I mean a sample statistic that is consistent to its true counterpart, and asymptotically normal with the other asmoment conditions. Let $K = K_1 + K_2$ be the length of the joint discrepancy vector. If $K = d$, the QIME estimator is defined as the parameter values $\hat{\boldsymbol{\theta}}$ that satisfies $l_n(\boldsymbol{\theta}) = 0$ (nullity of asmoment conditions). This is generally not possible when there are more asmoment conditions than parameters. For some $K \times K$ positive definite weighting matrix $\hat{W}$, define objective function:

$$J_n(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta})^{\mathrm{T}} \hat{W} l_n(\boldsymbol{\theta}) \tag{4.58}$$

This is a non-negative objective function generalising the notion of Euclidean length of quantile discrepancy vector. For example, if $\hat{W} = I_K$, then $J_n(\boldsymbol{\theta}) = \|l_n(\boldsymbol{\theta})\|^2$ coincide with Euclidean length. Typically, the objective function can assign different weights to different square discrepancy entries, and there are interaction terms between entries as well. QIME estimator is then defined as the minimiser of $J_n(\boldsymbol{\theta})$:

$$\hat{\boldsymbol{\theta}}_{\mathrm{QIME}} = \arg\min_{\boldsymbol{\theta}} \; J_n(\boldsymbol{\theta}) \tag{4.59}$$

The objective function defined in equation 4.58 is generally nonlinear. When $K = d$ (exact-identification), then there exists $\hat{\boldsymbol{\theta}}$ such that the joint conditions $l_n(\hat{\boldsymbol{\theta}}) = 0$. The QIME does not depend on the choice of weighting matrix $\hat{W}$. When $K > d$ (over-identification), the QIME depend on the weighting matrix and a numerical optimisation routine is required to solve for the maximiser.

Function $J_n : \mathbb{R}^d \to \mathbb{R}_+$ has gradient $\nabla J_n : \mathbb{R}^d \to \mathbb{R}^d$ and Hessian $\nabla^2 J_n : \mathbb{R}^d \to \mathbb{R}^{d\times d}$. Moreover, function $l_n : \mathbb{R}^d \to \mathbb{R}^K$ has Jacobian $\nabla l_n : \mathbb{R}^d \to \mathbb{R}^{K\times d}$ and Jacobian tranpose $\nabla^{\mathrm{T}} l_n : \mathbb{R}^d \to \mathbb{R}^{d\times K}$. The gradient and Hessian of J_n are:

$$\nabla J_n(\boldsymbol{\theta}) = 2\nabla^{\mathrm{T}} l_n(\boldsymbol{\theta}) \, \hat{W} \, l_n(\boldsymbol{\theta}) \tag{4.60}$$

$$\nabla^2 J_n(\boldsymbol{\theta}) = 2\nabla^{\mathrm{T}} l_n(\boldsymbol{\theta}) \, \hat{W} \, \nabla l_n(\boldsymbol{\theta}) + 2\mathcal{M} \odot \hat{W} \, l_n(\boldsymbol{\theta}) \tag{4.61}$$

[1] Asymptotic moment condition.

where $\mathcal{M}$ is a mysterious term, representing the total derivative of matrix-valued function $\nabla l_n : \mathbb{R}^d \to \mathbb{R}^{K\times d}$ and the multiplication rule for this term is also not proper. Moreover, it is clear the Jacobian of discrepancy function $l_n(\boldsymbol{\theta})$ is not a random function:

$$\nabla l_n(\boldsymbol{\theta}) = -\begin{bmatrix} \nabla q_n(\boldsymbol{\theta}) \\ \nabla m_n(\boldsymbol{\theta}) \end{bmatrix} = \begin{bmatrix} \nabla_{\boldsymbol{\theta}}^{\mathrm{T}} Q(\pi_1; \boldsymbol{\theta}) \\ \vdots \\ \nabla_{\boldsymbol{\theta}}^{\mathrm{T}} Q(\pi_{K_1}; \boldsymbol{\theta}) \\ \nabla_{\boldsymbol{\theta}}^{\mathrm{T}} \mu(\rho_0, \rho_1; \boldsymbol{\theta}) \\ \vdots \\ \nabla_{\boldsymbol{\theta}}^{\mathrm{T}} \mu(\rho_{K_2-1}, \rho_{K_2}; \boldsymbol{\theta}) \end{bmatrix} \tag{4.62}$$

and so is the $\mathcal{M}$ term. Denote $C(\boldsymbol{\theta}) = -\nabla l_n(\boldsymbol{\theta})$ and called **discrete C-matrix** from now on. In proving consistency and asymptotic normality of QIME, there is no need to invoke central limit theorem as in the case of GMM to show the $\mathcal{M}$ term and ∇l_n stabilises. The next several theorems in this section are immediate, followed by exactly the same arguments as QME theory.

Assumption 4.3.2 (No Multicollinearity)**.** $C = -\nabla l_n(\boldsymbol{\theta}_0)$ has independent columns at true parameters $\boldsymbol{\theta}_0$. Alternatively, C has no vector in the nullspace except 0.

We will implicitly assume this condition from now on because this condition is mild and are likely to be satisfied in all well-known distribution.

Theorem 4.3.3 (Consistency of QIME)**.** *For any weighting matrix $\hat{W}$ with $\hat{W} \xrightarrow{p} W$, the QIME estimator $\hat{\boldsymbol{\theta}}_n = \arg\min J_n(\boldsymbol{\theta})$ is a consistent estimator of the true parameters $\boldsymbol{\theta}_0$.*

Proof. Adapted from theorem 4.1.6. □

We have $\sqrt{n}\, l_n(\boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}(0, \Delta)$ where:

$$\Delta = \Delta(\boldsymbol{\theta}_0) = \begin{bmatrix} \Sigma & \Psi_a \\ \Psi_a^{\mathrm{T}} & \Omega \end{bmatrix} \tag{4.63}$$

The entries of Ω, Σ are provided in theorem 4.1.4, theorem 4.2.1 and theorem 4.2.2. The cross-covariance matrix:

$$\Psi_a = \begin{bmatrix} \psi_{11} & \psi_{11} & \cdots & \psi_{1K_2} \\ \psi_{21} & \psi_{22} & \cdots & \psi_{2K_2} \\ \vdots & \vdots & \cdots & \vdots \\ \psi_{K_1 1} & \psi_{K_1 2} & \cdots & \psi_{K_1 K_2} \end{bmatrix} \tag{4.64}$$

whose entries are provided in theorem 4.3.1

Theorem 4.3.4 (Asymptotic Normality of QIME)**.** *For any weighting matrix* $\hat{W}$ *with* $\hat{W} \xrightarrow{p} W$*, the QIME estimator* $\hat{\boldsymbol{\theta}}_n = \arg\min\ J_n(\boldsymbol{\theta})$ *is asymptotically normal:*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N\left(0, (C^{\mathrm{T}}WC)^{-1}C^{\mathrm{T}}W\Delta WC(C^{\mathrm{T}}WC)^{-1}\right)$$

where $C = -\nabla l_n(\boldsymbol{\theta}_0)$ *is discrete C-matrix at true parameters vector.*

Proof. Adapted from theorem 4.1.7. □

Theorem 4.3.5 (Optimal Weight of QIME)**.** *If weighting matrix* $\hat{W} \xrightarrow{p} \Delta^{-1}$*, then*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}(0, (C^{\mathrm{T}}\Delta^{-1}C)^{-1})$$

and the covariance matrix is the lowest among all choices of weighting matrix. In other words, the estimator corresponding to this weighting matrix is the most efficient.

Proof. Adapted from theorem 4.1.8. □

Two Step Joint Quantile Interquantile Mean Matching Estimator

To construct asymptotically efficient QIME estimator, we need a weighting matrix $\hat{W}$ that converges in probability to Δ^{-1}. To achieve this consistency for weighting matrix, we construct $\hat{W}$ as an estimate of Δ^{-1}. Given a preliminary weighting matrix

$\tilde{W}$, we construct QIME estimator corresponding to $\tilde{W}$:

$$J_n(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta})^{\mathrm{T}} \tilde{W} l_n(\boldsymbol{\theta}) \tag{4.65}$$

$$\tilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \ J_n(\boldsymbol{\theta}) \tag{4.66}$$

For any weighting matrix $\tilde{W}$, theorem 4.3.3 implies $\tilde{\boldsymbol{\theta}}$ is consistent. We compute the estimate for covariance matrix by equation 4.63: $\tilde{\Delta} = \Delta(\tilde{\boldsymbol{\theta}})$ and put $\hat{W} = \tilde{\Delta}^{-1}$. Then it is clear that $\hat{W} \xrightarrow{p} \Delta^{-1}$. Now we are ready to conduct the second step:

$$J_n(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta})^{\mathrm{T}} \hat{W} l_n(\boldsymbol{\theta}) \tag{4.67}$$

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \ J_n(\boldsymbol{\theta}) \tag{4.68}$$

Then one can show $\hat{\boldsymbol{\theta}}$ constructed by two step estimation is consistent and asymptotically most efficient.

Theorem 4.3.6 (Efficient Two-step QIME)**.** *The two step joint estimator $\hat{\boldsymbol{\theta}}$ constructed above is consistent estimator of $\boldsymbol{\theta}_0$ and asymptotically normal:*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}(0, (C^{\mathrm{T}} \Delta C)^{-1})$$

Proof. A direct application of theorem 4.3.4 with weighting matrix $\hat{W} \xrightarrow{p} \Delta^{-1}$ □

The result also show that the asymptotic behaviour of two-step estimator $\hat{\boldsymbol{\theta}}$ does not depend on the choice of preliminary weighting matrix $\hat{W}$. In finite samples, $\hat{\boldsymbol{\theta}}$, will be affected by the choice of $\hat{W}$. This undesirable dependency can be eliminated by iterating the estimation sequence until convergence. This is called **iterated QIME estimator**. Another strategy called **continuously-updated QIME estimator** can be repeated in the same way as its counterpart in GMM.

After obtaining the two step estimator $\hat{\boldsymbol{\theta}}$, the estimate for asymptotic covariance

matrix of $\sqrt{n}\,\hat{\boldsymbol{\theta}}$ is $(\hat{C}^{\mathrm{T}}\hat{\Delta}^{-1}\hat{C})^{-1}$ converging in probability to $(C^{\mathrm{T}}\Delta^{-1}C)^{-1}$ where:

$$\hat{C} = -\nabla l_n(\hat{\boldsymbol{\theta}}) \xrightarrow{p} C = -\nabla l_n(\boldsymbol{\theta}_0) \tag{4.69}$$

$$\hat{\Delta} = \Delta(\hat{\boldsymbol{\theta}}) \xrightarrow{p} \Delta = \Delta(\boldsymbol{\theta}_0) \tag{4.70}$$

Hypothesis testing

Again, the estimate of covariance matrix allows us to construct confidence intervals, confidence ellipsoid, hypothesis testing of parameters, and even functions of parameters (using Delta method). Hansen-Sargan Test and Distance test are also applicable.

4.4 Method of Proportion Matching

One popular method of summarising income data, which has been introduced in chapter 3, is using class frequency. Income range are partitioned into multiple class intervals and the information of how many people having income within each class interval is reported. The method repeats arguments from GMM theory by recognising identical pattern of moment conditions and asymptotic behaviour of sample proportions. Method of proportion matching serves as an alternative procedure for multinomial MLE introduced in previous chapter, and then provides new insights on the Pearson minimum χ^2 technique introduced in Hinkley and Cox (1979).

4.4.1 Statistical Model for Class Frequency Data

The detailed description for the model are given in chapter 3, section 3.1. For the sake of notation convenience, I intentionally adjust the number of income classes from K to $K + 1$. The model can be summarised as follows:

$$\textbf{K+1 income intervals} \quad [t_0, t_1),\ [t_1, t_2), \ldots, [t_K, t_{K+1}),\ t_0 = 0, t_{K+1} = \infty \tag{4.71}$$

$$\textbf{Exact incomes} \quad Z_1, Z_2, \ldots, Z_n \overset{iid}{\sim} F(\cdot; \boldsymbol{\theta}_0) \tag{4.72}$$

$$\textbf{Probabilities Vector} \quad \mathbf{p}(\boldsymbol{\theta}) = (p_1(\boldsymbol{\theta}), p_2(\boldsymbol{\theta}), ..., p_{K+1}(\boldsymbol{\theta})) \tag{4.73}$$

$$p_j(\boldsymbol{\theta}) = F(t_j; \boldsymbol{\theta}) - F(t_{j-1}; \boldsymbol{\theta}),\ j = 1, 2, \ldots K+1 \tag{4.74}$$

$$\textbf{Multinomial Counts} \quad \mathbf{N} = (N_1, N_2, \ldots, N_{K+1}) \sim \mathrm{Mult}(n, \mathbf{p}(\boldsymbol{\theta}_0)) \tag{4.75}$$

$$\textbf{Sample Proportions} \quad \hat{\mathbf{P}} = (\hat{p}_1, \hat{p}_2, \ldots, \hat{p}_{K+1}), \quad \hat{p}_i = N_i/N \tag{4.76}$$

4.4.2 Proportion Matching Estimator

The matching strategy I present next is for distributions whose support is $(0, \infty)$ and for non-overlapping class intervals. However, the extensions to general support and overlapping class intervals are straightforward.

Fundamental Asymptotic Result of Sample Proportions

Theorem 4.4.1 (Asymptotic Behaviour of Sample Proportions)**.** *Random vector* $\boldsymbol{N} = (N_1, N_2, \ldots, N_K, N_{K+1}) \sim \mathrm{Mult}(n, \boldsymbol{q})$ *where* $\boldsymbol{q} = (p_1, p_2, \ldots, p_{K+1})$ *adding up to unity. Let* $\hat{\boldsymbol{P}} = (N_1, N_2, \ldots N_K)/n$ *be vector of sample proportions and* $\boldsymbol{p} = (p_1, p_2, \ldots p_K)$ *be vector of true proportions, excluding the last proportion to avoid linear dependency. Then:* $\hat{\boldsymbol{P}} \xrightarrow{p} \boldsymbol{p}$ *and* $\sqrt{n}(\hat{\boldsymbol{P}} - \boldsymbol{p}) \xrightarrow{d} \mathrm{N}(0, \Pi)$ *where*

$$\Pi = \begin{bmatrix} p_1(1-p_1) & -p_1 p_2 & \cdots & -p_1 p_K \\ -p_2 p_1 & p_2(1-p_2) & \cdots & -p_2 p_K \\ \vdots & \vdots & \cdots & \vdots \\ -p_K p_1 & -p_K p_2 & \cdots & p_K(1-p_K) \end{bmatrix}$$

Proof. $\mathrm{Mult}(n, \mathbf{q})$ random vector can be viewed as iid sum of $\mathrm{Mult}(1, \mathbf{q})$ random binary vectors. The covariance matrix for $\mathrm{Mult}(1, \mathbf{q})$ excluding the last random entry is exactly the matrix Π given above. The theorem is then an immediate consequence of vector LLN and vector CLT. □

Construction & Properties of Proportion Matching Estimator

Our objective is again repeating the theory of GMM to construct PME. Here we developed an estimator specifically for non-overlapping class intervals; however, the extension to overlapping intervals are straightforward. As an analogy for moment conditions vector, we define proportions discrepancy vector:

$$f_n(\boldsymbol{\theta}) = \hat{\mathbf{P}} - \mathbf{p}(\boldsymbol{\theta}) = \begin{bmatrix} \hat{p}(t_0, t_1) - p(t_0, t_1; \boldsymbol{\theta}) \\ \hat{p}(t_1, t_2) - p(t_1, t_2; \boldsymbol{\theta}) \\ \vdots \\ \hat{p}(t_{K-1}, t_K) - p(t_{K-1}, t_K; \boldsymbol{\theta}) \end{bmatrix} \tag{4.77}$$

where $\hat{p}_j = \hat{p}(t_{j-1}, t_j)$ and $p_j(\boldsymbol{\theta}) = p(t_{j-1}, t_j; \boldsymbol{\theta})$ (alternative notations). Using new notation, by theorem 4.4.1, we have $f_n(\boldsymbol{\theta}_0) \xrightarrow{p} 0$ and $\sqrt{n} f_n(\boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}(0, \Pi(\boldsymbol{\theta}_0))$.

If $K = d$, the PME estimator is defined as the parameter values $\hat{\boldsymbol{\theta}}$ that satisfies $f_n(\boldsymbol{\theta}) = 0$. This is generally not possible when there are more proportion conditions than parameters. For some $K \times K$ positive definite weighting matrix $\hat{W}$, define objective function:

$$J_n(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} f_n(\boldsymbol{\theta})^{\mathrm{T}} \hat{W} f_n(\boldsymbol{\theta}) \tag{4.78}$$

This is a nonegative objective function generalising the notion of Euclidean length of quantile discrepancy vector. For example, if $\hat{W} = I_K$, then $J_n(\boldsymbol{\theta}) = \|f_n(\boldsymbol{\theta})\|^2$ coincide with Euclidean length. Typically, the objective function can assign different weights to different square discrepancy entries, and there are interaction terms between entries as well. PME estimator is the minimiser of $J_n(\boldsymbol{\theta})$:

$$\hat{\boldsymbol{\theta}}_{\mathrm{PME}} = \arg\min_{\boldsymbol{\theta}} \ J_n(\boldsymbol{\theta}) \tag{4.79}$$

Function $J_n : \mathbb{R}^d \to \mathbb{R}_+$ has gradient $\nabla J_n : \mathbb{R}^d \to \mathbb{R}^d$ and Hessian $\nabla^2 J_n : \mathbb{R}^d \to \mathbb{R}^{d \times d}$. Moreover, function $f_n : \mathbb{R}^d \to \mathbb{R}^K$ has Jacobian $\nabla f_n : \mathbb{R}^d \to \mathbb{R}^{K \times d}$ and Jacobian tranpose $\nabla^{\mathrm{T}} f_n : \mathbb{R}^d \to \mathbb{R}^{d\times K}$. The gradient and Hessian of J_n are:

$$\nabla J_n(\boldsymbol{\theta}) = 2\nabla^{\mathrm{T}} f_n(\boldsymbol{\theta})\, \hat{W}\, f_n(\boldsymbol{\theta}) \tag{4.80}$$

$$\nabla^2 J_n(\boldsymbol{\theta}) = 2\nabla^{\mathrm{T}} f_n(\boldsymbol{\theta})\, \hat{W}\, \nabla f_n(\boldsymbol{\theta}) + 2\mathcal{M} \odot \hat{W}\, f_n(\boldsymbol{\theta}) \tag{4.81}$$

One feature in PME that I find easier than GMM is that the Jacobian of discrepancy function $f_n(\boldsymbol{\theta})$ is no longer a random function:

$$\nabla f_n(\boldsymbol{\theta}) = -\begin{bmatrix} \nabla^{\mathrm{T}}_{\boldsymbol{\theta}} p(t_0, t_1; \boldsymbol{\theta}) \\ \nabla^{\mathrm{T}}_{\boldsymbol{\theta}} p(t_1, t_2; \boldsymbol{\theta}) \\ \vdots \\ \nabla^{\mathrm{T}}_{\boldsymbol{\theta}} p(t_{K-1}, t_K; \boldsymbol{\theta}) \end{bmatrix} = -\begin{bmatrix} \nabla^{\mathrm{T}} p_1(\boldsymbol{\theta}) \\ \nabla^{\mathrm{T}} p_2(\boldsymbol{\theta}) \\ \vdots \\ \nabla^{\mathrm{T}} p_K(\boldsymbol{\theta}) \end{bmatrix} \tag{4.82}$$

and so is the $\mathcal{M}$ term. Denote $D(\boldsymbol{\theta}) = -\nabla f_n(\boldsymbol{\theta})$ and called **discrete D-matrix** from now on. The next several theorems in this section are immediate, followed by exactly the same arguments as QME theory.

Assumption 4.4.2 (No Multicollinearity)**.** $D = -\nabla f_n(\boldsymbol{\theta}_0)$ has independent columns at true parameters $\boldsymbol{\theta}_0$. Alternatively, D has no vector in the nullspace except 0.

This is a mild regularity condition since every well-known distributions satisfy this assumption. We will implicitly assume this condition from now on.

Theorem 4.4.3 (Consistency of PME)**.** *For any weighting matrix $\hat{W}$ with $\hat{W} \xrightarrow{p} W$, the PME $\hat{\boldsymbol{\theta}}_n = \arg\min\, J_n(\boldsymbol{\theta})$ is consistent estimator of the true parameters $\boldsymbol{\theta}_0$.*

Proof. Adapted from theorem 4.1.6. □

By theorem 4.4.1, we have $\sqrt{n}\, f_n(\boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}(0, \Pi)$ where:

$$\Pi = \Pi(\boldsymbol{\theta}_0) = \mathrm{diag}(\mathbf{p}(\boldsymbol{\theta}_0)) - \mathbf{p}(\boldsymbol{\theta}_0)\mathbf{p}(\boldsymbol{\theta}_0)^{\mathrm{T}} \tag{4.83}$$

Theorem 4.4.4 (Asymptotic Normality of PME)**.** *For any weighting matrix $\hat{W}$ with*

$\hat{W} \xrightarrow{p} W$, *the PME estimator* $\hat{\boldsymbol{\theta}}_n = \arg\min\ J_n(\boldsymbol{\theta})$ *is asymptotically normal:*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N\left(0, (D^{\mathrm{T}}WD)^{-1}D^{\mathrm{T}}W\Pi WD(D^{\mathrm{T}}WD)^{-1}\right)$$

where $D = -\nabla f_n(\boldsymbol{\theta}_0)$ *is discrete D-matrix at true parameters vector.*

Proof. Adapted from theorem 4.1.7. □

Theorem 4.4.5 (Optimal Weight of PME)**.** *If weighting matrix* $\hat{W} \xrightarrow{p} \Pi^{-1}$, *then*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}(0, (D^{\mathrm{T}}\Pi^{-1}D)^{-1})$$

and the covariance matrix is the lowest among all choices of weighting matrix. In other words, the estimator corresponding to this weighting matrix is the most efficient.

Proof. Adapted from theorem 4.1.8 □

Two Step Proportion Matching Estimator

To construct asymptotically efficient PME estimator, we need a weighting matrix $\hat{W}$ that converges in probability to Π^{-1}. To achieve this consistency for weighting matrix, we construct $\hat{W}$ as an estimate of Π^{-1}. Given a preliminary weighting matrix $\tilde{W}$, we construct PME estimator corresponding to $\tilde{W}$:

$$J_n(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} f_n(\boldsymbol{\theta})^{\mathrm{T}}\tilde{W}f_n(\boldsymbol{\theta}) \tag{4.84}$$

$$\tilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}}\ J_n(\boldsymbol{\theta}) \tag{4.85}$$

For any weighting matrix $\tilde{W}$, theorem 4.1.6 implies $\tilde{\boldsymbol{\theta}}$ is consistent. We compute estimate for covariance matrix by equation 4.83: $\tilde{\Pi} = \Pi(\tilde{\boldsymbol{\theta}})$ and put $\hat{W} = \tilde{\Pi}^{-1}$. Then it is clear that $\hat{W} \xrightarrow{p} \Pi^{-1}$. Now we are ready to conduct the second step:

$$J_n(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} f_n(\boldsymbol{\theta})^{\mathrm{T}} \hat{W} f_n(\boldsymbol{\theta}) \tag{4.86}$$

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \; J_n(\boldsymbol{\theta}) \tag{4.87}$$

Then one can show $\hat{\boldsymbol{\theta}}$ constructed by two step estimation is consistent and asymptotically most efficient.

Theorem 4.4.6. *The two step estimator $\hat{\boldsymbol{\theta}}$ constructed above is consistent estimator of $\boldsymbol{\theta}_0$ and asymptotically normal:*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}(0, (D^{\mathrm{T}} \Pi^{-1} D)^{-1})$$

Proof. A direct application of theorem 4.4.4 with weighting matrix $\hat{W} \xrightarrow{p} \Pi^{-1}$. □

The result also show that the asymptotic behaviour of two-step estimator $\hat{\boldsymbol{\theta}}$ does not depend on the choice of preliminary weighting matrix $\hat{W}$. In finite samples, $\hat{\boldsymbol{\theta}}$, will be affected by the choice of $\hat{W}$. This undesirable dependency can be eliminated by iterating the estimation sequence until convergence. This is called **iterated PME estimator**. Another strategy called **continuously-updated PME estimator** can be repeated in the same way as its counterpart in GMM.

After obtaining the two step estimator $\hat{\boldsymbol{\theta}}$, the estimate for asymptotic covariance matrix of $\sqrt{n}\,\hat{\boldsymbol{\theta}}$ is $(\hat{D}^{\mathrm{T}} \hat{\Pi}^{-1} \hat{D})^{-1}$ converging in probability to $(D^{\mathrm{T}} \Pi^{-1} D)^{-1}$ where:

$$\hat{D} = -\nabla f_n(\hat{\boldsymbol{\theta}}) \xrightarrow{p} D = -\nabla f_n(\boldsymbol{\theta}_0) \tag{4.88}$$

$$\hat{\Pi} = \Pi(\hat{\boldsymbol{\theta}}) \xrightarrow{p} \Pi = \Pi(\boldsymbol{\theta}_0) \tag{4.89}$$

Hypothesis testing

Confidence intervals, confidence ellipsoid, hypothesis testing can be done using the consistent estimate of covariance matrix. Hansen-Sargan Test and Distance test are also applicable.

4.5 Joint Method of Grouped Data

If we are endowed with *class frequency data*, *quantile data* and *interquantile mean data*, what is the best estimation strategy to utilise all the given information? Successfully discovering the asymptotic relationships between sample proportion and sample quantile with sample interquantile mean is the key to estimate parameters of income distribution with highest efficiency. The strategy again repeats arguments from Generalised Method of Moment (GMM) by recognising similarity of moment conditions and joint asymptotic behaviours of quantile and interquantile mean and sample proportion.

4.5.1 Statistical Model

Refer to subsections 4.3.1 for the joint model of quantile data and interquantile mean data, and subsection 4.4.1 for the model of proportion data.

4.5.2 Joint Estimator of Type *QIP*

Fundamental Results of Proportion, Quantile and Interquantile Mean

Suppose $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} F$, then the empirical CDF of F, evaluated at y is defined as:

$$\hat{F}_n(b) = \frac{1}{n}\sum_{i=1}^{n} I_{\{Y_i \leq b\}} \tag{4.90}$$

It is simple to see that $n\hat{F}_n(b) \sim \text{Bin}(n, F(b))$. Similarly, for $a < b$ one can defined:

$$\hat{F}_n(a,b) = \frac{1}{n}\sum_{i=1}^{n} I_{\{a < Y_i \leq b\}} \tag{4.91}$$

Again, we have $n\hat{F}_n(a,b) \sim \text{Bin}(n, F(a,b))$ where $F(a,b) = F(b) - F(a) \in [0,1]$. There is a slight difference between the definitions of $\hat{F}_n(a,b)$ and $\hat{p}_n(a,b)$ (introduced in PME theory) appearing in the class bounds. In fact, $\hat{p}_n(a,b)$ is the sample proportion in $[a,b)$ instead of $(a,b]$. However, as the sample size tends to infinity, the difference becomes unnoticeable. We are now ready to state the first theorem concerning joint distribution of empirical CDF proportions.

Theorem 4.5.1 (Asymptotic Normality of Empirical CDFs)**.** *Suppose* $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} F$ *and* $\hat{F}_n(a)$, $\hat{F}_n(b)$ *are empirical CDF proportions where* $a < b$, *then:*

$$\sqrt{n}\left(\begin{bmatrix}\hat{F}_n(a)\\ \hat{F}_n(b)\end{bmatrix} - \begin{bmatrix}F(a)\\ F(b)\end{bmatrix}\right) \xrightarrow{d} N\left(\begin{bmatrix}0\\ 0\end{bmatrix}, \begin{bmatrix}F(a)(1-F(a)) & F(a)(1-F(b))\\ F(a)(1-F(b)) & F(b)(1-F(b))\end{bmatrix}\right)$$

Proof. With similar argument used in theorem 4.4.1, we have:

$$\sqrt{n}\left(\begin{bmatrix}\hat{F}_n(a)\\ \hat{F}_n(a,b)\end{bmatrix} - \begin{bmatrix}F(a)\\ F(a,b)\end{bmatrix}\right) \xrightarrow{d} N\left(\begin{bmatrix}0\\ 0\end{bmatrix}, \begin{bmatrix}F(a)(1-F(a)) & -F(a)F(a,b)\\ -F(a)F(a,b) & F(a,b)(1-F(a,b))\end{bmatrix}\right) \tag{4.92}$$

Since the map $(x, y) \mapsto (x, x + y)$ is a linear invertible map with Jacobian:

$$J = \begin{bmatrix}1 & 0\\ 1 & 1\end{bmatrix}, \tag{4.93}$$

the resulting random vector $(\hat{F}_n(a), \hat{F}_n(b))$ that was sent from $(\hat{F}_n(a), \hat{F}_n(a, b))$ is asymptotically bivariate normal with mean $(F(a), F(b))$. The asymptotic variances associated with $\hat{F}_n(a)$ and $\hat{F}_n(a, b)$ are clear because $n\hat{F}_n(a) \sim \text{Bin}(n, F(a))$ and $n\hat{F}_n(a, b) \sim \text{Bin}(n, F(a, b))$. As $\hat{F}_n(b) = \hat{F}_n(a) + \hat{F}_n(a, b)$, the asymptotic covariance associated with $\hat{F}_n(a)$ and $\hat{F}_n(b)$ is:

$$F(a)(1 - F(a)) - F(a)F(a, b) = F(a)(1 - F(b)) \tag{4.94}$$

which completes the proof. □

There are two simple remarks. Firstly, if it is not certain $a < b$, one can use the expression:

$$F(a) \wedge F(b) - F(a)F(b) \tag{4.95}$$

for the asymptotic covariance instead. Secondly, the theorem only shows asymptotic behaviour of two empirical CDF proportions, but the extension to multiple such sample statistics is straightforward. In this case, the asymptotic distribution of multiple sample proportions is multivariate normal with known covariance structure. I hope your intuition says it is correct.

The next important asymptotic result is due to Bahadur (1966), revealing a *strong duality* between sample quantile and empirical CDF proportion. By strong duality, I mean in proving asymptotic behaviours with magnifier $n^{1/2}$, one is still able treat sample quantile as empirical CDF and vice versa.

Theorem 4.5.2 (Bahadur)**.** *Let $p \in (0,1)$. Suppose that the distribution function F is twice differentiable at ξ_p, with $F'(\xi_p) = f(\xi_p) > 0$. Then:*

$$\hat{\xi}_{p,n} - \xi_p = \frac{p - \hat{F}_n(\xi_p)}{f(\xi_p)} + O_p(n^{-3/4}(\ln n)^{3/4})$$

Proof. See Serfling (2009, p. 91) or Bahadur (1966). □

Theorem 4.5.3 (Asymptotic Normality of Empirical CDF and Quantile)**.** *Let $p \in (0,1)$. $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} F$ with density function f and quantile function Q. Let $\hat{\xi}_{p,n}$ be the sample p^{th} quantile and $\hat{F}_n(a)$ be the empirical CDF proportion corresponding to the iid sample of size n, then:*

$$\sqrt{n}\left(\begin{bmatrix}\hat{\xi}_{p,n}\\ \hat{F}_n(a)\end{bmatrix} - \begin{bmatrix}\xi_p\\ F(a)\end{bmatrix}\right) \xrightarrow{d} N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\frac{p(1-p)}{f(Q(p))^2} & -\frac{p\wedge F(a) - pF(a)}{f(Q(p))}\\ -\frac{p\wedge F(a)-pF(a)}{f(Q(p))} & F(a)(1-F(a))\end{bmatrix}\right)$$

Proof. Let $\hat{\xi}_{F(a),n}$ be the sample $F(a)$-quantile corresponding to the sample of size n. Using theorem 4.1.4:

$$\sqrt{n}\left(\begin{bmatrix}\hat{\xi}_{p,n}\\ \hat{\xi}_{F(a),n}\end{bmatrix} - \begin{bmatrix}\xi_p\\ \xi_{F(a)}\end{bmatrix}\right) \xrightarrow{d} N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}\frac{p(1-p)}{f(Q(p))^2} & \frac{p\wedge F(a)-pF(a)}{f(Q(p))f(a)}\\ \frac{p\wedge F(a)-pF(a)}{f(Q(p))f(a)} & \frac{F(a)(1-F(a))}{f^2(a)}\end{bmatrix}\right) \tag{4.96}$$

Consider linear transformation $(x, y) \mapsto (x, -f(a)y)$. By theorem 4.5.2, up to a "noise" component of order $O_p(n^{-3/4}(\ln n)^{3/4})$, the transformation sends vector

$\left(\hat{\xi}_{p,n},\, \hat{\xi}_{F(a),n}\right)$ to $\left(\hat{\xi}_{p,n},\, \hat{F}_n(a)\right)$, and vector $\left(\xi_p, \xi_{F(a)}\right)$ to $(\xi_p, F(a))$. Since the noise component is negligible even if magnified by $\sqrt{n}$, the Delta Method 4.1.12 is applicable:

$$\sqrt{n}\left(\begin{bmatrix}\hat{\xi}_{p,n}\\ \hat{F}_n(a)\end{bmatrix} - \begin{bmatrix}\xi_p\\ F(a)\end{bmatrix}\right) \xrightarrow{d} N\left(0, J\Sigma J^{\mathrm{T}}\right) \tag{4.97}$$

where Σ is the covariance matrix that appears in equation 4.96, and J is the Jacobian matrix of the transformation at (ξ_p, a):

$$J = \begin{bmatrix}1 & 0\\ 0 & -f(a)\end{bmatrix} \tag{4.98}$$

Simple computation of matrix product yields the desired result. □

Theorem 4.5.4 (Asymptotic Normality of Proportion and Quantile)**.** *Let $p \in (0,1)$. $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} F$ with density function f and quantile function Q. Let $\hat{\xi}_{p,n}$ be the sample p^{th} quantile and $\hat{F}_n(a,b)$ be the empirical proportion in interval $(a,b]$ corresponding to the iid sample of size n, then $\left(\hat{\xi}_{p,n}, \hat{F}_n(a,b)\right)$ is asymptotically bivariate normal with mean $(\xi_p, F(a,b))$ and covariance ψ/n where:*

$$\psi = \frac{p \wedge F(a) - pF(a)}{f(Q(p))} - \frac{p \wedge F(b) - pF(b)}{f(Q(p))}$$

Proof. Write $\hat{F}_n(a,b) = \hat{F}_n(b) - \hat{F}_n(a)$ and the conclusion follows immediately from theorem 4.5.3 and the bilinear property of covariance. □

Let do a simple check for the theorem 4.5.4 to see how it agrees with our intuition. Consider $p < F(a) < F(b)$, then the covariance numerator is $pF(b) - pF(a) > 0$. If many observations are within $(a,b]$, then it seems likely that high $\hat{F}_n(a,b)$ will drag the $\hat{\xi}_p$ up, causing positive correlation. If $F(a) < F(b) < p$, then the covariance numerator is $(p-1)(F(b)-F(a)) < 0$. Indeed, many observations within $(a,b]$ now tend to drag the $\hat{\xi}_p$ down, causing negative correlation. In the case of $F(a) \leq p \leq F(b)$, the covariance sign is ambiguous.

Theorem 4.5.5 (Asymptotic Normality of Interquantile Mean and Proportion). *Let $p_1, p_2 \in (0,1)$ with $p_1 < p_2$. $Y_1, Y_2, \ldots, Y_n \overset{iid}{\sim} F$ with density function f and quantile function Q. Let $\hat{\mu}_n(p_1, p_2)$ be the sample (p_1, p_2)-interquantile mean and $\hat{F}_n(a, b)$ be the empirical proportion between a and b corresponding to the iid sample of size n, then $\left(\hat{\mu}_n(p_1, p_2), \hat{F}_n(a, b)\right)$ is asymptotically bivariate normal with mean $(\mu(p_1, p_2), F(a, b))$ and covariance ψ/n where:*

$$\psi = \frac{1}{p_2 - p_1} \int_{p_1}^{p_2} \frac{p \wedge F(a) - pF(a)}{f(Q(p))} - \frac{p \wedge F(b) - pF(b)}{f(Q(p))} \, dp$$

Proof. We use the same proof technique in theorem 4.2.1, expressing the sample interquantile mean as:

$$\hat{\mu}_n(p_1, p_2) = \frac{1}{p_2 - p_1} \int_{p_1}^{p_2} \hat{\xi}_{p,n} \, dp \tag{4.99}$$

The conclusion follows from bilinear property of covariance and theorem 4.5.4. □

The asymptotic covariance between sample interquantile mean and sample proportion derived in previous theorem is expressed in integral form and can be computed symbolically by considering 6 distinct configuration. Each ordering of p_1, p_2, a, b given $p_1 < p_2$ and $a < b$ will result in a different formula and computational details are deferred to Appendix A.4.

The extension to multiple sample statistics are straightforward in the previous five theorems. In this setting, the random vector of these sample statistics are consistent to their population counterparts and asymptotically normal with known covariance structure. I hope your intuition says the extension is valid.

Construction & Properties of the Joint Estimator of type QIP

The previous theorems important role in the development of the QIPME , which repeats the construction of GMM estimator. As a variation for moment condition vector, we define the joint discrepancy vector:

$$l_n(\boldsymbol{\theta}) = \begin{bmatrix} q_n(\boldsymbol{\theta}) \\ m_n(\boldsymbol{\theta}) \\ f_n(\boldsymbol{\theta}) \end{bmatrix} \tag{4.100}$$

where $q_n(\boldsymbol{\theta})$, $m_n(\boldsymbol{\theta})$ and $f_n(\boldsymbol{\theta})$ are discrepancy vectors associated with quantile data and interquantile mean data, and sample proportion data (computed from class frequency data). The size for $q_n(\boldsymbol{\theta})$, $m_n(\boldsymbol{\theta})$ and $f_n(\boldsymbol{\theta})$ are $K_1 \times 1$ and $K_2 \times 1$ and $K_3 \times 1$, respectively.

Let $K = K_1 + K_2 + K_3$ be the length of the joint discrepancy vector. If $K = d$, the QIPME estimator is defined as the parameter values $\hat{\boldsymbol{\theta}}$ that satisfies $l_n(\boldsymbol{\theta}) = 0$ (nullity of asmoment conditions). This is generally not possible when there are more asmoment conditions than parameters. For some $K \times K$ positive definite weighting matrix $\hat{W}$, define objective function:

$$J_n(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta})^{\mathrm{T}} \hat{W} l_n(\boldsymbol{\theta}) \tag{4.101}$$

This is a non-negative objective function generalising the notion of Euclidean length of quantile discrepancy vector. For example, if $\hat{W} = I_K$, then $J_n(\boldsymbol{\theta}) = \|l_n(\boldsymbol{\theta})\|^2$ coincide with Euclidean length. Typically, the objective function can assign different weights to different square discrepancy entries, and there are interaction terms between entries as well. QIPME estimator is then defined as the minimiser of $J_n(\boldsymbol{\theta})$:

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \ J_n(\boldsymbol{\theta}) \tag{4.102}$$

If $K = d$, then there is $\hat{\boldsymbol{\theta}}$ such that the joint conditions $l_n(\hat{\boldsymbol{\theta}}) = 0$. The QIPME does not depend on the weighting matrix $\hat{W}$. The objective function defined in equation 4.3 is generally nonlinear and a numerical optimisation routine is required to solve for the maximiser.

Function $J_n : \mathbb{R}^d \to \mathbb{R}_+$ has gradient $\nabla J_n : \mathbb{R}^d \to \mathbb{R}^d$ and Hessian $\nabla^2 J_n : \mathbb{R}^d \to$

$\mathbb{R}^{d\times d}$. Moreover, function $l_n : \mathbb{R}^d \to \mathbb{R}^K$ has Jacobian $\nabla l_n : \mathbb{R}^d \to \mathbb{R}^{K\times d}$ and Jacobian tranpose $\nabla^{\mathrm{T}} l_n : \mathbb{R}^d \to \mathbb{R}^{d\times K}$. The gradient and Hessian of J_n are:

$$\nabla J_n(\boldsymbol{\theta}) = 2\nabla^{\mathrm{T}} l_n(\boldsymbol{\theta})\, \hat{W}\, l_n(\boldsymbol{\theta}) \tag{4.103}$$

$$\nabla^2 J_n(\boldsymbol{\theta}) = 2\nabla^{\mathrm{T}} l_n(\boldsymbol{\theta})\, \hat{W}\, \nabla l_n(\boldsymbol{\theta}) + 2\mathcal{M} \odot \hat{W}\, l_n(\boldsymbol{\theta}) \tag{4.104}$$

where $\mathcal{M}$ is a mysterious term, representing the total derivative of matrix-valued function $\nabla l_n : \mathbb{R}^d \to \mathbb{R}^{K\times d}$ and the multiplication rule for this term is also not proper. Moreover, Jacobian of discrepancy function $l_n(\boldsymbol{\theta})$ is not a random function:

$$\nabla l_n(\boldsymbol{\theta}) = -\begin{bmatrix} \nabla q_n(\boldsymbol{\theta}) \\ \nabla m_n(\boldsymbol{\theta}) \\ \nabla f_n(\boldsymbol{\theta}) \end{bmatrix} \tag{4.105}$$

and so is the $\mathcal{M}$ term. Denote $E(\boldsymbol{\theta}) = -\nabla l_n(\boldsymbol{\theta})$ and called **discrete E-matrix** from now on. The next several theorems in this section are immediate, followed by exactly the same arguments as QME theory.

Assumption 4.5.6 (No Multicollinearity)**.** $E = -\nabla l_n(\boldsymbol{\theta}_0)$ has independent columns at true parameters $\boldsymbol{\theta}_0$. Alternatively, E has no vector in the nullspace except 0.

It is interesting to see that, by piling up the discrepancy vectors to construct a joint discrepancy vector, the "no multicollinearity" assumption is easier to be satisfied. In fact, if at least one of $\nabla q_n(\boldsymbol{\theta}_0)$, $\nabla m_n(\boldsymbol{\theta}_0)$, $\nabla f_n(\boldsymbol{\theta}_0)$, has independent columns, then the assumption is automatically satisfied.

Theorem 4.5.7 (Consistency of QIPME)**.** *For any weighting matrix $\hat{W}$ with $\hat{W} \xrightarrow{p} W$, the QIPME $\hat{\boldsymbol{\theta}}_n = \arg\min\, J_n(\boldsymbol{\theta})$ is consistent estimator of the true parameters $\boldsymbol{\theta}_0$.*

Proof. Adapted from theorem 4.1.6. □

We have $\sqrt{n}\, l_n(\boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}(0, \Delta)$ where:

$$\Delta = \Delta(\boldsymbol{\theta}_0) = \begin{bmatrix} \Sigma & \Psi_a & \Psi_b \\ \Psi_a^{\mathrm{T}} & \Omega & \Psi_c \\ \Psi_b^{\mathrm{T}} & \Psi_c^{\mathrm{T}} & \Pi \end{bmatrix} \tag{4.106}$$

The entries of Σ is supplied by theorem 4.1.3 and 4.1.4. The entries of Ω is supplied by theorem 4.2.1 and 4.2.2. The entries of Π is supplied by theorem 4.4.1. The entries of Ψ_a is supplied by theorem 4.3.1. The entries of Ψ_b is supplied by theorem 4.5.4. The entries of Ψ_c is supplied by theorem 4.5.5.

Theorem 4.5.8 (Asymptotic Normality of QIPME)**.** *For any weighting matrix $\hat{W}$ with $\hat{W} \xrightarrow{p} W$, QIPME estimator $\hat{\boldsymbol{\theta}}_n = \arg\min\, J_n(\boldsymbol{\theta})$ is asymptotically normal:*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} N\left(0, (E^{\mathrm{T}}WE)^{-1}E^{\mathrm{T}}W\Delta WE(E^{\mathrm{T}}WE)^{-1}\right)$$

where $E = -\nabla l_n(\boldsymbol{\theta}_0)$ is discrete E-matrix at true parameters vector.

Proof. Adapted from theorem 4.1.7 □

Theorem 4.5.9 (Optimal Weight of QIPME)**.** *If weighting matrix $\hat{W} \xrightarrow{p} \Delta^{-1}$, then*

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}(0, (E^{\mathrm{T}}\Delta^{-1}E)^{-1})$$

and the covariance matrix is the lowest among all choices of weighting matrix. In other words, the estimator corresponding to this weighting matrix is the most efficient.

Proof. Adapted from theorem 4.1.8. □

Two Step Joint Estimator

To construct asymptotically efficient QIPME estimator, we need a weighting matrix $\hat{W}$ that converges in probability to Δ^{-1}. To achieve this consistency for weighting matrix, we construct $\hat{W}$ as an estimate of Δ^{-1}. Given a preliminary weighting matrix

$\tilde{W}$, we construct QIPME estimator corresponding to $\tilde{W}$:

$$J_n(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta})^{\mathrm{T}} \tilde{W} l_n(\boldsymbol{\theta}) \tag{4.107}$$

$$\tilde{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \ J_n(\boldsymbol{\theta}) \tag{4.108}$$

For any weighting matrix $\tilde{W}$, theorem 4.5.7 implies $\tilde{\boldsymbol{\theta}}$ is consistent. We compute the estimate for covariance matrix by equation 4.106: $\tilde{\Delta} = \Delta(\tilde{\boldsymbol{\theta}})$ and put $\hat{W} = \tilde{\Delta}^{-1}$. Then it is clear that $\hat{W} \xrightarrow{p} \Delta^{-1}$. Now we are ready to conduct the second step:

$$J_n(\boldsymbol{\theta}) = \min_{\boldsymbol{\theta}} l_n(\boldsymbol{\theta})^{\mathrm{T}} \hat{W} l_n(\boldsymbol{\theta}) \tag{4.109}$$

$$\hat{\boldsymbol{\theta}} = \arg\min_{\boldsymbol{\theta}} \ J_n(\boldsymbol{\theta}) \tag{4.110}$$

Then one can show $\hat{\boldsymbol{\theta}}$ constructed by two step estimation is consistent and asymptotically most efficient.

Theorem 4.5.10 (Efficient Two-step QIPME)**.** *The two step joint estimator $\hat{\boldsymbol{\theta}}$ constructed above is consistent estimator of $\boldsymbol{\theta}_0$ and asymptotically normal:*

$$\sqrt{n}(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}(0, (E^{\mathrm{T}} \Delta E)^{-1})$$

Proof. A direct application of theorem 4.5.8 with weighting matrix $\hat{W} \xrightarrow{p} \Delta^{-1}$. □

The result also show that the asymptotic behaviour of two-step estimator $\hat{\boldsymbol{\theta}}$ does not depend on the choice of preliminary weighting matrix $\hat{W}$. In finite samples, $\hat{\boldsymbol{\theta}}$, will be affected by the choice of $\hat{W}$. This undesirable dependency can be eliminated by interating the estimation sequence until convergence. This is called **iterated QIPME estimator**. Another strategy called **continuously-updated QIPME estimator** can be repeated in the same way as its counterpart in GMM.

After obtaining the two step estimator $\hat{\boldsymbol{\theta}}$, the estimate for asymptotic covariance

matrix of $\sqrt{n}\,\hat{\boldsymbol{\theta}}$ is $(\hat{E}^{\mathrm{T}}\hat{\Delta}^{-1}\hat{E})^{-1}$ converging in probability to $(E^{\mathrm{T}}\Delta^{-1}E)^{-1}$ where:

$$\hat{E} = -\nabla l_n(\hat{\boldsymbol{\theta}}) \xrightarrow{p} E = -\nabla l_n(\boldsymbol{\theta}_0) \tag{4.111}$$

$$\hat{\Delta} = \Delta(\hat{\boldsymbol{\theta}}) \xrightarrow{p} \Delta = \Delta(\boldsymbol{\theta}_0) \tag{4.112}$$

Hypothesis testing

The estimate of covariance matrix enables us to construct confidence intervals, confidence ellipsoid, hypothesis testing. Furthermore, since GMM and QIPME have the same mathematical pattern, Hansen-Sargan Test and Distance test are also applicable.

Chapter 5

Limiting Asymptotic Efficiency

Grouping data by reporting quantiles, interquantile means or sample proportions can be viewed as a *lossy compression* of original information. In the previous chapter, we introduced a GMM-based method as a general framework to estimate the underlying distribution using any single type of aggregated data, or even a combination of them. In this chapter, we will show that in addition to having many desirable properties such as consistency and normality, the proposed estimators are also efficient. The information loss during the compression process is not as bad as you expect, and a lot can be recovered by utilising all the given information with optimal weighting scheme and joint estimation.

5.1 The Fundamental Efficiency Lemma

In this section, we will state several important results necessary for subsequent efficiency analysis of matching estimators. The first result is due to Cramér (2016) and Rao (1992) which essentially say MLE is asymptotically the most efficient.

Theorem 5.1.1 (Cramer-Rao Lower Bound). *Let $\boldsymbol{X} = (X_1, X_2, \ldots, X_n) \overset{iid}{\sim} \mathring{f}(\cdot;\boldsymbol{\theta})$ with information matrix $\mathring{I}(\boldsymbol{\theta})$ and joint information matrix $I(\boldsymbol{\theta}) = n\mathring{I}(\boldsymbol{\theta})$. Suppose random variable $Z = Z(\boldsymbol{X})$ has $\mathrm{E}(Z) = g(\boldsymbol{\theta})$, then $\mathrm{Var}(Z)$ is bounded from below via:*

$$\mathrm{Var}(Z) \geq \nabla^{\mathrm{T}} g(\boldsymbol{\theta}) I^{-1}(\boldsymbol{\theta}) \nabla g(\boldsymbol{\theta})$$

Proof. See Kroese, Chan et al. (2014, p. 171). □

If $g(\boldsymbol{\theta}) = \theta_j$, $Z(\mathbf{X})$ is consistent to θ_j (E(Z) and its standard deviation is proportional to $cn^{-1/2}$ (c is some constant). We can then write $\mathrm{E}(Z) = \theta_j + b_n(\boldsymbol{\theta})$ where b_n is the bias and $b_n(\boldsymbol{\theta}) \to 0$ as n tends to infinity. Apply the Cramer-Rao lower bound we obtain:

$$\lim_{n\to\infty} \mathrm{Var}(\sqrt{n}Z) \geq \mathring{I}^{-1}(\boldsymbol{\theta})_{jj} = \lim_{n\to\infty} \mathrm{Var}(\sqrt{n}\,\hat{\theta}_j^{\mathrm{MLE}}) \tag{5.1}$$

which roughly says among all standard estimators of $\boldsymbol{\theta}_j$, the MLE has the lowest asymptotic variance. In other words, Cramer-Rao theorem establishes the lower bound for asymptotic efficiency of any standard estimators you could ever think of. By **standard estimator**, I mean a consistent estimator whose standard deviation is proportional to $c\,n^{-1/2}$. Very often, the estimator you have is also asymptotically normal (as shown in previous chapter).

Lemma 5.1.2 (Positive Semi-definiteness of Matrix Difference)**.** *If matrices A and B are positive definite and $A-B$ is positive (semi)definite, then $B^{-1}-A^{-1}$ is positive (semi)definite.*

Proof. The proof is deferred to Appendix A.5. □

We usually face the question of whether one estimator is more efficient than another in terms of their asymptotic covariance matrices. The problem reduces to showing the difference between two covariance matrix is at least positive semi-definite, which is not always simple to show in a direct way. Therefore, lemma 5.1.2 is extremely useful in providing an alternative indirect way of proving matrix inequality.

Lemma 5.1.3 (Block Matrix Inversion)**.** *If a matrix is partitioned into 4 blocks and matrices A, D, Schur complement $D - CA^{-1}B$ are non-singular, it can be inverted blockwise as follows:*

$$\begin{bmatrix} A & B \\ C & D \end{bmatrix}^{-1} = \begin{bmatrix} A^{-1} + A^{-1}B(D-CA^{-1}B)^{-1}CA^{-1} & -A^{-1}B(D-CA^{-1}B)^{-1} \\ -(D-CA^{-1}B)^{-1}CA^{-1} & (D-CA^{-1}B)^{-1} \end{bmatrix}$$

Proof. See Lu and Shiou (2002) for an insightful proof. □

Block matrix inversion is the key to prove the next statement, which I shall call "the fundamental efficiency lemma". It is fundamental in proving efficiency improvement from adding more asmoment conditions. Later on, I will show: as long as the discrepancy vector holds a valid asymptotic covariance structure, using additional asmoment conditions always leads to efficiency gain.

Lemma 5.1.4 (Fundamental Efficiency Lemma)**.** *Suppose Σ is a K by K positive definite matrix, nested as a block inside another positive definite matrix Σ_R of size $K+1$ by $K+1$:*

$$\Sigma_R = \begin{bmatrix} \Sigma & s \\ s^{\mathrm{T}} & \sigma^2_{K+1} \end{bmatrix} = \left[\begin{array}{cccc|c} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1K} & \sigma_{1,K+1} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2K} & \sigma_{2,K+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \sigma_{K1} & \sigma_{K2} & \cdots & \sigma^2_{KK} & \sigma_{K,K+1} \\ \hline \sigma_{K+1,1} & \sigma_{K+1,2} & \cdots & \sigma_{K+1,K} & \sigma^2_{K+1} \end{array}\right]$$

Suppose a is a column vector of size d, A is a K by d matrix and put:

$$A_R = \begin{bmatrix} A \\ a^{\mathrm{T}} \end{bmatrix}, \quad A_R^{\mathrm{T}} = \begin{bmatrix} A^{\mathrm{T}} & a \end{bmatrix}$$

under the assumption that A has independent columns (A_R has independent columns as a result). Then we have the following matrix inequality:

$$(A_R^{\mathrm{T}} \Sigma_R^{-1} A_R)^{-1} \leq (A^{\mathrm{T}} \Sigma^{-1} A)^{-1}$$

The inequality still holds true even if the block Σ is not nested on the upper left of Σ_R. In this case, matrix A is required to be nested inside A_R in accordance with Σ.

Proof. Since A, A_R have full column rank, both $A_R^{\mathrm{T}} \Sigma_R^{-1} A_R$ and $A^{\mathrm{T}} \Sigma^{-1} A$ are positive definite. By invoking lemma 5.1.2, the matrix inequality we aim to prove is equivalent to:

$$A_R^{\mathrm{T}}\Sigma_R^{-1}A_R \geq A^{\mathrm{T}}\Sigma^{-1}A \tag{5.2}$$

Putting $\sigma^2 = \sigma_{K+1}^2$ to shorten the notation. Using block matrix inversion (lemma 5.1.3), the inverse of Σ_R is:

$$\Sigma_R^{-1} = \begin{bmatrix} \Sigma^{-1} + \Sigma^{-1}s(\sigma^2 - s^{\mathrm{T}}\Sigma^{-1}s)^{-1}s^{\mathrm{T}}\Sigma^{-1} & -\Sigma^{-1}s(\sigma^2 - s^{\mathrm{T}}\Sigma^{-1}s)^{-1} \\ -(\sigma^2 - s^{\mathrm{T}}\Sigma^{-1}s)^{-1}s^{\mathrm{T}}\Sigma^{-1} & (\sigma^2 - s^{\mathrm{T}}\Sigma^{-1}s)^{-1} \end{bmatrix} \tag{5.3}$$

By assumption Σ_R is positive $K+1$ by $K+1$ positive definite matrix, therefore Σ_R^{-1} is also positive definite. We then deduce that the last diagonal entry of Σ_R^{-1}, namely $(\sigma^2 - s^{\mathrm{T}}\Sigma^{-1}s)^{-1}$ is positive.
With simple matrix block multiplication, one can show the matrix inequality 5.2 can be expanded and simplified as:

$$A^{\mathrm{T}}\Sigma^{-1}s\, s^{\mathrm{T}}\Sigma^{-1}A + aa^{\mathrm{T}} - as^{\mathrm{T}}\Sigma^{-1}A - A^{\mathrm{T}}\Sigma^{-1}sa^{\mathrm{T}} \geq 0 \tag{5.4}$$

The conclusion is clear by recognising the LHS can be factored as:

$$(s^{\mathrm{T}}\Sigma^{-1}A - a^{\mathrm{T}})^{\mathrm{T}}(s^{\mathrm{T}}\Sigma^{-1}A - a^{\mathrm{T}}) \tag{5.5}$$

To prove the inequality for general nesting situation, let P_k be the $K+1$ by $K+1$ permutation matrix that switches the last row and the k^{th} row by pre-multiplication. Using the fact that $P_k^{\mathrm{T}} = P_k^{-1}$, we have:

$$(P_kA_R)^{\mathrm{T}}(P_k\Sigma_RP_k^{\mathrm{T}})^{-1}(P_kA_R) = A_R^{\mathrm{T}}\Sigma_RA_R \tag{5.6}$$

where $P_k\Sigma_RP_k^{\mathrm{T}}$ represents the general nesting of Σ inside Σ_R, and P_kA_R represents the general nesting of A inside A_R, yielding the desired conclusion. □

The *Fundamental efficiency lemma* concludes this brief section. Frankly, the theorem is purely a linear algebra proposition without any mention of "efficiency". By exploring the question of efficiency gain, this mathematical pattern is extracted. If you ever heard that using more valid instrument variables results in more asymptotic efficiency, then I encourage you to pause and see the connection between this result and the fundamental lemma.

5.2 Estimation Efficiency using Grouped Data

According to Cramer-Rao lower bound, MLE is the most asymptotically efficient among all standard estimators. This is the major reason why MLE is considered as a benchmark method in parametric estimation and inference. In this section, we will analyse various efficiency problems that either compare asymptotic efficiency between MLE and matching estimator, or between matching estimator and itself with extra addition of asmoment conditions. The results proved in this section have an important role to play in establishing quadratic matching estimation is a general efficient alternative to the benchmark method MLE, especially when handling grouped data.

5.2.1 Efficiency Results of Multinomial MLE and PME

We use the statistical model for generating class frequency data in chapter 3 section 3.1 with $K+1$ income classes for notation simplicity. The corresponding stochastic log-likelihood function is provided in equation 3.7:

$$\ln f(\mathbf{N};\boldsymbol{\theta}) = \ln\left(\frac{n!}{N_1!N_2!...N_{K+1}!}\right) + N_1 \ln p_1(\boldsymbol{\theta}) + ... + N_k \ln p_{K+1}(\boldsymbol{\theta}) \tag{5.7}$$

where $p_j(\boldsymbol{\theta}) = p(t_{j-1}, t_j; \boldsymbol{\theta}) = F(t_j; \boldsymbol{\theta}) - F(t_{j-1}; \boldsymbol{\theta})$ for $j = 1, 2, \ldots K+1$. Take the gradient of both side with respect to $\boldsymbol{\theta}$, we obtain the Efficient Score S_C for class frequency data:

$$S_C(\boldsymbol{\theta}; \mathbf{N}) = \sum_{j=1}^{K+1} N_j \frac{\nabla p_j(\boldsymbol{\theta})}{p_j(\boldsymbol{\theta})} \tag{5.8}$$

It is not needed but interesting to see that $\mathrm{E}_{\boldsymbol{\theta}} S_C(\boldsymbol{\theta}; \mathbf{N}) = \sum \nabla p_j(\boldsymbol{\theta}) = \nabla 1 = 0$ as expected. The information matrix associated with class frequency model is:

$$I_C(\boldsymbol{\theta}) = \mathrm{E}\, S_C(\boldsymbol{\theta}; N) S_C(\boldsymbol{\theta}; N)^{\mathrm{T}} \tag{5.9}$$

$$= \sum_{j=1}^{K+1} \frac{\nabla p_j(\boldsymbol{\theta}) \nabla^{\mathrm{T}} p_j(\boldsymbol{\theta})}{p_j(\boldsymbol{\theta})^2} \mathrm{E}(N_j^2) + \sum_{j \neq k} \frac{\nabla p_j(\boldsymbol{\theta}) \nabla^{\mathrm{T}} p_k(\boldsymbol{\theta})}{p_j(\boldsymbol{\theta}) p_k(\boldsymbol{\theta})} \mathrm{E}(N_j N_k) \tag{5.10}$$

As $\mathrm{E}(N_j^2) = np_j(1 - p_j + np_j)$ and $\mathrm{E}(N_j N_k) = n(n-1)p_j p_k$ for $j \neq k$, the information matrix can be simplified to:

$$I_C(\boldsymbol{\theta}) = n \sum_{j=1}^{K+1} \frac{\nabla p_j(\boldsymbol{\theta}) \nabla^{\mathrm{T}} p_j(\boldsymbol{\theta})}{p_j(\boldsymbol{\theta})} \tag{5.11}$$

Maximum likelihood theory provides $\sqrt{n}(\hat{\boldsymbol{\theta}}^{\mathrm{MLE}} - \boldsymbol{\theta}_0) \xrightarrow{d} \mathrm{N}(0, (I_C(\boldsymbol{\theta}_0)/n)^{-1})$. The natural question to ask is: if we refine the class intervals, is there a gain in asymptotic efficiency. The answer is *Yes* and shown in the following theorem.

Theorem 5.2.1 (Efficiency Improvement Multinomial MLE)**.** *Let $T = \{t_0, t_1, \ldots t_K\}$ is the interval bounds for class frequency model with information matrix $I_C(\boldsymbol{\theta})$. A refinement of T is finite set of point T' such as $T \subseteq T'$, and denote $J_C(\boldsymbol{\theta})$ is the information matrix associated with interval bounds T', then:*

$$I_C^{-1}(\boldsymbol{\theta}) \geq J_C^{-1}(\boldsymbol{\theta})$$

In other words, knowledge on additional class intervals improves asymptotic efficiency of the estimator. Moreover, if the refinement T is dense (in the support) when K tends to infinity, then:

$$\lim_{K \to \infty} I_C(\boldsymbol{\theta})/n = \mathring{I}(\boldsymbol{\theta})$$

Proof. It is enough to show the efficiency improvement when T' contains just one point more than T. Suppose the refinement happens at $t_{j-1} < t_j$ in T, which becomes $t_{j-1} < t < t_j$ in T', the difference between two information matrices is:

$$J_C(\boldsymbol{\theta}) - I_C(\boldsymbol{\theta}) = \frac{\nabla p_a(\boldsymbol{\theta}) \nabla^{\mathrm{T}} p_a(\boldsymbol{\theta})}{p_a(\boldsymbol{\theta})} + \frac{\nabla p_b(\boldsymbol{\theta}) \nabla^{\mathrm{T}} p_b(\boldsymbol{\theta})}{p_b(\boldsymbol{\theta})} - \frac{\nabla p_j(\boldsymbol{\theta}) \nabla^{\mathrm{T}} p_j(\boldsymbol{\theta})}{p_j(\boldsymbol{\theta})} \tag{5.12}$$

where $p_a(\boldsymbol{\theta}) = p(t_{j-1}, t; \boldsymbol{\theta})$ and $p_b(\boldsymbol{\theta}) = p(t, t_{j+1}; \boldsymbol{\theta})$.

Put $x_1 = \nabla p_a(\boldsymbol{\theta})$, $x_2 = \nabla p_b(\boldsymbol{\theta})$, $p_1 = p_a(\boldsymbol{\theta})$, $p_2 = p_b(\boldsymbol{\theta})$ and $x = \nabla p_j(\boldsymbol{\theta})$, $p = p_j(\boldsymbol{\theta})$. Then $x = x_1 + x_2$ and $p = p_1 + p_2$. Simple manipulation shows:

$$
\begin{aligned}
&\frac{x_1 x_1^{\mathrm{T}}}{p_1} + \frac{x_2 x_2^{\mathrm{T}}}{p_2} - \frac{x x^{\mathrm{T}}}{p} \\
=& \left(\sqrt{\frac{p_2}{p_1(p_1+p_2)}}x_1 - \sqrt{\frac{p_1}{p_2(p_1+p_2)}}x_2\right)^{\mathrm{T}} \left(\sqrt{\frac{p_2}{p_1(p_1+p_2)}}x_1 - \sqrt{\frac{p_1}{p_2(p_1+p_2)}}x_2\right)
\end{aligned}
\tag{5.13}
$$

which is at least positive semidefinite. Together with lemma 5.1.2, the equation above shows the improvement in asymptotic efficiency of MLE by using a more refined collection of class intervals.

Without loss of generality, assume the distribution support is $(0, \infty)$. Recall the integral form of the information matrix associated with distribution function $F(\cdot; \boldsymbol{\theta})$ and density $f(\cdot; \boldsymbol{\theta})$:

$$
\mathring{I}(\boldsymbol{\theta}) = \int_0^{\infty} \frac{\nabla_{\boldsymbol{\theta}} f(y; \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}}^{\mathrm{T}} f(y; \boldsymbol{\theta})}{f(y; \boldsymbol{\theta})}\, dy \tag{5.14}
$$

As $K \to \infty$ and $T_K = (t_0, t_1, \ldots, t_K)$ is dense in $(0, \infty)$ in limit, the information matrix corresponding to T_K (equation 5.11) is:

$$
\frac{1}{n} I_C(\boldsymbol{\theta}) = \sum_{j=1}^{K} \frac{\nabla p_j(\boldsymbol{\theta}) \nabla^{\mathrm{T}} p_j(\boldsymbol{\theta})}{p_j(\boldsymbol{\theta})} \approx \sum_{j=1}^{K} \frac{\nabla_{\boldsymbol{\theta}} f(y_j; \boldsymbol{\theta}) \nabla_{\boldsymbol{\theta}}^{\mathrm{T}} f(y_j; \boldsymbol{\theta})}{f(y_j; \boldsymbol{\theta})} (t_j - t_{j-1}) \tag{5.15}
$$

where $p_j(\boldsymbol{\theta}) \approx f(y_j; \boldsymbol{\theta})(t_j - t_{j-1})$ and $\nabla p_j(\boldsymbol{\theta}) \approx \nabla f(y_j; \boldsymbol{\theta})(t_j - t_{j-1})$ for some $y_j \in [t_{j-1}, t_j]$ provided $t_j - t_{j-1}$ is small for all j. By assumption, $\lim T$ is dense in the support, this expression therefore converges to the integral specified in 5.14, coinciding with the information matrix associated with class frequency model. □

The next theorem shows the continuously-updated PME coincides with Pearson minimum χ^2 estimator introduced in Hinkley and Cox (1979). PME estimator thus provides new insight by demonstrating a well-developed estimator is an instance of the general quadratic matching framework.

Theorem 5.2.2 (Continuous-updated PME is Pearson χ^2 Estimator)**.** *Suppose there are $K+1$ income classes. Continuously-updated PME $\hat{\boldsymbol{\theta}}_n$ coincides with Pearson minimum χ^2 estimator $\tilde{\boldsymbol{\theta}}_n$ which minimises the expression:*

$$T(\boldsymbol{\theta}) = \sum_{j=1}^{K+1} \frac{(\hat{p}_j - p_j(\boldsymbol{\theta}))^2}{p_j(\boldsymbol{\theta})}$$

Proof. Recall the continuously-updated PME $\hat{\boldsymbol{\theta}}_n$ minimise the objective function:

$$J(\boldsymbol{\theta}) = f_n(\boldsymbol{\theta})^{\mathrm{T}} \Pi^{-1}(\boldsymbol{\theta}) f_n(\boldsymbol{\theta}) \tag{5.16}$$

where f_n and Π are specified in expression 4.77 and theorem 4.4.1. Beware that the last sample proportion is excluded from f_n and Π, and the inverse covariance matrix is:

$$\Pi^{-1} = \begin{bmatrix} \left(\frac{1}{p_1} + \frac{1}{p_{K+1}}\right) & \frac{1}{p_{K+1}} & \cdots & \frac{1}{p_{K+1}} \\ \frac{1}{p_{K+1}} & \left(\frac{1}{p_2} + \frac{1}{p_{K+1}}\right) & \cdots & \frac{1}{p_{K+1}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{p_{K+1}} & \frac{1}{p_{K+1}} & \cdots & \left(\frac{1}{p_K} + \frac{1}{p_{K+1}}\right) \end{bmatrix} \tag{5.17}$$

The continuously-updated PME's objective function then becomes:

$$J(\boldsymbol{\theta}) = \sum_{j=1}^{K} \left(\frac{1}{p_j(\boldsymbol{\theta})} + \frac{1}{p_{K+1}(\boldsymbol{\theta})}\right) (\hat{p}_j - p_j(\boldsymbol{\theta}))^2 + \sum_{j \neq k \leq K} \frac{(\hat{p}_j - p_j(\boldsymbol{\theta}))(\hat{p}_k - p_k(\boldsymbol{\theta}))}{p_{K+1}(\boldsymbol{\theta})} \tag{5.18}$$

Simple manipulation confirms:

$$J(\boldsymbol{\theta}) = \sum_{j=1}^{K} \frac{1}{p_j(\boldsymbol{\theta})} (\hat{p}_j - p_j(\boldsymbol{\theta}))^2 + \frac{1}{p_{K+1}(\boldsymbol{\theta})} \left(\sum_{j=1}^{K} (\hat{p}_j - p_j(\boldsymbol{\theta}))\right)^2 \tag{5.19}$$

coincides with $T(\boldsymbol{\theta})$ as desired by recognising $\sum_{j=1}^{K}(\hat{p}_j - p_j(\boldsymbol{\theta})) = p_{K+1}(\boldsymbol{\theta}) - \hat{p}_{K+1}$. □

Hinkley and Cox (1979) establishes that: for any fixed number $K+1$ of class intervals, the Pearson minimum-χ^2 is asymptotically equivalent to the multinomial MLE, and less efficient than the MLE using micro-observations. Since continuously-updated PME coincides with Pearson minimum-χ^2 estimator, it implies the statement holds true for continuously-updated PME as well. In a similar fashion, the next theorem shows the equivalence of asymptotic efficiency between multinomial MLE and two-step PME.

Theorem 5.2.3 (Equivalence of Two-Step PME and Multinomial MLE)**.** *For a fixed finite number of income classes, say $K+1$, the asymptotic efficiency of two-step PME (or any PME with weighting matrix $\hat{W} \xrightarrow{p} \Pi^{-1}$) and multinomial MLE is the same.*

Proof. Recall the asymptotic covariance matrix associated with two-step PME (see theorem 4.4.6 or theorem 4.4.5) is $(D^{\mathrm{T}}\Pi^{-1}D)^{-1}$ where:

$$D = D(\boldsymbol{\theta}) = \begin{bmatrix} \nabla^{\mathrm{T}} p_1(\boldsymbol{\theta}) \\ \nabla^{\mathrm{T}} p_2(\boldsymbol{\theta}) \\ \vdots \\ \nabla^{\mathrm{T}} p_K(\boldsymbol{\theta}) \end{bmatrix} \tag{5.20}$$

The covariance matrix inverse is:

$$\begin{aligned} D^{\mathrm{T}}\Pi^{-1}D = &\sum_{j=1}^{K} \left(\frac{1}{p_j(\boldsymbol{\theta})} + \frac{1}{p_{K+1}(\boldsymbol{\theta})} \right) \nabla p_j(\boldsymbol{\theta}) \nabla^{\mathrm{T}} p_j(\boldsymbol{\theta}) \\ &+ \sum_{(j \neq k) \leq K} \frac{1}{p_{K+1}(\boldsymbol{\theta})} \nabla p_j(\boldsymbol{\theta}) \nabla^{\mathrm{T}} p_k(\boldsymbol{\theta}) \end{aligned} \tag{5.21}$$

With the same manipulation in the previous theorem 5.2.2, one can show:

$$D^{\mathrm{T}}\Pi^{-1}D = \sum_{j=1}^{K+1} \frac{1}{p_j(\boldsymbol{\theta})} \nabla p_j(\boldsymbol{\theta}) \nabla^{\mathrm{T}} p_j(\boldsymbol{\theta}) \tag{5.22}$$

which coincides with $I_C(\boldsymbol{\theta})/n$, specified in equation 5.11, as desired. □

Theorem 5.2.4 (Efficiency Improvement PME)**.** *Let $T = \{t_0, t_1, \ldots t_K\}$ is the interval bounds for class frequency model. The PME asymptotic covariance matrix associated with T is $(D^{\mathrm{T}}\Pi^{-1}D)^{-1}$. A refinement of T is finite set of point T' such as $T \subseteq T'$, and let $(D_R^{\mathrm{T}}\Pi_R^{-1}D_R)^{-1}$ is the asymptotic covariance matrix associated with* *interval bounds T', then:*

$$(D^{\mathrm{T}}\Pi^{-1}D)^{-1} \geq (D_R^{\mathrm{T}}\Pi_R^{-1}D_R)^{-1}$$

In other words, knowledge on additional class intervals improves asymptotic efficiency of the PME. Moreover, if the refinement T is dense (in the support) when K tends to infinity, then:

$$\lim_{K\to\infty}(D^{\mathrm{T}}\Pi^{-1}D)^{-1} = \mathring{I}^{-1}(\boldsymbol{\theta})$$

Proof. Theorem 5.2.3 shows the asymptotic equivalence between two-step PME and Multinomial MLE. The conclusions then follow directly from theorem 5.2.1. Alternatively, the fundamental efficiency lemma 5.1.4 can be invoked to prove the first statement. □

5.2.2 Efficiency Results of QME and IME

The next three theorems are direct application of the fundamental efficiency lemma, showing essentially that using a more refined collection of groups improves asymptotic efficiency. The result can be extended to any joint estimator as long as the sample statistics are consistent jointly asymptotically normal, and the covariance structure behind those sample statistics remains valid when adding more asmoment condition.

Theorem 5.2.5 (Efficiency Improvement QME). *Let $(A^{\mathrm{T}}\Sigma^{-1}A)^{-1}$ be the covariance matrix associated with the most asymptotically efficient QME estimator using sample quantiles at $P = \{\pi_1, \pi_2, \ldots, \pi_K\}$. A refinement P' of P is a set of percentage points such that $P \subseteq P'$. Let $(A_R^{\mathrm{T}}\Sigma_R^{-1}A_R)^{-1}$ be the covariance matrix associated with the most asymptotically efficient QME estimator using sample quantiles at refinement P', then:*

$$(A_R^{\mathrm{T}}\Sigma_R^{-1}A_R)^{-1} \leq (A^{\mathrm{T}}\Sigma^{-1}A)^{-1}$$

Proof. Followed from lemma 5.1.4 □

Theorem 5.2.6 (Efficiency Improvement IME). *Let $(B^{\mathrm{T}}\Omega^{-1}B)^{-1}$ be the covariance matrix associated with the most asymptotically efficient IME estimator using sample* *non-overlapping interquantile means at* $P = \{\rho_0, \rho_1, \rho_2 \ldots, \rho_K\}$*. A refinement* P' *of* P *is a set of percentage points such that* $P \subseteq P'$*. Let* $(B_R^{\mathrm{T}} \Omega_R^{-1} B_R)^{-1}$ *be the covariance matrix associated with the most asymptotically efficient IME estimator using sample non-overlapping interquantile means at refinement* P'*, then:*

$$(B_R^{\mathrm{T}} \Omega_R^{-1} B_R)^{-1} \leq (B^{\mathrm{T}} \Omega^{-1} B)^{-1}$$

Proof. Followed from lemma 5.1.4 □

Theorem 5.2.7 (Efficiency Improvement QIME)**.** *Let* $(C^{\mathrm{T}} \Delta^{-1} C)^{-1}$ *be the covariance matrix associated with the most asymptotically efficient QIME estimator using sample quantiles at* $P_1 = \{\pi_1, \pi_2, \ldots, \pi_K\}$*, and non-overlapping interquantile means at* $P_2 = \{\rho_0, \rho_1, \rho_2 \ldots, \rho_K\}$*. Let* $(C_R^{\mathrm{T}} \Delta_R^{-1} C_R)^{-1}$ *be the covariance matrix associated with the most asymptotically efficient QIME estimator using sample quantiles at refinement* P_1'*, and sample non-overlapping interquantile means at refinement* P_2'*, then:*

$$(C_R^{\mathrm{T}} \Delta_R^{-1} C_R)^{-1} \leq (C^{\mathrm{T}} \Delta^{-1} C)^{-1}$$

Proof. Followed from lemma 5.1.4 □

I will conclude this section by two efficiency results and a conjecture. The two theorems surprisingly demonstrate that the QME and IQME can achieve their Cramer-Rao lower bound in limit as the number of groups tend to infinity. Before delving into the last two efficiency results, some useful conventions in notation are required. First, sometimes I will suppress the dependence of functions F, f, Q on $\boldsymbol{\theta}$ or $\boldsymbol{\theta}_0$, but readers should constantly keep in mind this dependence. Suppression of parameter vector $\boldsymbol{\theta}$ also serves another purpose. For example, $\nabla_{\boldsymbol{\theta}} Q(F(y;\boldsymbol{\theta});\boldsymbol{\theta})$ and $\nabla_{\boldsymbol{\theta}} Q(F(y);\boldsymbol{\theta})$ denote different expression. The former expression is the gradient of $Q(F(y;\boldsymbol{\theta});\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, while the latter expression is the gradient of $\nabla_{\boldsymbol{\theta}} Q(\pi;\boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$, and evaluated at $\pi = F(y;\boldsymbol{\theta})$.

I also want to offer an alternative quantile view of Fisher's Information matrix. Suppose $Y_1, Y_2, ..., Y_n \overset{iid}{\sim} F(\cdot;\boldsymbol{\theta}_0)$ with density function $f(\cdot;\boldsymbol{\theta})$ and quantile function

$Q(\cdot;\boldsymbol{\theta})$. Maximum likelihood estimator $\hat{\boldsymbol{\theta}}_{\text{MLE}}$ is asymptotically normal $\sqrt{n}(\hat{\boldsymbol{\theta}}_{\text{MLE}} - \boldsymbol{\theta}_0) \xrightarrow{d} \text{N}(0, \mathring{I}^{-1}(\boldsymbol{\theta}_0))$ where $\mathring{I}(\boldsymbol{\theta})$ is information matrix, defined as:

$$\mathring{I}(\boldsymbol{\theta}) = \int_{-\infty}^{\infty} \nabla_{\boldsymbol{\theta}} \ln f(y;\boldsymbol{\theta})\, \nabla_{\boldsymbol{\theta}}^{\text{T}} \ln f(y;\boldsymbol{\theta})\, f(y;\boldsymbol{\theta})\, dy \tag{5.23}$$

By making a change of variable $y = Q(p;\boldsymbol{\theta})$, $p = F(y;\boldsymbol{\theta})$, the information matrix can be alternatively expressed as:

$$\mathring{I}(\boldsymbol{\theta}) = \int_0^1 \nabla_{\boldsymbol{\theta}} \ln f(Q(p);\boldsymbol{\theta}).\nabla_{\boldsymbol{\theta}}^{\text{T}} \ln f(Q(p);\boldsymbol{\theta})\, dp \tag{5.24}$$

$$= \int_0^1 \frac{\nabla_{\boldsymbol{\theta}} f(Q(p);\boldsymbol{\theta})}{f(Q(p);\boldsymbol{\theta})}.\frac{\nabla_{\boldsymbol{\theta}}^{\text{T}} f(Q(p);\boldsymbol{\theta})}{f(Q(p);\boldsymbol{\theta})}\, dp \tag{5.25}$$

Let's make a detour by looking at information matrix from quantile perspective. Firstly, dimension analysis of equation 5.25 reveals that $\mathring{I}(\boldsymbol{\theta})$ is dimensionless in terms of Y, and contains an inverse squared measurement unit of $\boldsymbol{\theta}$ as expected. Secondly, if you consider $R = \nabla_{\boldsymbol{\theta}}^{\text{T}} \ln f(Q(p);\boldsymbol{\theta})_{p\in(0,1)}$ as a generalised matrix of size $(0,1) \times d$, then there are d columns in R. The i^{th} column captures the effect of a slight change in $\boldsymbol{\theta}_j$ in the distributional shape (measured by log-quantile function). Matrix R is *ill-behaved* whenever its columns are nearly dependent, and it intuitively means the distributional shape's change by one parameter can be closely imitated by adjustment in the other parameters. The situation resembles multicollinearity problem encountered in econometrics. Thirdly, near linear dependency in R, which causes the problem of difficult identification, can be revealed from the information matrix $\mathring{I}(\boldsymbol{\theta}) = R^{\text{T}}R$. If R is ill-behaved then $\mathring{I}(\boldsymbol{\theta})$ is near-singular and the Cramer-Rao lower bound for any standard estimators' covariance matrices $\mathring{I}^{-1}(\boldsymbol{\theta})$ blows up, making your estimation extremely imprecise.

From the quantile outlook of information matrix, I propose a Φ-measure[1] of "multicollinearity" of a continuous distribution with information matrix $\mathring{I}(\boldsymbol{\theta})$ as:

$$\Phi(\boldsymbol{\theta}) = \lambda_{\min}\big(\mathring{I}(\boldsymbol{\theta})\big) = \rho\big(\mathring{I}^{-1}(\boldsymbol{\theta})\big) \tag{5.26}$$

[1]in honour of my thesis supervisor Dr Fedor Iskhakov, who also pointed out this problem.

where $\lambda_{\min}$ denotes the minimum eigenvalue, and ρ denotes the spectral radius (also known as maximum eigenvalue). Very small Φ is interpreted as excessive "multi-collinearity", causing serious problem in estimation.

Theorem 5.2.8 (QME and Cramer-Rao Lower Bound)**.** *Let $Y_1, Y_2, ..., Y_n \overset{iid}{\sim} F(\cdot;\boldsymbol{\theta}_0)$ with density function $f(\cdot;\boldsymbol{\theta})$ and quantile function $Q(\cdot;\boldsymbol{\theta})$. For any K with $K \geq d$, a set of percentage points $P_K = \{\pi_1, \pi_2, \ldots, \pi_K\} \subset (0,1)$, such that $\pi_1 < \pi_2 < \ldots < \pi_K$, is chosen to compute sample quantiles. Let $\overset{\circ}{I}(\boldsymbol{\theta})$ be the information matrix associated with distribution function $F(\cdot;\boldsymbol{\theta})$, and $(A_K^{\mathrm{T}}\Sigma_K^{-1}A_K)^{-1}$ be the asymptotic covariance matrix associated with the most asymptotically efficient QME estimator using sample quantiles at percentage points set P_K. Then:*

$$(A_K^{\mathrm{T}}\Sigma_K^{-1}A_K)^{-1} \geq \overset{\circ}{I}{}^{-1}(\boldsymbol{\theta}_0),$$

and as long as $\lim_{K\to\infty} \max\left\{|\pi_{j+1} - \pi_j| : j = 0, 1, 2, \ldots K\right\} = 0, (\pi_0 = 0, \pi_{K+1} = 1)$ holds, the Cramer-Rao lower bound is achievable in limit:

$$\lim_{K\to\infty} (A_K^{\mathrm{T}}\Sigma_K^{-1}A_K)^{-1} = \overset{\circ}{I}{}^{-1}(\boldsymbol{\theta}_0)$$

Proof. The asymptotic covariance matrix of sample quantiles at $P_K = \{\pi_1, \pi_2, \ldots, \pi_K\}$

$$\Sigma_K = \begin{bmatrix} \frac{\pi_1(1-\pi_1)}{f(Q(\pi_1))^2} & \frac{\pi_1(1-\pi_2)}{f(Q(\pi_1))f(Q(\pi_2))} & \cdots & \frac{\pi_1(1-\pi_K)}{f(Q(\pi_1))f(Q(\pi_K))} \\ \frac{\pi_1(1-\pi_2)}{f(Q(\pi_2))f(Q(\pi_1))} & \frac{\pi_2(1-\pi_2)}{f(Q(\pi_2))^2} & \cdots & \frac{\pi_2(1-\pi_K)}{f(Q(\pi_2))f(Q(\pi_K))} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\pi_1(1-\pi_K)}{f(Q(\pi_K))f(Q(\pi_1))} & \frac{\pi_2(1-\pi_K)}{f(Q(\pi_K))f(Q(\pi_2))} & \cdots & \frac{\pi_K(1-\pi_K)}{f(Q(\pi_K))^2} \end{bmatrix} \tag{5.27}$$

has inverse:

$$\Sigma_K^{-1} = \begin{bmatrix} \frac{f(Q(\pi_1))^2\pi_2}{(\pi_1-0)(\pi_2-\pi_1)} & -\frac{f(Q(\pi_1))f(Q(\pi_2))}{\pi_2-\pi_1} & 0 & \cdots & 0 \\ -\frac{f(Q(\pi_1))f(Q(\pi_2))}{\pi_2-\pi_1} & \frac{f(Q(\pi_2))^2(\pi_3-\pi_1)}{(\pi_2-\pi_1)(\pi_3-\pi_2)} & -\frac{f(Q(\pi_2))f(Q(\pi_3))}{\pi_3-\pi_2} & \cdots & 0 \\ 0 & -\frac{f(Q(\pi_2))f(Q(\pi_3))}{\pi_3-\pi_2} & \frac{f(Q(\pi_3))^2(\pi_4-\pi_2)}{(\pi_3-\pi_2)(\pi_4-\pi_3)} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{f(Q(\pi_K))^2(1-\pi_{K-1})}{(\pi_K-\pi_{K-1})(1-\pi_K)} \end{bmatrix} \tag{5.28}$$

So the inverse of covariance is a banded matrix, whose bandwidth is 1.

Note that the covariance matrix and its inverse are dependent on true parameters $\boldsymbol{\theta}_0$. The matrix Σ_K^{-1} has $\mathrm{LDL}^{\mathrm{T}}$- decomposition (with sign reversal on L) where:

$$L_K^{\mathrm{T}} = \begin{bmatrix} -1 & \frac{f(Q(\pi_2))/\pi_2}{f(Q(\pi_1))/\pi_1} & 0 & \cdots & 0 & 0 \\ 0 & -1 & \frac{f(Q(\pi_3))/\pi_3}{f(Q(\pi_2))/\pi_2} & \cdots & 0 & 0 \\ 0 & 0 & -1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -1 & \frac{f(Q(\pi_K))/\pi_K}{f(Q(\pi_{K-1}))/\pi_{K-1}} \\ 0 & 0 & 0 & \cdots & 0 & -1 \end{bmatrix} \tag{5.29}$$

$$D_K = \begin{bmatrix} \frac{f(Q(\pi_1))^2\pi_2}{\pi_1(\pi_2-\pi_1)} & 0 & 0 & \cdots & 0 \\ 0 & \frac{f(Q(\pi_2))^2\pi_3}{\pi_2(\pi_3-\pi_2)} & 0 & \cdots & 0 \\ 0 & 0 & \frac{f(Q(\pi_3))^2\pi_4}{\pi_3(\pi_4-\pi_3)} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \frac{f(Q(\pi_K))^2}{\pi_K(1-\pi_K)} \end{bmatrix} \tag{5.30}$$

The diagonal entries of D_K are called the pivots of Σ_K^{-1}. One passing comment is all the pivots of the inverse covariance matrix are positive, consistent with the positive definiteness of Σ_K and Σ_K^{-1}. We want to investigate $A_K^{\mathrm{T}} L_K D_K L_K^{\mathrm{T}} A_K$ of size $d \times d$, where:

$$A_K = \begin{bmatrix} \nabla_{\boldsymbol{\theta}}^{\mathrm{T}}(\pi_1; \boldsymbol{\theta}_0) \\ \nabla_{\boldsymbol{\theta}}^{\mathrm{T}}(\pi_2; \boldsymbol{\theta}_0) \\ \vdots \\ \nabla_{\boldsymbol{\theta}}^{\mathrm{T}}(\pi_K; \boldsymbol{\theta}_0) \end{bmatrix} = \begin{bmatrix} Q_{\theta_1}(\pi_1; \boldsymbol{\theta}) & \cdots & Q_{\theta_d}(\pi_1; \boldsymbol{\theta}) \\ Q_{\theta_1}(\pi_2; \boldsymbol{\theta}) & \cdots & Q_{\theta_d}(\pi_2; \boldsymbol{\theta}) \\ \cdots & \ddots & \cdots \\ Q_{\theta_1}(\pi_K; \boldsymbol{\theta}) & \cdots & Q_{\theta_d}(\pi_K; \boldsymbol{\theta}) \end{bmatrix} \tag{5.31}$$

of size K by d.

Moreover, D_K can be factored out as product of three diagonal matrices $D_K = E_K F_K E_K$ whose diagonal entries of F_K are $\pi_j \pi_{j+1}(\pi_{j+1} - \pi_j)$ and diagonal entries of E_K are $\frac{f(Q(\pi_j))}{\pi_j(\pi_{j+1}-\pi_j)}$, with index j running from 1 to K. When $K \to \infty$ provided that the set limit $\lim P_K$ is dense in $(0, 1)$, it is straightforward to see that matrix

$\lim E_K L_K^{\mathrm{T}}$ represents a linear differential operator T that send $g(p)$, $p \in (0,1)$ to:

$$\frac{d}{dp}\left[\frac{g(p)f(Q(p))}{p}\right], \quad p \in (0,1) \tag{5.32}$$

This is indeed a linear operator because $T(g_1+g_2) = Tg_1 + Tg_2$ and $T(cg) = c(Tg)$ where c is a scalar, and g_1, g_2, g are differentiable functions in $(0,1)$.
Denote $\tilde{A}^{\mathrm{T}}\tilde{\Sigma}^{-1}\tilde{A}$ as $\lim_{K\to\infty} A_K^{\mathrm{T}}\Sigma_K^{-1}A_K$. The tilda indicates we take a matrix, say A, to its extreme continuous version $\tilde{A}$ when $K=\infty$. We temporarily use $\boldsymbol{\theta}$ instead of $\boldsymbol{\theta}_0$ for simplicity, then it is apparent that:

$$\begin{aligned}\tilde{A}^{\mathrm{T}}\tilde{\Sigma}^{-1}\tilde{A} &= (\tilde{A}^{\mathrm{T}}\tilde{L}\tilde{E})\tilde{F}(\tilde{E}\tilde{L}^{\mathrm{T}}\tilde{A}) &(5.33)\\ &= \int_0^1 \frac{d}{dp}\left[\frac{\nabla_{\boldsymbol{\theta}}Q(p;\boldsymbol{\theta})f(Q(p))}{p}\right]\frac{d}{dp}\left[\frac{\nabla_{\boldsymbol{\theta}}Q(p;\boldsymbol{\theta})f(Q(p))}{p}\right]^{\mathrm{T}} p^2\,dp &(5.34)\end{aligned}$$

We aim to show expression 5.25 coincides with expression 5.34 for information matrix. Firstly, we are able to expand the integrand of the former expression by using:

$$\frac{d}{dp}\left[\frac{\nabla_{\boldsymbol{\theta}}Q(p;\boldsymbol{\theta})f(Q(p))}{p}\right] = -\left(\frac{1}{p}\frac{\nabla_{\boldsymbol{\theta}}f(Q(p);\boldsymbol{\theta})}{f(Q(p);\boldsymbol{\theta})} + \frac{1}{p^2}\nabla_{\boldsymbol{\theta}}Q(p;\boldsymbol{\theta})\,f(Q(p))\right) \tag{5.35}$$

By using expression 5.25, it remains to be shown that:

$$\begin{aligned}&\int_0^1 \frac{1}{p}\nabla_{\boldsymbol{\theta}}Q(p;\boldsymbol{\theta}).\nabla_{\boldsymbol{\theta}}^{\mathrm{T}}f(Q(p);\boldsymbol{\theta})\,dp + \int_0^1 \frac{1}{p}\nabla_{\boldsymbol{\theta}}f(Q(p);\boldsymbol{\theta}).\nabla_{\boldsymbol{\theta}}^{\mathrm{T}}Q(p;\boldsymbol{\theta})\,dp\\ &\quad + \int_0^1 \frac{1}{p^2}\nabla_{\boldsymbol{\theta}}Q(p;\boldsymbol{\theta}).\nabla_{\boldsymbol{\theta}}^{\mathrm{T}}Q(p;\boldsymbol{\theta})f^2(Q(p))\,dp = \mathbf{0}_{d\times d}\end{aligned} \tag{5.36}$$

Details of proving equation 5.36 is deferred to Appendix A.6, so the validity of the second statement is fulfilled. Finally, the first statement of the theorem is a direct consequence of the second statement and theorem 5.2.5. □

Theorem 5.2.9 (IME and Cramer-Rao Lower Bound). *Let $Y_1, Y_2, ..., Y_n \overset{iid}{\sim} F(\cdot;\boldsymbol{\theta}_0)$ with density function $f(\cdot;\boldsymbol{\theta})$ and quantile function $Q(\cdot;\boldsymbol{\theta})$. For any K with $K \geq d$, a set of percentage points $P_K = \{\rho_0, \rho_1, \ldots, \rho_K\} \subset [0,1]$, such that $\rho_0 < \rho_1 < \ldots < \rho_K$, is chosen to compute sample interquantile means. Let $\mathring{I}(\boldsymbol{\theta})$ be the information matrix* *associated with distribution function $F(\cdot;\boldsymbol{\theta})$, and $(B_K^{\mathrm{T}}\Omega_K^{-1}B_K)^{-1}$ be the asymptotic covariance matrix associated with the most asymptotically efficient IME estimator using sample interquantile mean at percentage points set P_K. Then:*

$$(B_K^{\mathrm{T}}\Omega_K^{-1}B_K)^{-1} \geq \mathring{I}^{-1}(\boldsymbol{\theta}_0),$$

and as long as $\lim_{K\to\infty} \max\left\{|\rho_{j+1}-\rho_j| : j = 0,1,2,\dots K\right\} = 0, (\rho_0 = 0, \rho_K = 1)$ holds, the Cramer-Rao lower bound is achievable in limit:

$$\lim_{K\to\infty}(B_K^{\mathrm{T}}\Omega_K^{-1}B_K)^{-1} = \mathring{I}^{-1}(\boldsymbol{\theta}_0)$$

Proof. (Heuristic) Let $\tilde{B}$ and $\tilde{\Omega}$ be the continuous version of discrete matrices B and Ω previously defined in subsection 4.2.2 as $K \to \infty$ with $\lim P_K$ is dense in $(0,1)$. Let $\tilde{A}$ and $\tilde{\Sigma}$ be the continuous version of discrete matrices A and Σ defined in theorem 5.2.8. Due to the fundamental connection between sample quantile and sample interquantile mean (see equation 4.31), and connection between quantile and interquantile mean (see equation 2.4) we have:

$$B_K = M_K\tilde{A} \tag{5.37}$$

$$\Omega_K = M_K\tilde{\Sigma}M_K^{\mathrm{T}} \tag{5.38}$$

for some linear operator M_K formulated using percentage points set P_K. The inverse covariance matrix of IME is $B_K^{\mathrm{T}}\Omega_K^{-1}B_K = \tilde{A}^{\mathrm{T}}M_K^{\mathrm{T}}(M_K\tilde{\Sigma}M_K^{\mathrm{T}})^{-1}M_K\tilde{A}$. Because

$$\begin{aligned}&\tilde{A}^{\mathrm{T}}\tilde{\Sigma}^{-1}\tilde{A} - \tilde{A}^{\mathrm{T}}M_K^{\mathrm{T}}(M_K\tilde{\Sigma}M_K^{\mathrm{T}})^{-1}M_K\tilde{A}\\ &\quad = \tilde{A}^{\mathrm{T}}\tilde{\Sigma}^{-1/2}\big(I - \Sigma^{1/2}M_K^{\mathrm{T}}(M_K^{\mathrm{T}}\tilde{\Sigma}M_K^{\mathrm{T}})^{-1}M_K\tilde{\Sigma}^{1/2}\big)\tilde{\Sigma}^{-1/2}\tilde{A}\end{aligned} \tag{5.39}$$

is idempotent, and $\tilde{A}^{\mathrm{T}}\tilde{\Sigma}^{-1}\tilde{A} = \mathring{I}(\boldsymbol{\theta})$ as showed in theorem 5.2.8, we have $B_K^{\mathrm{T}}\Omega_K^{-1}B_K \leq I(\boldsymbol{\theta}_0)$ as desired. The Cramer-Rao lower bound is clearly achievable in limit. □

We close off this chapter by an efficiency conjecture.

Conjecture 5.2.10 (QME and MLE Asymptotic Equivalence)**.** *Let* $Y_1, Y_2, ..., Y_n \overset{iid}{\sim} F(\cdot;\boldsymbol{\theta}_0)$ *with density function* $f(\cdot;\boldsymbol{\theta})$ *and quantile function* $Q(\cdot;\boldsymbol{\theta})$*. Instead of using maximum likelihood, we perform the following quantile matching strategy: match* Y_1 *with* $Q(1/n;\boldsymbol{\theta})$*, match* Y_2 *with* $Q(2/n;\boldsymbol{\theta})$*, match* Y_3 *with* $Q(3/n;\boldsymbol{\theta})$ *and so on. Suppose for any given* n*, the matching estimator we use is the most asymptotically efficient (using two step QME for instance), then the corresponding estimator, say* $\hat{\boldsymbol{\theta}}_n$*, is consistent to* $\boldsymbol{\theta}_0$*, asymptotically normal, and asymptotically equivalent to its MLE counterpart.*

The matching scheme in the conjecture above can be considered as generalised non-linear least square. If the conjecture is correct, then QME, despite being unfeasible with microdata when the sample size is very large, is on par with the MLE and hence being asymptotically the most efficient.

Chapter 6

Application: Estimating Income Distributions in Australia

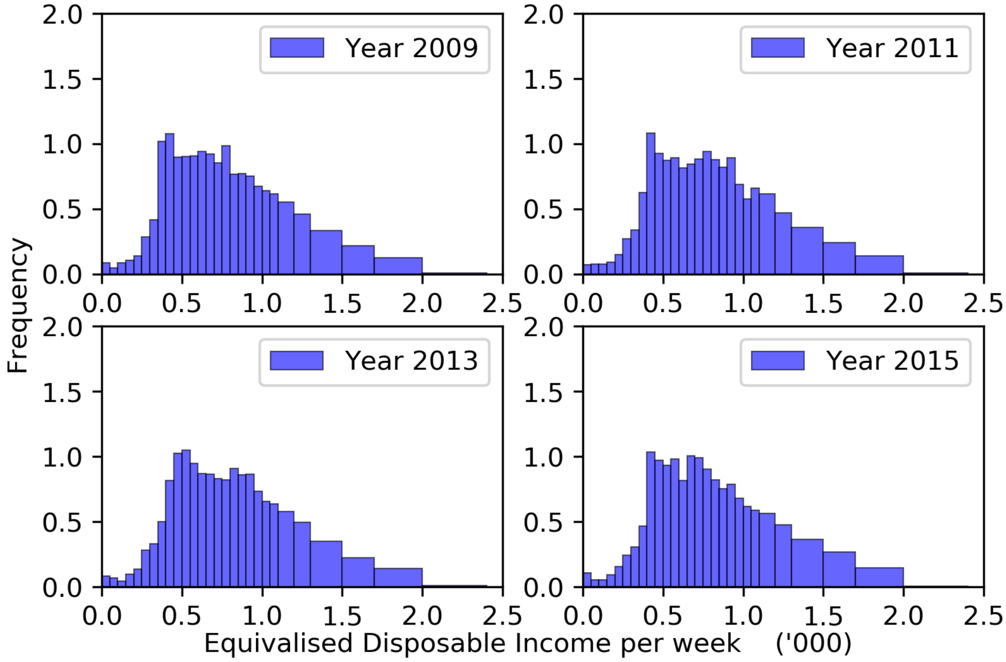
In this chapter, the proposed methods are employed to recover the Australian income distributions over a span of 14 periods using GB2, Log-normal and Gamma distributions. The estimated distributions then enable us to explore the concentration levels of income and inequality dynamics of Australia over time.

6.1 Data Description

We use data from ABS survey on *Australian Household Income and Wealth*. There are three types of income data available in the survey: class frequency data, interquantile mean data and quantile data. The comprehensive dataset extends over 14 periods from 1994 to 2015 with the absence of data for several years in between. A time series of Gini statistics, computed based on microdata and provided by ABS, is considered an important empirical benchmark for our estimation.

The income data used in our estimation is *equivalised disposable income*, which is the total income of a household, after tax and other deductions, that is available for spending or saving, divided by the number of household members converted into equalised adults. Household members are equalised or made equivalent by weighting

Figure 6-1: CLASS FREQUENCY PLOTS



Notes: Frequency plots are drawn for 4 years 2009, 2011, 2013, 2015. The plot is normalised so that the blue (or shaded) area in each graph is 1.

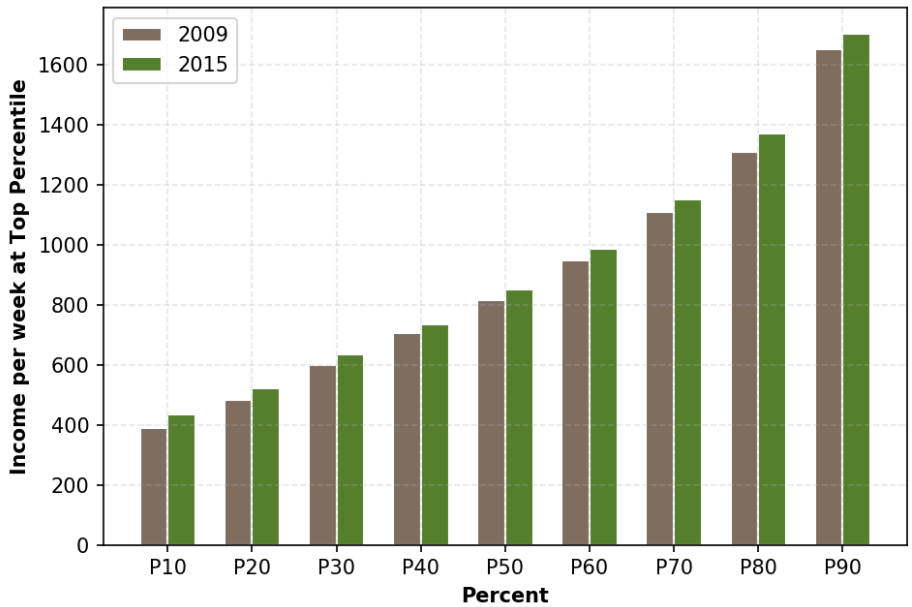
each according to their age, using the so-called modified OECD equivalence scale[1]. The data are also adjusted to CPI factor, so the comparison over time is possible and meaningful.

Australian income data in class frequency format can be found in Appendix B, table B.1. In this table, income range is subdivided into 29 intervals, including a special income range, namely "No Income". Since the fraction of people having zero income is extremely small, we shall ignore that income group from now on. Frequency graphs for individual equivalised income data are shown in Figure 6-1 using data on four recent periods.

Australian income data summarised by quantiles are placed in Table B.2. In this table, sample quantiles are reported at 0.1, 0.2, 0.3, . . . , 0.9, also known as the 10^{th}, 20^{th}, . . . , 90^{th} percentiles. The reported quantiles take into account people with

[1]See `http://www.oecd.org/eco/growth/OECD-Note-EquivalenceScales.pdf`

Figure 6-2: SAMPLE QUANTILES BAR PLOT

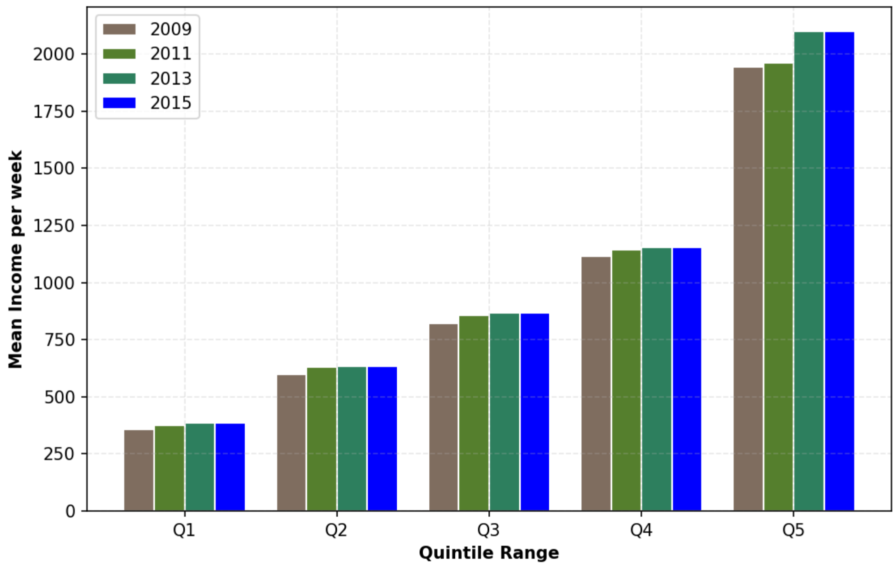


Notes: The chart depicts the income levels at 9 given quantiles for 2 periods: 2009 and 2015. There is a noticeable rise in earnings between two periods, especially at high income groups P80 and P90.

no income without adjustment. However, since the portion of null-income people is negligible, the specification of positive support is justifiable. Figure 6-2 shows the bar plot for sample quantiles in 2009 and 2015. The figure depicts a consistent improvement in income levels at every given percentile, indicating the rightward of the entire income distribution.

Australian income data summarised by interquantile means are placed in Table B.3. In this table, the whole income sample is partitioned into 5 equal subgroups, also known as quintiles, and the sample averages for each group are computed. Again the reported statistics do not adjust for the portion of null-income people. However, ignoring this issue is acceptable since the null-income group is tiny in comparison to the whole population. Figure 6-3 again depicts the steady improvement in average real income at any given quintile. Both Figure 6-2 and Figure 6-3 show a noticeable

Figure 6-3: SAMPLE INTERQUANTILE MEANS BAR PLOT

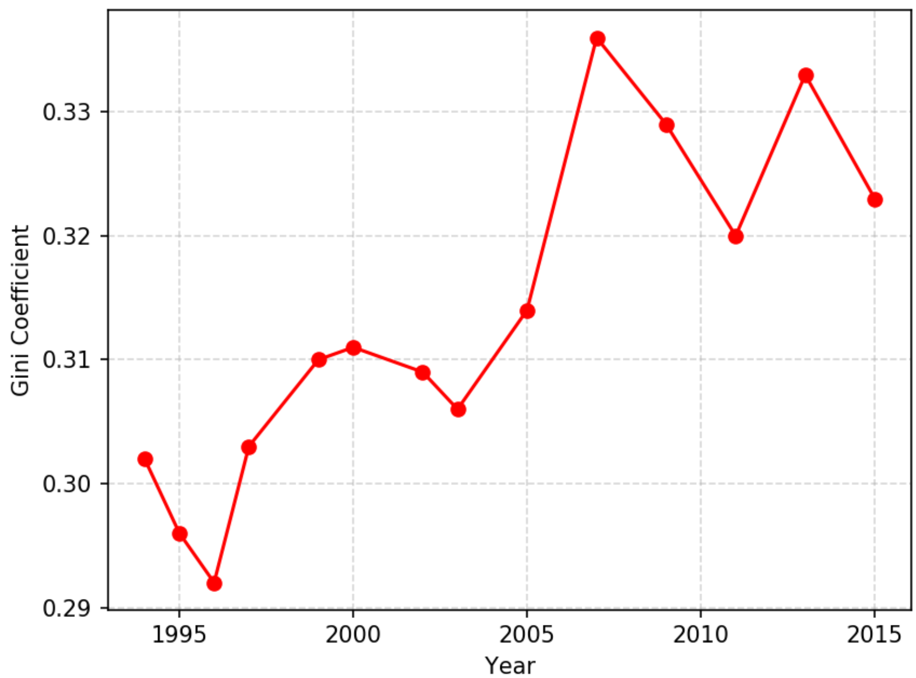


Notes: The chart depicts the average income at 5 quintiles for 4 periods: 2009, 2011, 2013, 2015. There is a consistent increase in earnings over 4 periods, indicating a right-shift of the underlying income distribution.

increase in income level for people with very high earnings. The illustration gives us some signals that the very high income group is getting richer at a faster pace than the rest of the population, which is one major determinant of rising inequality (Koechlin, 2013).

Gini coefficient is a measure of inequality that is dominantly used by the ABS. With income microdata, ABS computed and reported the Gini statistics in the survey. Gini coefficient will, later on, be an important benchmark to assess the fitness of our estimation by verifying the closeness between estimated Gini coefficients and census Gini coefficients. Figure 6-4 displays a time series graph of Gini coefficients in Australia for household income from 1994-2015.

Figure 6-4: GINI STATISTICS, Australia, 1994-1995 to 2015-2016



Notes: Gini statistics computed using micro-observations and reported by the ABS. The picture reveals a slight growth of income inequality in Australia in terms of Gini coefficient.

6.2 Estimating Income Distributions in Australia

To make the analysis tractable, there are two important assumptions we impose: random sampling and correct specification. Firstly, we assume the sample of roughly 20 million observations ABS obtained is a random sampling from a hypothetical distribution without worrying about the census nature of the data. This viewpoint is previously discussed in chapter 4. Furthermore, this assumption is likely to be wrong because ABS obtained income data for around 8 million households and employed OECD scales to equivalise the earnings for each household member. Instead of having 20 million observations, the ABS's effective number of data points is only 8 million. Nevertheless, assuming random sampling is standard and necessary for many econometric analyses, and we proceed the estimation by assuming iid from now on.

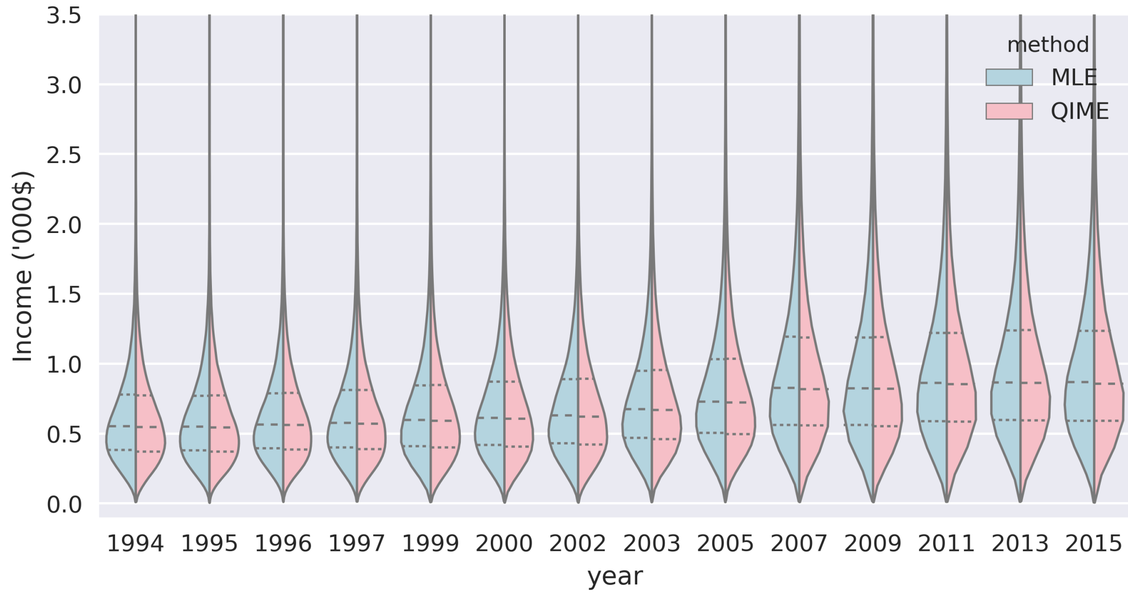
Secondly, we assume that the income distribution has a specified functional form. The use of Generalised Beta is justified at this point since the family has superb flexibility and encompasses a wide range of well-known distributions (see chapter 2). However, the parametrised form of Generalised Beta family exhibits excessive multicollinearity (see chapter 5 for a short discussion on how to measure the degree of multicollinearity). To circumvent the problem, we employ a more restricted class of distributions, namely GB2. The use of GB2 family is slightly more helpful but still exhibits a considerable degree of multicollinearity. To my surprise, the estimation for means, Gini coefficients, and other inequality indices are insensitive to all the distributions used in this paper.

We will proceed the estimation in two parts. In part I, various distributions such as GB2, Gamma, Log-normal, mixture of Log-normal, and mixture of Gamma serve as a descriptive model for income. Except for GB2 family, all the other distributions are estimated using EM algorithm developed in chapter 3. Concerning part II, we employ QME, IME and the joint QIME developed in chapter 4 to estimate the parameters of Gamma, Log-normal, and GB2 distributions. Detailed estimation outputs with standard errors for part I and part II can be found in Appendix C and Appendix D.

Estimation results for both parts are succinctly depicted in Figure 6-5. The figure plots the estimated GB2 density functions using MLE and joint QIME side by side in vertical direction. Both estimation techniques, despite being operated on different kinds of data, show excellent agreement in terms of distributional shapes. Moreover, the plot reveals a steady upward movement in the underlying income distribution over time. Based on the estimated quantiles (marked by horizontal dotted line), there is a marked increase in overall income level from 2005 to 2007, followed by a period of no improvement in income distribution and then the income distribution slowly moves upwards after that. This feature is consistent with the event that Australian economy was only moderately affected by the Great Financial Crisis in 2008 without falling into recession[2]. The same plots repeated for Gamma and Log-normal distributions

[2] `https://www.worldfinance.com/special-reports/recession-proof-australia`

Figure 6-5: SERIES OF GB2 INCOME DISTRIBUTIONS, 1994 to 2015



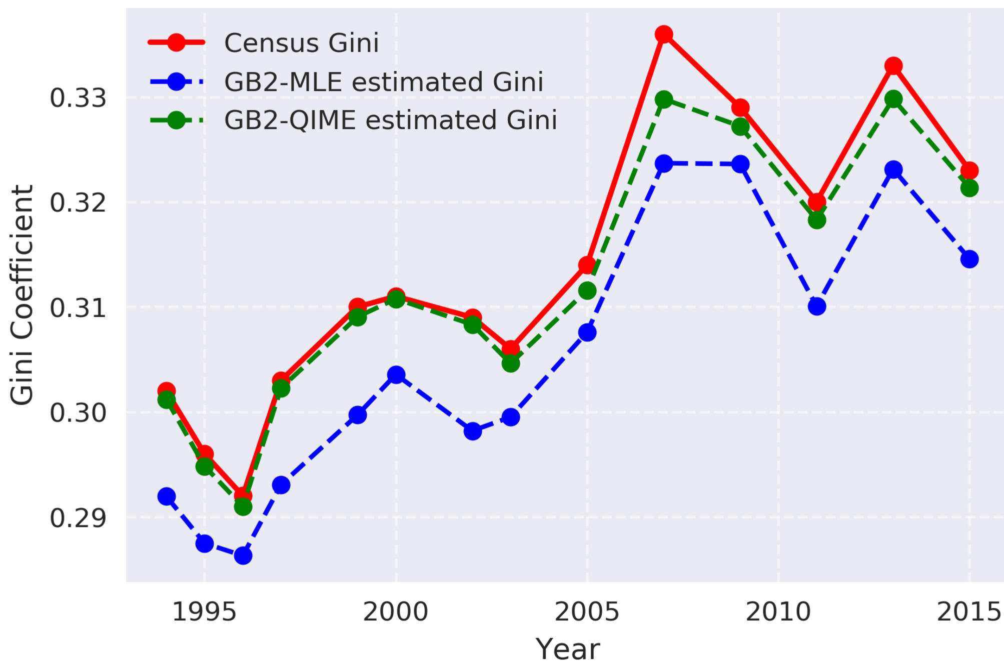
Notes: A violin plot is drawn to analyse the Australian income dynamics over 14 periods using MLE and QIME with GB2 as a model of income distribution. The symmetric density graphs in all periods show excellent agreement between two methods. The picture also reveals a consistent upward shift of overall income level over time. Horizontal dotted lines indicate the estimated first quartile, the median, and the third quartile.

can be found in Appendix E with similar highlighted features.

6.3 Gini Coefficients as Benchmark and Inequality measures

Another way of assessing the quality of our estimation is using census estimates of Gini coefficient. The parameters of GB2 family are estimated using two distinct techniques, namely MLE and QIME and we then extract the Gini statistics from the estimated distributions. Figure 6-6 then plots the census series of Gini coefficients in conjunction with two series of estimated Gini statistics. The extracted Gini coefficients from QIME track the census series exceptionally closed while the extracted Gini coefficient from MLE is only able to mimic the overall trend. It seems the series

Figure 6-6: CENSUS END ESTIMATED GINI COEFFICIENTS, 1994 to 2015

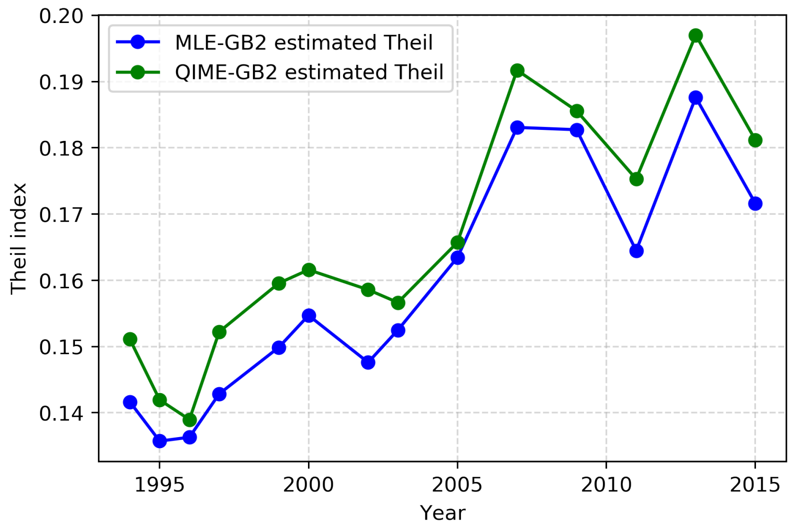


Notes: A time series plot are drawn to compare the census Gini coefficients (provided by the ABS) and the estimated Gini coefficients using MLE and QIME. GB2 is employed as a model for the hypothetical income distribution. Estimated Gini statistics using QIME track the census Gini coefficient remarkably closed.

associated with GB2 estimates consistently underestimate their census versions in every period. One possible explanation is the truncation process of null-income group. Despite being tiny in size, the deletion of low income group can potentially bias our estimates. The comparison between census estimates and extracted estimates using MLE and QIME are repeated for Gamma and Log-normal distributions, and related figures can be found in the Appendix E.

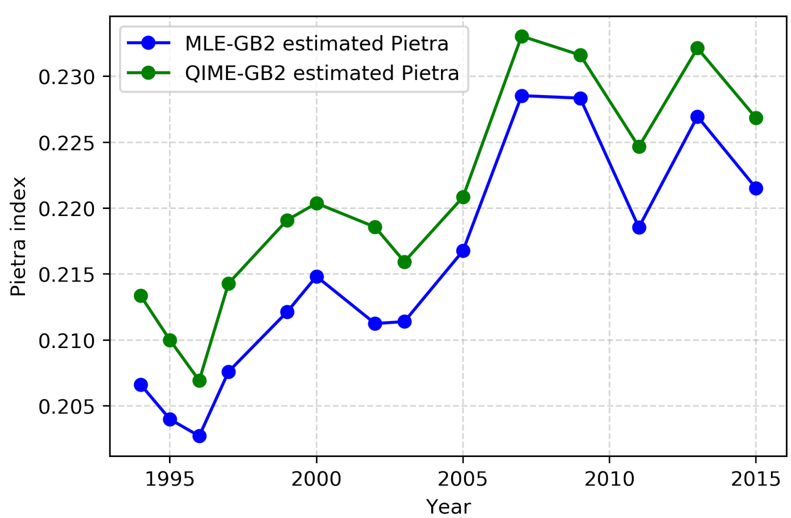
Finally, the estimated income distributions allow us to look at the inequality trend with regards to other indices such as Theil index and Pietra index (introduced in chapter 2). Figure 6-7 shows the movement of Theil index and Figure 6-8 shows the movement of Pietra index. The estimated series show a consistent upward trend in income inequality. Similar figures using different distributional assumptions can be found in Appendix E.

Figure 6-7: CENSUS END ESTIMATED THEIL COEFFICIENTS, 1994 to 2015



Notes: A time series plot are drawn to depict the movement of Theil index over time, extracted from MLE and QIME using GB2 distribution. Both estimated Theil series exhibit an increasing trend in equality consistent with the Gini measure.

Figure 6-8: ESTIMATED PIETRA COEFFICIENTS, 1994 to 2015



Notes: A time series plot is drawn to depict the movement of Pietra index over time, extracted from MLE and QIME using GB2 distribution. Both estimated Pietra series exhibit an increasing trend in equality consistent with the Gini measure.

Chapter 7

Conclusion

This thesis has made contribution to the branch of distributional analysis in inequality literature. By applying and extending the GMM framework (Hansen, 1982), I obtain a new estimation framework for analysing grouped data, followed by a fruitful series of results that are both of practical and theoretical interest. There is still much room for improvement in this research direction. On the theoretical side, many efficiency questions still wait to be discovered. To name a few, what is the convergence rate of the asymptotic covariance matrix to its lowest bound as the number of groups tend to infinity? Is the efficiency hypothesis in chapter 5 correct? How to incorporate the new data type, which I shall call truncated mean data, into the quadratic matching framework. An example of this data type can be seen at Chotikapanich (2008, p. 75). Can the associated estimator achieve its highest limiting efficiency? On the practical side, the optimisation routines I employ in this thesis are gradient free and prone to producing sub-optimal outcomes; therefore, finding an efficient implementation of new estimation techniques are the priority for future research. With efficient implementation, one can potentially tackle many inequality questions using various available sources of income data.

Appendix A

Mathematical Supplement

A.1 Proof of Theorem 2.1.4

From the parametric equation of the Lorenz curve in theorem 2.1.2, the area under the Lorenz curve is obtained by:

$$A = \int_0^\infty \pi_2(y) f_Y(y)\, dy = \frac{1}{\mathrm{E}Y} \int_0^\infty \int_0^t t f_Y(t) f_Y(y)\, dt\, dy \tag{A.1}$$

Switching the order of integration with proper adjustments on the bounds, A can be also expressed as:

$$\begin{aligned} A &= \frac{1}{\mathrm{E}Y} \int_0^\infty \int_t^\infty t f_Y(t) f_Y(y)\, dy\, dt & \text{(A.2)} \\ &= \frac{1}{\mathrm{E}Y} \int_0^\infty t f_Y(t)(1 - F_Y(t))\, dt & \text{(A.3)} \\ &= 1 - \frac{1}{\mathrm{E}Y} \int_0^\infty t f_Y(t) F_Y(t)\, dt & \text{(A.4)} \end{aligned}$$

Let I be the integral involved in equation A.4. Using integration by part:

$$I = -t F_Y(t)(1 - F_Y(t))\Big|_0^\infty + \int_0^\infty (1 - F_Y(t))(F(t) + t f(t))\, dt \tag{A.5}$$

With simple manipulation, one can show

$$2I = \mathrm{E}Y + \int_0^\infty F_Y(t)(1 - F_Y(t))\,dt \tag{A.6}$$

It completes the proof because

$$G = 1 - 2A = 1 - 2\left(1 - \frac{1}{\mathrm{E}Y} I\right) = \frac{1}{\mathrm{E}Y}\int_0^\infty F_Y(t)(1 - F_Y(t))\,dt \tag{A.7}$$

A.2 Proof of Equation 4.31

We wish to show, for $p_1, p_2 \in (0,1)$ with $p_1 < p_2$, and for any given $\epsilon > 0$:

$$\frac{1}{\lfloor np_2 \rfloor - \lfloor np_1 \rfloor} \sum_{i=\lfloor np_1 \rfloor + 1}^{\lfloor np_2 \rfloor} Y_{i:n} = \frac{1}{p_2 - p_1}\int_{p_1}^{p_2} \hat{\xi}_p\,dp + o_p(n^{-1+\epsilon}) \tag{A.8}$$

I will denote the left hand side expression A. From the definition $\hat{\xi}_p = \inf\{y : \hat{F}_n(y) \geq p\}$, we obtain a simple fact that $\hat{\xi}_p = Y_{i:n}$ for any $p \in (\frac{i-1}{n}, \frac{i}{n}]$, i running from 1 to n. Therefore the right hand side integral can be expanded into:

$$\frac{1}{p_2 - p_1}\int_{p_1}^{p_2} \hat{\xi}_p\,dp = \frac{1}{p_2 - p_1}\left(\int_{\lfloor np_1 \rfloor/n}^{\lfloor np_2 \rfloor/n} \hat{\xi}_p\,dp + \int_{\lfloor np_1 \rfloor/n}^{p_1} \hat{\xi}_p\,dp - \int_{\lfloor np_2 \rfloor/n}^{p_2} \hat{\xi}_p\,dp\right) \tag{A.9}$$

With some thinking, one can rewrite each component in the RHS of equation A.9 as:

$$B = \frac{1}{p_2 - p_1}\int_{\lfloor np_1 \rfloor/n}^{\lfloor np_2 \rfloor/n} \hat{\xi}_p\,dp = \frac{1}{np_2 - np_1}\sum_{i=\lfloor np_1 \rfloor + 1}^{\lfloor np_2 \rfloor} Y_{i:n} \tag{A.10}$$

$$C = \int_{\lfloor np_1 \rfloor/n}^{p_1} \hat{\xi}_p\,dp = \left(p_1 - \frac{\lfloor np_1 \rfloor}{n}\right) Y_{\lfloor np_1 \rfloor + 1:n} \tag{A.11}$$

$$D = \int_{\lfloor np_2 \rfloor/n}^{p_2} \hat{\xi}_p\,dp = \left(p_2 - \frac{\lfloor np_2 \rfloor}{n}\right) Y_{\lfloor np_2 \rfloor + 1:n} \tag{A.12}$$

The problem is reduced to show $A - B - C + D$ is of probability order $o_p\left(n^{-1+\epsilon}\right)$. In fact since $(p - \lfloor np \rfloor/n) \propto 1/n$ and $Y_{\lfloor np \rfloor + 1:n} \sim o_p(n^\epsilon)$ for all p, it must be the case

C, $D \sim o_p(n^{-1+\epsilon})$ as required. Furthermore:

$$A - B = \left(\frac{1}{\lfloor np_2 \rfloor - \lfloor np_1 \rfloor} - \frac{1}{np_2 - np_1} \right) \sum_{i=\lfloor np_1 \rfloor + 1}^{\lfloor np_2 \rfloor} Y_{i:n} \tag{A.13}$$

The deterministic factor is $\propto 1/n^2$ while the random factor is $\sim o_p(n^{1+\epsilon})$. Therefore we can deduce $A - B \sim o_p(n^{-1+\epsilon})$, completing the proof. In proving asymptotic behaviours, we often magnify $\hat{\xi}_p$ by $n^{1/2}$ and the corresponding magnified "noise" term remains negligible (of order $o_p(n^{-1/2+\epsilon})$). Therefore, I will simply ignore the noise component and regard the equation 4.31 as true equality throughout the paper.

A.3 Proof of Equation 4.36

The double integral on B is on a rectangle $(p_1, p_2) \times (p_1, p_2)$ and the integrand is multiplicative in p and q, therefore it can be factored out as a product of two identical simple integral. After that, we make a simple change of variable, for instance $p = Q(y)$ for each simple integral, the result for B is:

$$B = \int_{p_1}^{p_2} \frac{p}{f(Q(p))}\, dp \int_{p_1}^{p_2} \frac{q}{f(Q(q))}\, dq \tag{A.14}$$

$$= \Big(p_2 Q(p_2) - p_1 Q(p_1) - (p_2 - p_1)\mu(p_1, p_2) \Big)^2 \tag{A.15}$$

By the same change of variable and integration by part, A can be broken into 3 pieces:

$$A = \int_{p_1}^{p_2} \frac{1}{f(Q(q))} \Big(qQ(q) - p_1 Q(p_1) - (q - p_1)\mu(p_1, q) \Big)\, dq \tag{A.16}$$

$$= A_1 - A_2 - A_3 \tag{A.17}$$

With the same change of variable, the result of A_1 is:

$$A_1 = \int_{p_1}^{p_2} \frac{qQ(q)}{f(Q(q))}\, dq = \int_{Q(p_1)}^{Q(p_2)} yF(y)\, dy$$

$$= \frac{1}{2} \Big(p_2 Q^2(p_2) - p_1 Q^2(p_1) - (p_2 - p_1)\mu^{(2)}(p_1, p_2) \Big) \tag{A.18}$$

Similarly, The result of A_2 is:

$$A_2 = \int_{p_1}^{p_2} \frac{p_1 Q(p_1)}{f(Q(q))}\, dq = p_1 Q(p_1) \int_{Q(p_1)}^{Q(p_2)} 1\, dy$$
$$= p_1 Q(p_1)\Big(Q(p_2) - Q(p_1)\Big) \tag{A.19}$$

Similarly, the result of A_3 is:

$$A_3 = \int_{p_1}^{p_2} \frac{(q - p_1)\mu(p_1, q)}{f(Q(q))}\, dq = \int_{p_1}^{p_2} \int_{p_1}^{q} \frac{Q(p)}{f(Q(q))}\, dp\, dq \tag{A.20}$$

We switch the order of integration from $dp\, dq$ to $dq\, dp$ with proper adjustment of integration bounds:

$$A_3 = \int_{p_1}^{p_2} \int_{p}^{p_2} \frac{Q(p)}{f(Q(q))}\, dq\, dp = \int_{p_1}^{p_2} Q(p)(Q(p_2) - Q(p))\, dp$$
$$= Q(p_2)(p_2 - p_1)\mu(p_1, p_2) - (p_2 - p_1)\mu^{(2)}(p_1, p_2) \tag{A.21}$$

I leave readers confirming the following calculation, completing the proof:

$$2(A_1 - A_2 - A_3) - B = (p_2 - p_1)\sigma_H^2 + (Q(p_1) - \mu_H)^2 p_1(1 - p_1)$$
$$+ (Q(p_2) - \mu_H)^2 p_2(1 - p_2) - 2(Q(p_1) - \mu_H)(Q(p_2) - \mu_H)p_1(1 - p_2) \tag{A.22}$$

A.4 Evaluation of Covariance in Theorem 4.5.5

There are 6 cases in consideration. To compute the integral for the covariance in each case, we make a change of variable $p = F(y)$, $dp = f(y)dy$ and obtain the formula as follows:

(1) If $F(a) \leq F(b) \leq p_1 \leq p_2$, the cross covariance is:

$$\psi = \frac{F(a) - F(b)}{p_2 - p_1}\Big[(1 - p_2)Q(p_2) - (1 - p_1)Q(p_1) + (p_2 - p_1)\mu(p_1, p_2)\Big] \tag{A.23}$$

(2) If $F(a) \le p_1 \le F(b) \le p_2$, the cross covariance is:

$$\begin{aligned}\psi = &\frac{F(a)}{p_2 - p_1}\Big[(1-p_2)Q(p_2) - (1-p_1)Q(p_1) + (p_2 - p_1)\mu(p_1, p_2)\Big] \\ &- \frac{1-F(b)}{p_2-p_1}\Big[F(b)b - p_1 Q(p_1) - (F(b) - p_1)\mu(p_1, F(b))\Big] \\ &- \frac{F(b)}{p_2 - p_1}\Big[(1-p_2)Q(p_2) - (1-F(b))b + (p_2 - F(b))\mu(F(b), p_2)\Big] \end{aligned} \tag{A.24}$$

(3) If $F(a) \le p_1 \le p_2 \le F(b)$, the cross covariance is:

$$\begin{aligned}\psi = &\frac{F(a)}{p_2 - p_1}\Big[(1-p_2)Q(p_2) - (1-p_1)Q(p_1) + (p_2 - p_1)\mu(p_1, p_2)\Big] \\ &- \frac{1-F(b)}{p_2-p_1}\Big[p_2 Q(p_2) - p_1 Q(p_1) - (p_2 - p_1)\mu(p_1, p_2)\Big] \end{aligned} \tag{A.25}$$

(4) If $p_1 \le F(a) \le F(b) \le p_2$, the cross covariance is:

$$\begin{aligned}\psi = &\Bigg(\frac{1-F(a)}{p_2-p_1}\Big[F(a)a - p_1 Q(p_1) - (F(a) - p_1)\mu(p_1, F(a))\Big] + \\ &\frac{F(a)}{p_2 - p_1}\Big[(1-p_2)Q(p_2) - (1-F(a))a + (p_2 - F(a))\mu(F(a), p_2)\Big]\Bigg) \\ &- \Bigg(\frac{1-F(b)}{p_2-p_1}\Big[F(b)b - p_1 Q(p_1) - (F(b) - p_1)\mu(p_1, F(b))\Big] + \\ &\frac{F(b)}{p_2 - p_1}\Big[(1-p_2)Q(p_2) - (1-F(b))b + (p_2 - F(b))\mu(F(b), p_2)\Big]\Bigg) \end{aligned} \tag{A.26}$$

(5) If $p_1 \le F(a) \le p_2 \le F(b)$, the cross covariance is:

$$\begin{aligned}\psi = &\Bigg(\frac{1-F(a)}{p_2-p_1}\Big[F(a)a - p_1 Q(p_1) - (F(a) - p_1)\mu(p_1, F(a))\Big] + \\ &\frac{F(a)}{p_2 - p_1}\Big[(1-p_2)Q(p_2) - (1-F(a))a + (p_2 - F(a))\mu(F(a), p_2)\Big]\Bigg) \\ &- \frac{1-F(b)}{p_2-p_1}\Big[p_2 Q(p_2) - p_1 Q(p_1) - (p_2 - p_1)\mu(p_1, p_2)\Big] \end{aligned} \tag{A.27}$$

(6) If $p_1 \leq p_2 \leq F(a) \leq F(b)$, the cross covariance is:

$$\psi = \frac{F(b) - F(a)}{p_2 - p_1}\Big[p_2 Q(p_2) - p_1 Q(p_1) - (p_2 - p_1)\mu(p_1, p_2)\Big] \tag{A.28}$$

A.5 Proof of Lemma 5.1.2

The following explanation is taken from *StackExchange*[1]: Since $A - B \geq 0$, we have (conjugating with $B^{-1/2}$) that $B^{-1/2}AB^{-1/2} - I \geq 0$. This tells you that all the eigenvalues of the positive definite matrix $B^{-1/2}AB^{-1/2}$ are at least 1. Now note that $B^{-1/2}AB^{-1/2} = (B^{-1/2}A^{1/2})(A^{1/2}B^{-1/2})$. Since commuting the product of two matrices does not change the set of eigenvalues, we obtain $A^{1/2}B^{-1/2}B^{-1/2}A^{1/2} = A^{1/2}B^{-1}A^{1/2}$ also has all eigenvalues at least 1. So $A^{1/2}B^{-1}A^{1/2} - I \geq 0$. Finally, we conjugate with $A^{-1/2}$ and obtain $B^{-1} - A^{-1} \geq 0$.

A.6 Proof of Equation 5.36

Making a change of variable $p = F(y; \boldsymbol{\theta})$, $y = Q(p; \boldsymbol{\theta})$, equation 5.36 is equivalent to:

$$\int_{-\infty}^{\infty} \frac{f(y;\boldsymbol{\theta})}{F(y;\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} Q(F(y);\boldsymbol{\theta}).\nabla_{\boldsymbol{\theta}}^{\mathrm{T}} f(y;\boldsymbol{\theta})\, dy + \int_{-\infty}^{\infty} \frac{f(y;\boldsymbol{\theta})}{F(y;\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} f(y;\boldsymbol{\theta}).\nabla_{\boldsymbol{\theta}}^{\mathrm{T}} Q(F(y);\boldsymbol{\theta})\, dy \\ + \int_{-\infty}^{\infty} \frac{f^3(y;\boldsymbol{\theta})}{F^2(y;\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} Q(F(y);\boldsymbol{\theta}).\nabla_{\boldsymbol{\theta}}^{\mathrm{T}} Q(F(y);\boldsymbol{\theta})\, dy = \mathbf{0}_{d\times d} \tag{A.29}$$

Using the fact that $y = Q(F(y;\boldsymbol{\theta});\boldsymbol{\theta})$, taking the gradient with respect to $\boldsymbol{\theta}$ for both sides yields:

$$0 = Q'(F(y);\boldsymbol{\theta})\nabla_{\boldsymbol{\theta}} F(y;\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}} Q(F(y);\boldsymbol{\theta}) \tag{A.30}$$

Recognising that $Q'(F(y);\boldsymbol{\theta}) = 1/f(y;\boldsymbol{\theta})$ by inverse function theorem, we then have:

$$\nabla_{\boldsymbol{\theta}} Q(F(y);\boldsymbol{\theta}) = -\frac{\nabla_{\boldsymbol{\theta}} F(y;\boldsymbol{\theta})}{f(y;\boldsymbol{\theta})} \tag{A.31}$$

[1] `https://math.stackexchange.com/questions/435831`

With this neat expression, equation A.29 can be rewritten as:

$$\begin{aligned}
&\int_{-\infty}^{\infty} \frac{f(y;\boldsymbol{\theta})}{F^2(y;\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} F(y;\boldsymbol{\theta}).\nabla_{\boldsymbol{\theta}}^{\mathrm{T}} F(y;\boldsymbol{\theta})\, dy \\
&= \int_{-\infty}^{\infty} \frac{\nabla_{\boldsymbol{\theta}} F(y;\boldsymbol{\theta}).\nabla_{\boldsymbol{\theta}}^{\mathrm{T}} f(y;\boldsymbol{\theta})}{F(y;\boldsymbol{\theta})}\, dy + \int_{-\infty}^{\infty} \frac{\nabla_{\boldsymbol{\theta}} f(y;\boldsymbol{\theta}).\nabla_{\boldsymbol{\theta}}^{\mathrm{T}} F(y;\boldsymbol{\theta})}{F(y;\boldsymbol{\theta})}\, dy
\end{aligned} \tag{A.32}$$

We almost reach the end of the proof. By switching the order of differentiation:

$$\frac{d}{dy} \nabla_{\boldsymbol{\theta}} F(y;\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} f(y;\boldsymbol{\theta}) \tag{A.33}$$

Finally, using integration by part, we have:

$$\begin{aligned}
&\int_{-\infty}^{\infty} \frac{\nabla_{\boldsymbol{\theta}} F(y;\boldsymbol{\theta}).\nabla_{\boldsymbol{\theta}}^{\mathrm{T}} f(y;\boldsymbol{\theta})}{F(y;\boldsymbol{\theta})}\, dy = \int_{-\infty}^{\infty} \frac{\nabla_{\boldsymbol{\theta}} F(y;\boldsymbol{\theta})}{F(y;\boldsymbol{\theta})}.\, d\nabla_{\boldsymbol{\theta}}^{\mathrm{T}} F(y;\boldsymbol{\theta}) \\
&= \left[\frac{\nabla_{\boldsymbol{\theta}} F(y;\boldsymbol{\theta}).\nabla_{\boldsymbol{\theta}}^{\mathrm{T}} F(y;\boldsymbol{\theta})}{F(y;\boldsymbol{\theta})}\right]_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \frac{d}{dy}\left[\frac{\nabla_{\boldsymbol{\theta}} F(y;\boldsymbol{\theta})}{F(y;\boldsymbol{\theta})}\right].\nabla_{\boldsymbol{\theta}}^{\mathrm{T}} F(y;\boldsymbol{\theta})\, dy \\
&= -\left(\int_{-\infty}^{\infty} \frac{\nabla_{\boldsymbol{\theta}} f(y;\boldsymbol{\theta}).\nabla_{\boldsymbol{\theta}}^{\mathrm{T}} F(y;\boldsymbol{\theta})}{F(y;\boldsymbol{\theta})}\, dy - \int_{-\infty}^{\infty} \frac{f(y;\boldsymbol{\theta})}{F^2(y;\boldsymbol{\theta})} \nabla_{\boldsymbol{\theta}} F(y;\boldsymbol{\theta}).\nabla_{\boldsymbol{\theta}}^{\mathrm{T}} F(y;\boldsymbol{\theta})\, dy\right)
\end{aligned} \tag{A.34}$$

as desired.

Appendix B

Tables: Dataset

Table B.1: PERSONS BY INCOME RANGE, Australia, 1994-95 to 2015-16

	1994	1995	1996	1997	1999	2000	2002	2003	2005	2007	2009	2011	2013	2015
No income	170.7	129.6	138.6	169.6	159.5	157.8	175.4	87.3	73.7	72.6	89.0	87.4	86.4	69.7
$1 - 49	113.7	82.0	77.6	88.3	110.9	99.5	101.0	91.3	88.9	78.8	94.2	80.4	95.3	122.4
$50 - 99	84.8	82.9	53.7	72.0	63.3	63.3	76.0	52.0	62.3	42.3	53.7	86.5	75.2	62.3
$100 - 149	70.0	159.0	128.3	107.8	136.1	136.2	129.7	90.9	64.7	49.8	90.6	86.8	51.0	60.6
$150 - 199	278.8	298.5	196.6	233.1	236.9	231.9	178.2	108.0	93.9	104.8	112.9	102.7	111.7	108.6
$200 - 249	442.1	471.1	421.7	412.4	400.1	360.1	340.0	233.5	183.4	168.7	152.0	165.6	156.2	180.3
$250 - 299	1307.6	1299.3	1100.6	1015.8	927.8	888.4	866.6	521.2	365.4	255.2	309.1	298.0	318.3	284.6
$300 - 349	1606.9	1596.1	1657.8	1682.7	1601.2	1558.7	1445.5	1280.5	1003.4	775.8	450.5	373.9	373.2	353.8
$350 - 399	1309.8	1315.3	1411.4	1305.5	1399.2	1298.5	1219.3	1289.8	1208.9	986.0	1093.5	693.7	565.0	543.3
$400 - 449	1344.2	1407.8	1281.8	1434.1	1225.7	1302.5	1211.4	1221.3	1085.2	866.0	1158.8	1193.6	922.4	1199.7
$450 - 499	1072.4	1211.6	1156.0	1145.2	1149.0	1058.6	1193.7	1102.6	1032.0	845.9	963.4	1024.1	1157.9	1128.7
$500 - 549	1050.9	1106.1	1270.8	1148.8	983.9	1110.1	1141.5	987.5	986.9	1034.5	967.9	967.2	1183.0	1082.2
$550 - 599	1019.8	989.8	1062.7	894.8	1031.6	1000.9	1104.3	1164.0	1017.0	786.5	976.5	985.2	1072.4	1139.6
$600 - 649	1009.3	1043.2	1060.7	974.5	1015.7	1055.6	1059.6	995.6	1018.9	1029.2	1010.7	899.2	980.4	948.3
$650 - 699	976.9	851.4	980.4	950.3	970.7	861.8	943.2	998.9	1028.6	909.9	993.5	934.8	974.1	1165.7
$700 - 749	832.0	901.6	741.0	914.8	870.4	967.6	966.7	1084.4	1101.3	897.2	916.0	973.8	937.1	1151.8
$750 - 799	806.2	681.8	843.1	808.0	904.0	836.7	834.8	987.3	865.0	945.8	1058.4	1040.6	925.6	1047.7
$800 - 849	613.4	656.0	632.0	755.2	770.5	813.4	729.9	848.2	822.7	858.6	821.7	968.0	1023.7	954.9
$850 - 899	517.4	580.9	628.4	628.4	623.7	692.6	770.2	851.7	910.9	855.0	829.8	907.9	973.3	875.4
$900 - 949	455.5	535.0	589.0	553.6	505.5	615.3	632.9	640.6	690.7	811.1	806.6	984.8	975.0	915.1
$950 - 999	417.5	403.4	462.8	395.8	585.0	459.6	570.8	619.0	767.7	829.4	726.2	762.2	828.4	791.9
$1,000 - 1,049	339.9	396.1	329.9	442.6	470.5	455.2	474.9	546.8	627.6	640.9	686.9	639.1	738.7	716.7
$1,050 - 1,099	275.5	364.6	287.9	340.6	409.7	396.8	422.5	479.4	556.7	586.9	662.0	729.6	716.2	682.3
$1,100 - 1,199	463.8	399.5	458.0	563.2	518.3	651.2	729.1	829.1	954.4	1079.0	1189.8	1362.7	1309.2	1308.9
$1,200 - 1,299	317.8	261.9	356.0	341.8	450.7	486.7	542.0	570.0	730.4	934.4	991.8	1042.0	1114.2	1105.7
$1,300 - 1,499	374.3	300.7	334.5	419.6	575.6	591.5	651.0	772.6	974.5	1370.6	1435.3	1588.5	1577.1	1701.3
$1,500 - 1,699	127.1	169.5	211.1	169.0	198.9	258.8	343.3	436.4	496.7	931.7	939.5	1077.1	1008.4	1238.8
$1,700 - 1,999	98.8	85.5	97.4	160.6	133.2	181.1	215.1	314.2	479.4	692.4	796.6	925.0	961.9	1020.7
$2,000 or more	111.2	81.2	119.2	148.2	224.7	268.3	234.4	402.1	639.6	1204.1	1212.4	1208.7	1468.0	1332.9
All persons*	17437.6	17731.8	17950.4	18106.7	18492.8	18700.9	19127.6	19518.9	19857.1	20570.5	21500.3	22101.7	22592.9	23224.2

Notes: Equivalised disposable income data per week are summarised in *class frequency format*, with count unit is in 1000. There are 29 income classes including negative range and around 20 millions observations each year. Data is taken from ABS at `http://abs.gov.au/household-income`.
*Excluding people with negative income.

Table B.2: EQUIVALISED DISPOSABLE INCOME BY QUANTILES, Australia, 1994-95 to 2015-16

	1994	1995	1996	1997	1999	2000	2002	2003	2005	2007	2009	2011	2013	2015
10th (P10)	278	276	290	291	294	297	303	334	352	380	391	410	428	436
20th (P20)	332	330	343	346	354	361	370	407	439	492	485	512	527	523
30th (P30)	396	398	408	414	424	429	452	489	535	605	599	630	631	635
40th (P40)	463	461	484	481	503	517	534	583	634	714	707	746	751	737
50th (P50)	547	540	559	567	596	609	623	678	729	830	816	856	870	853
60th (P60)	633	624	641	661	687	709	721	771	844	951	950	979	991	987
70th (P70)	729	725	744	766	792	810	838	883	966	1104	1109	1135	1148	1153
80th (P80)	847	850	870	886	936	948	975	1039	1131	1308	1309	1339	1349	1371
90th (P90)	1051	1034	1059	1095	1144	1181	1209	1295	1422	1654	1654	1684	1741	1705

Notes: Equivalised disposable income data per week are summarised by reporting *sample quantiles*, with unit in $.

Data Source: `http://abs.gov.au/household-income`.

Table B.3: EQUIVALISED DISPOSABLE INCOME BY INTERQUANTILE MEANS, Australia, 1994-95 to 2015-16

	1994	1995	1996	1997	1999	2000	2002	2003	2005	2007	2009	2011	2013	2015
Lowest quintile	245	247	260	257	260	265	271	307	329	358	357	375	387	388
Second quintile	396	396	411	412	425	434	452	493	535	602	598	629	634	632
Third quintile	547	541	560	570	595	609	624	678	732	829	823	858	869	856
Fourth quintile	731	729	746	768	799	816	839	892	970	1109	1115	1145	1154	1162
Highest quintile	1168	1137	1168	1224	1293	1330	1356	1478	1650	2011	1946	1963	2101	2009

Notes: Equivalised disposable income data per week are summarised by reporting *sample interquantile means*, with unit in $.

Data Source: `http://abs.gov.au/household-income`.

Appendix C

Tables: Estimation Outputs Part I

Table C.1: ESTIMATED DISTRIBUTION FUNCTIONS 1994 INCOME

	GB2		G-G	LN-LN		G	LN	Expo
p	1.3290	ρ	0.9818	0.0181	a	3.5837		
	(2.81E-03)					(1.18E-03)		
q	2.4320	a_0, μ_0	0.2036	-0.5960	λ	5.7946		1.6101
	(7.49E-03)					(2.05E-03)		(3.87E-04)
a	2.2575	λ_0, σ_0^2	0.0605	0.2574	μ		-0.6242	
	(3.35E-03)						(1.35E-04)	
b	0.7642	a_1, λ_1	4.1225	-2.5158	σ^2		0.3163	
	(8.47E-04)						(1.09E-04)	
		λ_1, σ_1	6.7297	2.1867				
Mean*	0.6202					0.6184	0.6275	0.6211
	(8.48E-05)					(7.85E-05)	(9.13E-05)	(1.49E-04)
Gini**	0.2920					0.2878	0.3091	0.5000
	(5.66E-05)					(4.42E-05)	(5.07E-05)	(1.95E-17)
Theil	0.1416					0.1331	0.1582	0.4228
	(6.51E-05)					(4.18E-05)	(5.47E-05)	(1.12E-15)
Pietra	0.2066					0.2059	0.2215	0.3679
	(3.96E-05)					(3.24E-05)	(3.73E-05)	(9.62E-14)
-lnL	5.85E+05		4.33E+05	2.70E+05		7.06E+05	1.05E+06	6.13E+06

Sample size: 17,437,600 *Census estimate: 0.617 **Census estimate: 0.3020

Table C.2: ESTIMATED DISTRIBUTION FUNCTIONS 1995 INCOME

	GB2		G-G	LN-LN		G	LN	Expo
p	1.6288	ρ	0.9652	0.0203	a	3.6734		
	(4.04E-03)					(1.20E-03)		
q	3.5162	a_0, μ_0	0.7172	-0.6051	λ	6.0145		1.6323
	(1.39E-02)					(2.10E-03)		(3.89E-04)
a	1.9494	λ_0, σ_0^2	0.7890	0.2562	μ		-0.6334	
	(3.35E-03)						(1.32E-04)	
b	0.8663	a_1, λ_1	4.2005	-2.1556	σ^2		0.3068	
	(1.46E-03)						(1.05E-04)	
		λ_1, σ_1	6.9563	1.2533				
Mean*	0.6117					0.6108	0.6188	0.6126
	(8.00E-05)					(7.59E-05)	(8.77E-05)	(1.46E-04)
Gini**	0.2875					0.2845	0.3047	0.5000
	(5.21E-05)					(4.33E-05)	(4.95E-05)	(1.60E-17)
Theil	0.1357					0.1300	0.1534	0.4228
	(5.58E-05)					(4.05E-05)	(5.25E-05)	(3.31E-17)
Pietra	0.2040					0.2035	0.2182	0.3679
	(3.70E-05)					(3.17E-05)	(3.64E-05)	(6.96E-14)
-lnL	4.27E+05		3.45E+05	2.45E+05		5.01E+05	8.16E+05	6.21E+06

Sample size: 17,731,800   *Census estimate: 0.610   **Census estimate: 0.2960

Table C.3: ESTIMATED DISTRIBUTION FUNCTIONS 1996 INCOME

	GB2		G-G	LN-LN		G	LN	Expo
p	1.5862	ρ	0.8885	0.0181	a	3.7809		
	(3.67E-03)					(1.23E-03)		
q	2.5956	a_0, μ_0	1.5454	-0.5761	λ	5.9881		1.5765
	(8.35E-03)					(2.08E-03)		(3.74E-04)
a	2.1343	λ_0, σ_0^2	1.8248	0.2498	μ		-0.5960	
	(3.36E-03)						(1.28E-04)	
b	0.7411	a_1, λ_1	4.7456	-1.8775	σ^2		0.2941	
	(8.26E-04)						(1.00E-04)	
		λ_1, σ_1	7.8145	2.5303				
Mean*	0.6335					0.6314	0.6383	0.6343
	(8.41E-05)					(7.69E-05)	(8.78E-05)	(1.50E-04)
Gini**	0.2863					0.2807	0.2986	0.5000
	(5.53E-05)					(4.27E-05)	(4.84E-05)	(3.70E-17)
Theil	0.1363					0.1265	0.1471	0.4228
	(6.31E-05)					(3.93E-05)	(5.00E-05)	(2.14E-19)
Pietra	0.2027					0.2007	0.2137	0.3679
	(3.88E-05)					(3.12E-05)	(3.55E-05)	(1.35E-13)
-lnL	4.79E+05		4.20E+05	2.48E+05		6.21E+05	8.10E+05	6.59E+06

Sample size: 17,950,400 *Census estimate: 0.629 **Census estimate: 0.2920

Table C.4: ESTIMATED DISTRIBUTION FUNCTIONS 1997 INCOME

	GB2		G-G	LN-LN		G	LN	Expo
p	1.5057	ρ	0.9684	0.0189	a	3.5947		
	(3.38E-03)					(1.16E-03)		
q	2.6054	a_0, μ_0	0.5506	-0.5544	λ	5.5543		1.5366
	(8.33E-03)					(1.93E-03)		(3.63E-04)
a	2.1150	λ_0, σ_0^2	0.3899	0.2620	μ		-0.5784	
	(3.26E-03)						(1.32E-04)	
b	0.7841	a_1, λ_1	4.1820	-2.0546	σ^2		0.3128	
	(9.12E-04)						(1.06E-04)	
		λ_1, σ_1	6.5784	2.2416				
Mean*	0.6496					0.6472	0.6557	0.6508
	(8.81E-05)					(8.05E-05)	(9.30E-05)	(1.54E-04)
Gini**	0.2931					0.2874	0.3075	0.5000
	(5.63E-05)					(4.33E-05)	(4.95E-05)	(5.70E-17)
Theil	0.1428					0.1327	0.1564	0.4228
	(6.58E-05)					(4.09E-05)	(5.31E-05)	(4.46E-17)
Pietra	0.2076					0.2056	0.2202	0.3679
	(3.95E-05)					(3.17E-05)	(3.64E-05)	(6.94E-14)
-lnL	5.31E+05		4.35E+05	2.82E+05		6.62E+05	9.05E+05	6.32E+06

Sample size: 18,106,700 *Census estimate: 0.646 **Census estimate: 0.3030

Table C.5: ESTIMATED DISTRIBUTION FUNCTIONS 1999 INCOME

	GB2		G-G	LN-LN		G	LN	Expo
p	1.2667	ρ	0.9805	0.0239	a	3.4194		
	(2.64E-03)					(1.10E-03)		
q	2.3029	a_0, μ_0	0.1425	-0.5238	λ	5.0924		1.4775
	(6.85E-03)					(1.76E-03)		(3.46E-04)
a	2.2650	λ_0, σ_0^2	0.0020	0.2726	μ		-0.5487	
	(3.32E-03)						(1.35E-04)	
b	0.8243	a_1, λ_1	3.9750	-1.7810	σ^2		0.3351	
	(8.67E-04)						(1.13E-04)	
		λ_1, σ_1	6.0323	3.3056				
Mean*	0.6747					0.6715	0.6830	0.6768
	(9.33E-05)					(8.48E-05)	(9.99E-05)	(1.58E-04)
Gini**	0.2997					0.2942	0.3177	0.5000
	(5.78E-05)					(4.37E-05)	(5.06E-05)	(5.96E-17)
Theil	0.1499					0.1392	0.1676	0.4228
	(6.98E-05)					(4.24E-05)	(5.65E-05)	(5.39E-18)
Pietra	0.2121					0.2106	0.2278	0.3679
	(4.04E-05)					(3.21E-05)	(3.73E-05)	(9.50E-12)
-lnL	5.58E+05		3.78E+05	3.06E+05		6.88E+05	1.04E+06	6.08E+06

Sample size: 18,492,800   *Census estimate: 0.674   **Census estimate: 0.3100

Table C.6: ESTIMATED DISTRIBUTION FUNCTIONS 2000 INCOME

	GB2		G-G	LN-LN		G	LN	Expo
p	1.2915	ρ	0.9769	0.0250	a	3.3732		
	(2.74E-03)					(1.08E-03)		
q	2.1621	a_0, μ_0	0.2360	-0.5006	λ	4.8893		1.4360
	(6.34E-03)					(1.68E-03)		(3.35E-04)
a	2.2608	λ_0, σ_0^2	0.0180	0.2781	μ		-0.5237	
	(3.35E-03)						(1.35E-04)	
b	0.8066	a_1, λ_1	3.9307	-1.5805	σ^2		0.3385	
	(7.99E-04)						(1.14E-04)	
		λ_1, σ_1	5.8296	3.1718				
Mean*	0.6943					0.6899	0.7015	0.6964
	(9.83E-05)					(8.73E-05)	(1.03E-04)	(1.62E-04)
Gini**	0.3036					0.2961	0.3192	0.5000
	(6.02E-05)					(4.37E-05)	(5.06E-05)	(3.15E-17)
Theil	0.1547					0.1410	0.1692	0.4228
	(7.62E-05)					(4.27E-05)	(5.68E-05)	(4.72E-17)
Pietra	0.2148					0.2119	0.2289	0.3679
	(4.20E-05)					(3.21E-05)	(3.73E-05)	(6.46E-13)
-lnL	4.92E+05		3.57E+05	2.63E+05		6.42E+05	9.19E+05	5.99E+06

Sample size: 18,700,900 *Census estimate: 0.691 **Census estimate: 0.3110

Table C.7: ESTIMATED DISTRIBUTION FUNCTIONS 2002 INCOME

	GB2		G-G	LN-LN		G	LN	Expo
p	1.3242	ρ	0.9787	0.0193	a	3.4405		
	(2.88E-03)					(1.08E-03)		
q	2.5728	a_0, μ_0	0.2595	-0.4678	λ	4.8724		1.4048
	(8.47E-03)					(1.66E-03)		(3.23E-04)
a	2.1763	λ_0, σ_0^2	0.0477	0.2744	μ		-0.4975	
	(3.35E-03)						(1.33E-04)	
b	0.9095	a_1, λ_1	3.9400	-2.2357	σ^2		0.3346	
	(1.13E-03)						(1.11E-04)	
		λ_1, σ_1	5.6647	1.5892				
Mean*	0.7091					0.7061	0.7188	0.7119
	(9.51E-05)					(8.74E-05)	(1.03E-04)	(1.64E-04)
Gini**	0.2982					0.2934	0.3175	0.5000
	(5.55E-05)					(4.29E-05)	(4.97E-05)	(1.81E-17)
Theil	0.1476					0.1383	0.1673	0.4228
	(6.58E-05)					(4.14E-05)	(5.54E-05)	(5.44E-18)
Pietra	0.2113					0.2099	0.2276	0.3679
	(3.88E-05)					(3.15E-05)	(3.66E-05)	(1.15E-12)
-lnL	4.38E+05		3.04E+05	1.95E+05		5.37E+05	9.44E+05	6.16E+06

Sample size: 19,127,600 *Census estimate: 0.708 **Census estimate: 0.3090

Table C.8: ESTIMATED DISTRIBUTION FUNCTIONS 2003 INCOME

	GB2		G-G	LN-LN		G	LN	Expo
p	1.1788	ρ	0.9685	0.0247	a	3.5412		
	(2.37E-03)					(1.11E-03)		
q	1.7206	a_0, μ_0	0.3550	-0.3998	λ	4.6547		1.2965
	(4.48E-03)					(1.57E-03)		(2.97E-04)
a	2.5421	λ_0, σ_0^2	0.0522	0.2625	μ		-0.4180	
	(3.59E-03)						(1.29E-04)	
b	0.8139	a_1, λ_1	4.2383	-1.2047	σ^2		0.3228	
	(6.09E-04)						(1.06E-04)	
		λ_1, σ_1	5.7389	4.2265				
Mean*	0.7681					0.7608	0.7737	0.7713
	(1.08E-04)					(9.21E-05)	(1.08E-04)	(1.76E-04)
Gini**	0.2995					0.2894	0.3121	0.5000
	(6.26E-05)					(4.22E-05)	(4.87E-05)	(5.45E-22)
Theil	0.1525					0.1346	0.1614	0.4228
	(8.24E-05)					(4.02E-05)	(5.32E-05)	(1.76E-18)
Pietra	0.2114					0.2071	0.2236	0.3679
	(4.32E-05)					(3.10E-05)	(3.59E-05)	(9.84E-14)
-lnL	5.38E+05		3.97E+05	2.56E+05		7.34E+05	1.01E+06	6.64E+06

Sample size: 19,518,900 *Census estimate: 0.769 **Census estimate: 0.3060

Table C.9: ESTIMATED DISTRIBUTION FUNCTIONS 2005 INCOME

	GB2		G-G	LN-LN		G	LN	Expo
p	1.0460	ρ	0.9554	0.0303	a	3.4419		
	(2.06E-03)					(1.08E-03)		
q	1.4268	a_0, μ_0	0.4389	-0.3229	λ	4.1613		1.1839
	(3.53E-03)					(1.41E-03)		(2.70E-04)
a	2.7440	λ_0, σ_0^2	0.0674	0.2677	μ		-0.3376	
	(3.86E-03)						(1.31E-04)	
b	0.8464	a_1, λ_1	4.2726	-0.7145	σ^2		0.3370	
	(5.49E-04)						(1.11E-04)	
		λ_1, σ_1	5.3916	4.9610				
Mean*	0.8402					0.8271	0.8444	0.8446
	(1.26E-04)					(1.01E-04)	(1.20E-04)	(1.93E-04)
Gini**	0.3076					0.2933	0.3186	0.5000
	(6.97E-05)					(4.26E-05)	(4.96E-05)	(1.25E-18)
Theil	0.1634					0.1383	0.1685	0.4228
	(1.00E-04)					(4.11E-05)	(5.55E-05)	(3.30E-19)
Pietra	0.2168					0.2099	0.2284	0.3679
	(4.79E-05)					(3.13E-05)	(3.66E-05)	(1.17E-12)
-lnL	5.18E+05		3.62E+05	2.34E+05		7.36E+05	1.03E+06	6.45E+06

Sample size: 19,857,100 *Census estimate: 0.843 **Census estimate: 0.3140

Table C.10: ESTIMATED DISTRIBUTION FUNCTIONS 2007 INCOME

	GB2		G-G	LN-LN		G	LN	Expo
p	1.1054	ρ	0.9439	0.0232	a	3.2408		
	(2.41E-03)					(1.01E-03)		
q	1.4346	a_0, μ_0	0.4584	-0.2009	λ	3.4336		1.0215
	(4.05E-03)					(1.17E-03)		(2.32E-04)
a	2.5553	λ_0, σ_0^2	0.0315	0.3061	μ		-0.2129	
	(4.01E-03)						(1.34E-04)	
b	0.9418	a_1, λ_1	4.0325	-0.6280	σ^2		0.3667	
	(7.06E-04)						(1.21E-04)	
		λ_1, σ_1	4.5420	6.0217				
Mean*	0.9694					0.9438	0.9708	0.9789
	(1.65E-04)					(1.18E-04)	(1.44E-04)	(2.23E-04)
Gini**	0.3237					0.3016	0.3315	0.5000
	(8.36E-05)					(4.34E-05)	(5.14E-05)	(2.94E-18)
Theil	0.1831					0.1464	0.1834	0.4228
	(1.36E-04)					(4.32E-05)	(6.05E-05)	(1.30E-17)
Pietra	0.2285					0.2160	0.2379	0.3679
	(5.70E-05)					(3.19E-05)	(3.81E-05)	(6.97E-14)
-lnL	4.11E+05		2.72E+05	1.89E+05		5.87E+05	8.35E+05	5.91E+06

Sample size: 20,570,500   *Census estimate: 0.982   **Census estimate: 0.3360

Table C.11: ESTIMATED DISTRIBUTION FUNCTIONS 2009 INCOME

	GB2		G-G	LN-LN		G	LN	Expo
p	1.0414	ρ	0.9425	0.0311	a	3.2114		
	(2.04E-03)					(9.77E-04)		
q	1.4016	a_0, μ_0	0.4682	-0.1960	λ	3.4089		1.0250
	(3.71E-03)					(1.14E-03)		(2.28E-04)
a	2.6232	λ_0, σ_0^2	0.0508	0.2980	μ		-0.2163	
	(3.79E-03)						(1.33E-04)	
b	0.9567	a_1, λ_1	4.0135	-0.8719	σ^2		0.3734	
	(7.09E-04)						(1.20E-04)	
		λ_1, σ_1	4.5034	3.8558				
Mean*	0.9663					0.9421	0.9709	0.9756
	(1.60E-04)					(1.15E-04)	(1.43E-04)	(2.17E-04)
Gini**	0.3236					0.3029	0.3343	0.5000
	(8.14E-05)					(4.25E-05)	(5.06E-05)	(8.36E-22)
Theil	0.1827					0.1477	0.1867	0.4228
	(1.31E-04)					(4.25E-05)	(6.02E-05)	(8.07E-18)
Pietra	0.2284					0.2169	0.2400	0.3679
	(5.53E-05)					(3.13E-05)	(3.75E-05)	(1.20E-13)
-lnL	4.70E+05		3.17E+05	2.01E+05		6.42E+05	1.02E+06	6.15E+06

Sample size: 21,500,300 *Census estimate: 0.968 **Census estimate: 0.3290

Table C.12: ESTIMATED DISTRIBUTION FUNCTIONS 2011 INCOME

	GB2		G-G	LN-LN		G	LN	Expo
p	0.9487	ρ	0.9338	0.0299	a	3.3617		
	(1.88E-03)					(1.01E-03)		
q	1.4885	a_0, μ_0	0.6497	-0.1501	λ	3.4753		0.9986
	(4.23E-03)					(1.14E-03)		(2.19E-04)
a	2.7604	λ_0, σ_0^2	0.1980	0.2824	μ		-0.1825	
	(4.13E-03)						(1.28E-04)	
b	1.0683	a_1, λ_1	4.1958	-1.3788	σ^2		0.3606	
	(8.59E-04)						(1.15E-04)	
		λ_1, σ_1	4.5314	1.9377				
Mean*	0.9850					0.9673	0.9978	1.0014
	(1.46E-04)					(1.14E-04)	(1.42E-04)	(2.19E-04)
Gini**	0.3101					0.2965	0.3289	0.5000
	(7.15E-05)					(4.12E-05)	(4.92E-05)	(9.73E-20)
Theil	0.1645					0.1414	0.1803	0.4228
	(1.04E-04)					(4.03E-05)	(5.73E-05)	(2.48E-17)
Pietra	0.2185					0.2123	0.2360	0.3679
	(4.77E-05)					(3.03E-05)	(3.65E-05)	(1.15E-07)
-lnL	4.06E+05		2.61E+05	1.73E+05		5.34E+05	1.11E+06	6.57E+06

Sample size: 22,101,700 *Census estimate: 0.994 **Census estimate: 0.3200

Table C.13: ESTIMATED DISTRIBUTION FUNCTIONS 2013 INCOME

	GB2		G-G	LN-LN		G	LN	Expo
p	0.8124	ρ	0.9266	0.0421	a	3.3347		
	(1.46E-03)					(9.96E-04)		
q	1.0014	a_0, μ_0	0.5553	-0.1490	λ	3.3848		0.9742
	(2.28E-03)					(1.11E-03)		(2.12E-04)
a	3.2551	λ_0, σ_0^2	0.0878	0.2758	μ		-0.1644	
	(4.47E-03)						(1.28E-04)	
b	0.9495	a_1, λ_1	4.3714	-0.3554	σ^2		0.3636	
	(4.94E-04)						(1.15E-04)	
		λ_1, σ_1	4.7285	4.2966				
Mean*	1.0190					0.9852	1.0175	1.0265
	(1.75E-04)					(1.16E-04)	(1.44E-04)	(2.23E-04)
Gini**	0.3231					0.2976	0.3302	0.5000
	(8.72E-05)					(4.12E-05)	(4.92E-05)	(2.83E-19)
Theil	0.1876					0.1425	0.1818	0.4228
	(1.49E-04)					(4.04E-05)	(5.76E-05)	(7.00E-18)
Pietra	0.2270					0.2131	0.2370	0.3679
	(5.96E-05)					(3.03E-05)	(3.64E-05)	(7.12E-13)
-lnL	3.95E+05		2.11E+05	1.14E+05		6.33E+05	1.12E+06	6.66E+06

Sample size: 22,592,900 *Census estimate: 1.029 **Census estimate: 0.3330

Table C.14: ESTIMATED DISTRIBUTION FUNCTIONS 2015 INCOME

	GB2		G-G	LN-LN		G	LN	Expo
p	0.9314	ρ	0.9506	0.0228	a	3.3445		
	(1.64E-03)					(9.82E-04)		
q	1.3289	a_0, μ_0	0.3425	-0.1446	λ	3.4212		0.9865
	(3.36E-03)					(1.10E-03)		(2.11E-04)
a	2.8519	λ_0, σ_0^2	0.0192	0.2847	μ		-0.1722	
	(3.84E-03)						(1.26E-04)	
b	1.0230	a_1, λ_1	4.1673	-1.6382	σ^2		0.3647	
	(7.27E-04)						(1.13E-04)	
		λ_1, σ_1	4.4494	3.3199				
Mean*	0.9997					0.9776	1.0102	1.0137
	(1.53E-04)					(1.13E-04)	(1.41E-04)	(2.17E-04)
Gini**	0.3146					0.2972	0.3306	0.5000
	(7.63E-05)					(4.04E-05)	(4.83E-05)	(0.00E+00)
Theil	0.1716					0.1421	0.1824	0.4228
	(1.18E-04)					(3.96E-05)	(5.66E-05)	(2.19E-18)
Pietra	0.2215					0.2128	0.2373	0.3679
	(5.12E-05)					(2.97E-05)	(3.57E-05)	(7.28E-13)
-lnL	5.59E+05		3.15E+05	1.73E+05		7.26E+05	1.39E+06	6.99E+06

Sample size: 23,224,200 *Census estimate: 1.009 **Census estimate: 0.3230

Appendix D

Tables: Estimation Outputs Part II

Table D.1: ESTIMATED DISTRIBUTION FUNCTIONS 1994 INCOME

	QME			IME			QIME		
	GB2	G	LN	GB2	G	LN	GB2	G	LN
p, μ	89.2642		-0.6188	14.0633		-0.6336	1.3474		-0.6370
	(3.74E+01)		(1.31E-04)	(6.75E-01)		(1.34E-04)	(4.41E-03)		(1.36E-04)
q, σ^2	87.8264		0.2849	106.7812		0.3105	2.4829		0.3196
	(3.65E+01)		(1.15E-04)	(1.71E+01)		(1.07E-04)	(1.04E-02)		(1.10E-04)
a	0.2827	3.8105		0.5117	3.4520		2.1604	3.3396	
	(5.93E-02)	(1.47E-03)		(1.59E-02)	(1.14E-03)		(4.80E-03)	(1.10E-03)	
b, λ	0.5085	6.2649		29.6124	5.5914		0.7673	5.4092	
	(2.89E-02)	(2.59E-03)		(1.01E+01)	(1.99E-03)		(9.55E-04)	(1.91E-03)	
Mean*	0.6211	0.6082	0.6211	0.6176	0.6174	0.6198	0.6174	0.6174	0.6205
	(1.35E-03)	(2.72E-05)	(3.14E-04)	(1.50E-02)	(3.31E-05)	(3.14E-04)	(5.15E-04)	(3.48E-05)	(3.09E-04)
Gini**	0.2939	0.2797	0.2941	0.3004	0.2929	0.3064	0.3012	0.2974	0.3107
	(9.99E-04)	(4.39E-06)	(2.45E-04)	(1.81E-03)	(5.45E-06)	(2.50E-04)	(2.97E-05)	(5.82E-06)	(2.41E-04)
Theil	0.1423	0.1255	0.1424	0.1467	0.1379	0.1552	0.1511	0.1423	0.1598
	(1.02E-03)	(4.02E-06)	(2.49E-04)	(1.84E-03)	(5.25E-06)	(2.67E-04)	(3.05E-05)	(5.70E-06)	(2.61E-04)
Pietra	0.2102	0.2000	0.2104	0.2149	0.2096	0.2195	0.2134	0.2129	0.2226
	(7.32E-04)	(3.21E-06)	(1.80E-04)	(1.33E-03)	(4.00E-06)	(1.84E-04)	(2.16E-05)	(4.28E-06)	(1.77E-04)

Sample size: 17,437,600 *Census estimate: 0.617 **Census estimate: 0.3020

Table D.2: ESTIMATED DISTRIBUTION FUNCTIONS 1995 INCOME

	QME			IME			QIME		
	GB2	G	LN	GB2	G	LN	GB2	G	LN
p, μ	46.5979		-0.6255	11.7744		-0.6391	2.4982		-0.6414
	(1.24E+01)		(1.29E-04)	(4.86E-01)		(1.30E-04)	(1.40E-02)		(1.31E-04)
q, σ^2	123.8030		0.2824	145.6171		0.2980	6.2966		0.3049
	(5.56E+01)		(1.13E-04)	(2.97E+01)		(1.02E-04)	(5.55E-02)		(1.04E-04)
a	0.3247	3.8432		0.5589	3.5888		1.4416	3.5108	
	(5.17E-02)	(1.47E-03)		(1.52E-02)	(1.18E-03)		(5.27E-03)	(1.14E-03)	
b, λ	11.0499	6.3688		50.9605	5.8835		1.0926	5.7555	
	(1.15E+01)	(2.61E-03)		(2.11E+01)	(2.07E-03)		(5.18E-03)	(2.01E-03)	
Mean*	0.6138	0.6034	0.6161	0.6100	0.6100	0.6126	0.6100	0.6100	0.6133
	(2.15E-02)	(2.63E-05)	(3.10E-04)	(1.64E-02)	(3.03E-05)	(3.16E-04)	(2.86E-04)	(3.14E-05)	(3.12E-04)
Gini**	0.2901	0.2786	0.2929	0.2936	0.2877	0.3005	0.2948	0.2906	0.3038
	(3.50E-03)	(4.28E-06)	(2.47E-04)	(1.70E-03)	(5.00E-06)	(2.65E-04)	(5.66E-05)	(5.23E-06)	(2.57E-04)
Theil	0.1376	0.1245	0.1412	0.1396	0.1329	0.1490	0.1419	0.1357	0.1525
	(3.46E-03)	(3.90E-06)	(2.49E-04)	(1.68E-03)	(4.73E-06)	(2.76E-04)	(5.65E-05)	(5.00E-06)	(2.72E-04)
Pietra	0.2074	0.1991	0.2095	0.2100	0.2058	0.2151	0.2100	0.2079	0.2175
	(2.56E-03)	(3.13E-06)	(1.81E-04)	(1.25E-03)	(3.67E-06)	(1.94E-04)	(4.13E-05)	(3.84E-06)	(1.89E-04)

Sample size: 17,731,800 *Census estimate: 0.610 **Census estimate: 0.2960

Table D.3: ESTIMATED DISTRIBUTION FUNCTIONS 1996 INCOME

	QME			IME			QIME		
	GB2	G	LN	GB2	G	LN	GB2	G	LN
p, μ	62.0854		-0.5924	16.2724		-0.6044	2.2774		-0.6066
	(2.11E+01)		(1.26E-04)	(8.91E-01)		(1.27E-04)	(1.13E-02)		(1.28E-04)
q, σ^2	131.4453		0.2723	129.1664		0.2882	4.6489		0.2946
	(6.60E+01)		(1.08E-04)	(2.44E+01)		(9.82E-05)	(3.18E-02)		(9.96E-05)
a	0.2962	3.9736		0.4922	3.7034		1.6065	3.6225	
	(5.83E-02)	(1.51E-03)		(1.74E-02)	(1.21E-03)		(5.10E-03)	(1.18E-03)	
b, λ	7.0514	6.3962		38.8362	5.8880		0.9163	5.7591	
	(7.39E+00)	(2.60E-03)		(1.64E+01)	(2.06E-03)		(2.57E-03)	(2.00E-03)	
Mean*	0.6320	0.6212	0.6337	0.6292	0.6290	0.6311	0.6290	0.6290	0.6318
	(1.90E-02)	(2.59E-05)	(3.24E-04)	(1.92E-02)	(3.01E-05)	(3.31E-04)	(3.71E-04)	(3.12E-05)	(3.27E-04)
Gini**	0.2858	0.2743	0.2878	0.2904	0.2835	0.2958	0.2910	0.2864	0.2989
	(3.57E-03)	(3.98E-06)	(2.59E-04)	(2.07E-03)	(4.68E-06)	(2.77E-04)	(4.11E-05)	(4.89E-06)	(2.70E-04)
Theil	0.1336	0.1206	0.1361	0.1368	0.1290	0.1441	0.1390	0.1317	0.1473
	(3.48E-03)	(3.57E-06)	(2.57E-04)	(2.02E-03)	(4.35E-06)	(2.84E-04)	(4.07E-05)	(4.60E-06)	(2.80E-04)
Pietra	0.2043	0.1960	0.2058	0.2076	0.2027	0.2116	0.2069	0.2049	0.2139
	(2.61E-03)	(2.91E-06)	(1.90E-04)	(1.51E-03)	(3.42E-06)	(2.03E-04)	(2.99E-05)	(3.58E-06)	(1.98E-04)

Sample size: 17,950,400 *Census estimate: 0.629 **Census estimate: 0.2920

Table D.4: ESTIMATED DISTRIBUTION FUNCTIONS 1997 INCOME

	QME			IME			QIME		
	GB2	G	LN	GB2	G	LN	GB2	G	LN
p, μ	90.9109		-0.5760	15.8119		-0.5866	1.4779		-0.5896
	(3.80E+01)		(1.28E-04)	(8.31E-01)		(1.32E-04)	(5.04E-03)		(1.33E-04)
q, σ^2	88.1249		0.2847	113.0951		0.3121	2.6375		0.3192
	(3.62E+01)		(1.13E-04)	(1.90E+01)		(1.06E-04)	(1.13E-02)		(1.08E-04)
a	0.2814	3.8136		0.4836	3.4497		2.0516	3.3342	
	(5.85E-02)	(1.44E-03)		(1.64E-02)	(1.12E-03)		(4.67E-03)	(1.07E-03)	
b, λ	0.5032	6.0076		34.3107	5.3386		0.7959	5.1597	
	(3.46E-02)	(2.44E-03)		(1.29E+01)	(1.86E-03)		(1.03E-03)	(1.79E-03)	
Mean*	0.6483	0.6348	0.6481	0.6462	0.6462	0.6501	0.6462	0.6462	0.6505
	(1.55E-03)	(2.84E-05)	(3.21E-04)	(1.82E-02)	(3.48E-05)	(3.22E-04)	(4.89E-04)	(3.67E-05)	(3.18E-04)
Gini**	0.2937	0.2796	0.2940	0.3009	0.2930	0.3072	0.3023	0.2977	0.3105
	(9.99E-04)	(4.30E-06)	(2.41E-04)	(2.05E-03)	(5.35E-06)	(2.44E-04)	(2.93E-05)	(5.73E-06)	(2.37E-04)
Theil	0.1422	0.1254	0.1423	0.1474	0.1380	0.1560	0.1522	0.1425	0.1596
	(1.02E-03)	(3.94E-06)	(2.45E-04)	(2.09E-03)	(5.16E-06)	(2.61E-04)	(3.04E-05)	(5.62E-06)	(2.57E-04)
Pietra	0.2101	0.1999	0.2104	0.2153	0.2097	0.2200	0.2143	0.2131	0.2224
	(7.32E-04)	(3.14E-06)	(1.77E-04)	(1.50E-03)	(3.93E-06)	(1.79E-04)	(2.13E-05)	(4.21E-06)	(1.74E-04)

Sample size: 18,106,700 *Census estimate: 0.646 **Census estimate: 0.3030

Table D.5: ESTIMATED DISTRIBUTION FUNCTIONS 1999 INCOME

	QME			IME			QIME		
	GB2	G	LN	GB2	G	LN	GB2	G	LN
p, μ	29.3774		-0.5406	13.4001		-0.5534	1.3689		-0.5572
	(5.25E+00)		(1.30E-04)	(5.94E-01)		(1.33E-04)	(4.40E-03)		(1.35E-04)
q, σ^2	144.7326		0.2992	98.5713		0.3273	2.4950		0.3362
	(6.45E+01)		(1.17E-04)	(1.42E+01)		(1.10E-04)	(1.02E-02)		(1.12E-04)
a	0.3722	3.6453		0.5117	3.2866		2.0868	3.1795	
	(4.19E-02)	(1.36E-03)		(1.47E-02)	(1.06E-03)		(4.53E-03)	(1.01E-03)	
b, λ	43.6802	5.5125		30.2424	4.8736		0.8351	4.7145	
	(5.25E+01)	(2.22E-03)		(9.30E+00)	(1.69E-03)		(1.04E-03)	(1.62E-03)	
Mean*	0.6717	0.6613	0.6764	0.6746	0.6744	0.6772	0.6744	0.6744	0.6776
	(3.52E-02)	(3.21E-05)	(3.22E-04)	(1.50E-02)	(3.97E-05)	(3.22E-04)	(4.62E-04)	(4.18E-05)	(3.18E-04)
Gini**	0.2961	0.2856	0.3011	0.3077	0.2996	0.3142	0.3090	0.3043	0.3182
	(3.61E-03)	(4.65E-06)	(2.20E-04)	(1.72E-03)	(5.82E-06)	(2.23E-04)	(3.03E-05)	(6.22E-06)	(2.16E-04)
Theil	0.1431	0.1309	0.1496	0.1543	0.1445	0.1637	0.1595	0.1491	0.1681
	(3.64E-03)	(4.36E-06)	(2.30E-04)	(1.80E-03)	(5.76E-06)	(2.46E-04)	(3.21E-05)	(6.25E-06)	(2.41E-04)
Pietra	0.2118	0.2042	0.2155	0.2203	0.2146	0.2252	0.2191	0.2180	0.2281
	(2.64E-03)	(3.40E-06)	(1.62E-04)	(1.27E-03)	(4.28E-06)	(1.65E-04)	(2.20E-05)	(4.58E-06)	(1.59E-04)

Sample size: 18,492,800 *Census estimate: 0.674 **Census estimate: 0.3100

Table D.6: ESTIMATED DISTRIBUTION FUNCTIONS 2000 INCOME

	QME			IME			QIME		
	GB2	G	LN	GB2	G	LN	GB2	G	LN
p, μ	28.0105		-0.5196	14.2509		-0.5312	1.4200		-0.5346
	(4.77E+00)		(1.31E-04)	(6.60E-01)		(1.33E-04)	(4.63E-03)		(1.35E-04)
q, σ^2	145.1934		0.3057	91.5500		0.3307	2.5136		0.3389
	(6.39E+01)		(1.19E-04)	(1.25E+01)		(1.10E-04)	(1.03E-02)		(1.12E-04)
a	0.3754	3.5741		0.4979	3.2555		2.0449	3.1568	
	(4.04E-02)	(1.32E-03)		(1.48E-02)	(1.04E-03)		(4.47E-03)	(9.99E-04)	
b, λ	49.3045	5.2792		26.1579	4.7129		0.8427	4.5698	
	(5.88E+01)	(2.11E-03)		(7.63E+00)	(1.62E-03)		(1.03E-03)	(1.56E-03)	
Mean*	0.6879	0.6770	0.6930	0.6907	0.6908	0.6936	0.6908	0.6908	0.6941
	(3.65E-02)	(3.42E-05)	(3.24E-04)	(1.46E-02)	(4.16E-05)	(3.26E-04)	(4.58E-04)	(4.37E-05)	(3.22E-04)
Gini**	0.2990	0.2882	0.3042	0.3095	0.3010	0.3157	0.3108	0.3053	0.3194
	(3.59E-03)	(4.80E-06)	(2.12E-04)	(1.74E-03)	(5.90E-06)	(2.19E-04)	(3.00E-05)	(6.27E-06)	(2.12E-04)
Theil	0.1460	0.1334	0.1528	0.1563	0.1458	0.1653	0.1616	0.1501	0.1694
	(3.67E-03)	(4.55E-06)	(2.24E-04)	(1.84E-03)	(5.86E-06)	(2.42E-04)	(3.22E-05)	(6.33E-06)	(2.38E-04)
Pietra	0.2140	0.2062	0.2178	0.2216	0.2155	0.2263	0.2204	0.2187	0.2290
	(2.64E-03)	(3.52E-06)	(1.56E-04)	(1.28E-03)	(4.34E-06)	(1.61E-04)	(2.19E-05)	(4.62E-06)	(1.56E-04)

Sample size: 18,700,900 *Census estimate: 0.691 **Census estimate: 0.3110

Table D.7: ESTIMATED DISTRIBUTION FUNCTIONS 2002 INCOME

	QME			IME			QIME		
	GB2	G	LN	GB2	G	LN	GB2	G	LN
p, μ	18.9989		-0.4918	9.3834		-0.5050	1.3454		-0.5049
	(2.25E+00)		(1.29E-04)	(2.91E-01)		(1.32E-04)	(4.21E-03)		(1.33E-04)
q, σ^2	155.7913		0.3049	76.4180		0.3332	2.5162		0.3370
	(6.72E+01)		(1.18E-04)	(8.36E+00)		(1.10E-04)	(1.01E-02)		(1.11E-04)
a	0.4437	3.5827		0.6088	3.2680		2.0997	3.1894	
	(3.42E-02)	(1.31E-03)		(1.24E-02)	(1.03E-03)		(4.47E-03)	(9.98E-04)	
b, λ	73.4680	5.1495		20.4345	4.6134		0.8905	4.5022	
	(7.87E+01)	(2.04E-03)		(4.06E+00)	(1.57E-03)		(1.11E-03)	(1.52E-03)	
Mean*	0.7052	0.6957	0.7123	0.7084	0.7084	0.7129	0.7084	0.7084	0.7143
	(3.13E-02)	(3.51E-05)	(3.29E-04)	(8.48E-03)	(4.26E-05)	(3.30E-04)	(4.30E-04)	(4.43E-05)	(3.29E-04)
Gini**	0.2971	0.2879	0.3038	0.3071	0.3004	0.3168	0.3083	0.3038	0.3185
	(2.72E-03)	(4.73E-06)	(2.10E-04)	(1.05E-03)	(5.79E-06)	(2.14E-04)	(3.02E-05)	(6.08E-06)	(2.11E-04)
Theil	0.1435	0.1331	0.1524	0.1533	0.1453	0.1666	0.1586	0.1487	0.1685
	(2.74E-03)	(4.47E-06)	(2.22E-04)	(1.09E-03)	(5.74E-06)	(2.37E-04)	(3.18E-05)	(6.10E-06)	(2.37E-04)
Pietra	0.2125	0.2059	0.2175	0.2199	0.2151	0.2271	0.2186	0.2176	0.2284
	(1.99E-03)	(3.46E-06)	(1.55E-04)	(7.71E-04)	(4.26E-06)	(1.57E-04)	(2.19E-05)	(4.48E-06)	(1.56E-04)

Sample size: 19,127,600 *Census estimate: 0.708 **Census estimate: 0.3090

Table D.8: ESTIMATED DISTRIBUTION FUNCTIONS 2003 INCOME

	QME			IME			QIME		
	GB2	G	LN	GB2	G	LN	GB2	G	LN
p, μ	31.3041		-0.4143	7.1181		-0.4161	1.7106		-0.4142
	(5.74E+00)		(1.24E-04)	(1.30E-01)		(1.27E-04)	(6.01E-03)		(1.26E-04)
q, σ^2	141.1047		0.2892	11.4277		0.3120	2.3864		0.3079
	(6.08E+01)		(1.10E-04)	(2.55E-01)		(1.02E-04)	(9.34E-03)		(9.99E-05)
a	0.3691	3.7577		0.8760	3.4351		2.0041	3.3617	
	(4.24E-02)	(1.37E-03)		(9.10E-03)	(1.08E-03)		(4.46E-03)	(1.04E-03)	
b, λ	40.2805	5.0291		1.1675	4.4639		0.8130	4.3682	
	(4.59E+01)	(1.97E-03)		(1.25E-02)	(1.50E-03)		(7.55E-04)	(1.46E-03)	
Mean*	0.7588	0.7472	0.7636	0.7694	0.7695	0.7710	0.7689	0.7696	0.7709
	(3.63E-02)	(3.64E-05)	(3.61E-04)	(3.06E-04)	(4.48E-05)	(3.68E-04)	(5.33E-04)	(4.64E-05)	(3.72E-04)
Gini**	0.2920	0.2816	0.2962	0.3040	0.2936	0.3071	0.3046	0.2965	0.3052
	(3.42E-03)	(4.26E-06)	(2.26E-04)	(1.38E-04)	(5.20E-06)	(2.35E-04)	(2.56E-05)	(5.43E-06)	(2.41E-04)
Theil	0.1391	0.1272	0.1446	0.1524	0.1386	0.1560	0.1566	0.1414	0.1540
	(3.40E-03)	(3.93E-06)	(2.32E-04)	(1.45E-04)	(5.02E-06)	(2.51E-04)	(2.75E-05)	(5.30E-06)	(2.56E-04)
Pietra	0.2088	0.2013	0.2120	0.2172	0.2101	0.2200	0.2159	0.2123	0.2186
	(2.50E-03)	(3.12E-06)	(1.66E-04)	(1.01E-04)	(3.82E-06)	(1.73E-04)	(1.86E-05)	(3.99E-06)	(1.77E-04)

Sample size: 19,518,900 *Census estimate: 0.769 **Census estimate: 0.3060

Table D.9: ESTIMATED DISTRIBUTION FUNCTIONS 2005 INCOME

	QME			IME			QIME		
	GB2	G	LN	GB2	G	LN	GB2	G	LN
p, μ	13.2154		-0.3332	2.9633		-0.3330	1.6695		-0.3354
	(1.01E+00)		(1.26E-04)	(2.15E-02)		(1.29E-04)	(5.64E-03)		(1.29E-04)
q, σ^2	58.2588		0.3025	3.7811		0.3273	2.1279		0.3296
	(1.04E+01)		(1.15E-04)	(2.89E-02)		(1.06E-04)	(7.65E-03)		(1.06E-04)
a	0.5599	3.6062		1.4533	3.2878		2.0484	3.2222	
	(2.70E-02)	(1.30E-03)		(6.11E-03)	(1.02E-03)		(4.35E-03)	(9.90E-04)	
b, λ	10.6225	4.4283		0.8694	3.8999		0.8328	3.8214	
	(3.30E+00)	(1.72E-03)		(1.37E-03)	(1.30E-03)		(6.60E-04)	(1.27E-03)	
Mean*	0.8255	0.8144	0.8336	0.8429	0.8431	0.8442	0.8413	0.8432	0.8432
	(6.89E-03)	(4.45E-05)	(3.80E-04)	(4.71E-04)	(5.57E-05)	(3.88E-04)	(5.52E-04)	(5.75E-05)	(3.86E-04)
Gini**	0.2963	0.2870	0.3026	0.3122	0.2996	0.3142	0.3116	0.3024	0.3152
	(9.53E-04)	(4.58E-06)	(2.09E-04)	(4.32E-05)	(5.61E-06)	(2.16E-04)	(2.49E-05)	(5.85E-06)	(2.15E-04)
Theil	0.1429	0.1323	0.1512	0.1638	0.1444	0.1636	0.1657	0.1472	0.1648
	(9.58E-04)	(4.32E-06)	(2.20E-04)	(4.80E-05)	(5.55E-06)	(2.37E-04)	(2.80E-05)	(5.84E-06)	(2.37E-04)
Pietra	0.2119	0.2053	0.2167	0.2223	0.2145	0.2252	0.2208	0.2166	0.2259
	(6.98E-04)	(3.36E-06)	(1.54E-04)	(3.16E-05)	(4.13E-06)	(1.59E-04)	(1.81E-05)	(4.30E-06)	(1.58E-04)

Sample size: 19,857,100 *Census estimate: 0.843 **Census estimate: 0.3140

Table D.10: ESTIMATED DISTRIBUTION FUNCTIONS 2007 INCOME

	QME			IME			QIME		
	GB2	G	LN	GB2	G	LN	GB2	G	LN
p, μ	11.3095		-0.2102	1.3939		-0.2047	1.1597		-0.2080
	(7.78E-01)		(1.29E-04)	(5.03E-03)		(1.35E-04)	(3.06E-03)		(1.35E-04)
q, σ^2	116.8032		0.3299	1.6229		0.3711	1.4168		0.3738
	(3.55E+01)		(1.23E-04)	(5.77E-03)		(1.18E-04)	(3.90E-03)		(1.18E-04)
a	0.5493	3.3291		2.2156	2.9328		2.4866	2.8381	
	(2.50E-02)	(1.17E-03)		(5.07E-03)	(8.91E-04)		(4.38E-03)	(8.53E-04)	
b, λ	60.6790	3.5756		0.8923	2.9884		0.9103	2.8908	
	(3.77E+01)	(1.37E-03)		(5.83E-04)	(9.86E-04)		(5.21E-04)	(9.48E-04)	
Mean*	0.9423	0.9311	0.9558	0.9808	0.9814	0.9810	0.9745	0.9818	0.9791
	(2.03E-02)	(6.30E-05)	(4.06E-04)	(5.71E-04)	(8.65E-05)	(4.09E-04)	(5.85E-04)	(9.11E-05)	(4.07E-04)
Gini**	0.3061	0.2979	0.3154	0.3337	0.3158	0.3333	0.3298	0.3205	0.3345
	(1.50E-03)	(5.26E-06)	(1.79E-04)	(2.53E-05)	(6.90E-06)	(1.73E-04)	(2.49E-05)	(7.36E-06)	(1.73E-04)
Theil	0.1523	0.1427	0.1650	0.1958	0.1609	0.1855	0.1917	0.1660	0.1869
	(1.56E-03)	(5.17E-06)	(1.98E-04)	(3.28E-05)	(7.22E-06)	(2.05E-04)	(3.21E-05)	(7.84E-06)	(2.05E-04)
Pietra	0.2192	0.2133	0.2260	0.2364	0.2265	0.2393	0.2331	0.2300	0.2402
	(1.11E-03)	(3.87E-06)	(1.32E-04)	(1.84E-05)	(5.10E-06)	(1.28E-04)	(1.80E-05)	(5.45E-06)	(1.28E-04)

Sample size: 20,570,500 *Census estimate: 0.982 **Census estimate: 0.3360

Table D.11: ESTIMATED DISTRIBUTION FUNCTIONS 2009 INCOME

	QME			IME			QIME		
	GB2	G	LN	GB2	G	LN	GB2	G	LN
p, μ	34.3662		-0.2115	2.6190		-0.2120	1.2928		-0.2153
	(6.51E+00)		(1.26E-04)	(1.64E-02)		(1.30E-04)	(3.56E-03)		(1.31E-04)
q, σ^2	137.3288		0.3257	3.4775		0.3617	1.7010		0.3664
	(5.63E+01)		(1.19E-04)	(2.34E-02)		(1.13E-04)	(5.05E-03)		(1.13E-04)
a	0.3354	3.3694		1.4688	3.0008		2.2640	2.9168	
	(3.94E-02)	(1.16E-03)		(5.44E-03)	(8.92E-04)		(4.11E-03)	(8.58E-04)	
b, λ	51.7974	3.6272		1.0141	3.1014		0.9524	3.0139	
	(5.93E+01)	(1.35E-03)		(1.54E-03)	(9.99E-04)		(6.32E-04)	(9.65E-04)	
Mean*	0.9466	0.9289	0.9525	0.9673	0.9676	0.9693	0.9658	0.9678	0.9684
	(5.00E-02)	(6.03E-05)	(3.99E-04)	(3.82E-04)	(7.95E-05)	(4.02E-04)	(5.02E-04)	(8.31E-05)	(3.99E-04)
Gini**	0.3093	0.2962	0.3135	0.3266	0.3125	0.3294	0.3272	0.3166	0.3314
	(3.96E-03)	(5.03E-06)	(1.79E-04)	(4.25E-05)	(6.45E-06)	(1.77E-04)	(2.44E-05)	(6.83E-06)	(1.74E-04)
Theil	0.1569	0.1411	0.1628	0.1804	0.1575	0.1809	0.1856	0.1617	0.1832
	(4.22E-03)	(4.91E-06)	(1.96E-04)	(5.00E-05)	(6.68E-06)	(2.06E-04)	(2.94E-05)	(7.17E-06)	(2.05E-04)
Pietra	0.2215	0.2120	0.2246	0.2328	0.2240	0.2364	0.2316	0.2270	0.2378
	(2.92E-03)	(3.69E-06)	(1.32E-04)	(3.12E-05)	(4.76E-06)	(1.31E-04)	(1.77E-05)	(5.05E-06)	(1.29E-04)

Sample size: 21,500,300 *Census estimate: 0.968 **Census estimate: 0.3290

Table D.12: ESTIMATED DISTRIBUTION FUNCTIONS 2011 INCOME

	QME			IME			QIME		
	GB2	G	LN	GB2	G	LN	GB2	G	LN
p, μ	20.7609		-0.1731	1.6731		-0.1759	1.0811		-0.1787
	(2.48E+00)		(1.21E-04)	(6.99E-03)		(1.25E-04)	(2.66E-03)		(1.26E-04)
q, σ^2	151.3382		0.3087	2.3301		0.3442	1.4997		0.3488
	(6.01E+01)		(1.11E-04)	(1.06E-02)		(1.06E-04)	(4.06E-03)		(1.06E-04)
a	0.4242	3.5378		1.9409	3.1325		2.5721	3.0487	
	(3.27E-02)	(1.20E-03)		(5.13E-03)	(9.20E-04)		(4.34E-03)	(8.86E-04)	
b, λ	94.9863	3.6918		1.0421	3.1518		1.0058	3.0671	
	(9.65E+01)	(1.36E-03)		(1.01E-03)	(1.00E-03)		(5.86E-04)	(9.67E-04)	
Mean*	0.9721	0.9583	0.9814	0.9938	0.9939	0.9962	0.9923	0.9940	0.9958
	(4.24E-02)	(5.86E-05)	(4.19E-04)	(4.11E-04)	(7.74E-05)	(4.20E-04)	(5.03E-04)	(8.07E-05)	(4.17E-04)
Gini**	0.2994	0.2896	0.3056	0.3182	0.3064	0.3217	0.3183	0.3102	0.3238
	(2.74E-03)	(4.51E-06)	(1.92E-04)	(2.63E-05)	(5.85E-06)	(1.89E-04)	(2.38E-05)	(6.18E-06)	(1.86E-04)
Theil	0.1460	0.1347	0.1544	0.1720	0.1512	0.1721	0.1753	0.1551	0.1744
	(2.79E-03)	(4.29E-06)	(2.04E-04)	(2.99E-05)	(5.93E-06)	(2.14E-04)	(2.72E-05)	(6.34E-06)	(2.13E-04)
Pietra	0.2142	0.2072	0.2189	0.2258	0.2195	0.2307	0.2247	0.2223	0.2322
	(2.01E-03)	(3.30E-06)	(1.41E-04)	(1.91E-05)	(4.31E-06)	(1.40E-04)	(1.71E-05)	(4.56E-06)	(1.37E-04)

Sample size: 22,101,700 *Census estimate: 0.994 **Census estimate: 0.3200

Table D.13: ESTIMATED DISTRIBUTION FUNCTIONS 2013 INCOME

	QME			IME			QIME		
	GB2	G	LN	GB2	G	LN	GB2	G	LN
p, μ	91.7820		-0.1543	1.6089		-0.1571	0.9111		-0.1577
	(3.39E+01)		(1.18E-04)	(6.07E-03)		(1.28E-04)	(2.06E-03)		(1.27E-04)
q, σ^2	81.4726		0.3043	1.6583		0.3684	1.0462		0.3630
	(2.83E+01)		(1.08E-04)	(5.81E-03)		(1.12E-04)	(2.43E-03)		(1.10E-04)
a	0.2764	3.5796		2.1498	3.0565		3.0304	2.9006	
	(4.96E-02)	(1.21E-03)		(4.91E-03)	(8.87E-04)		(4.80E-03)	(8.32E-04)	
b, λ	0.5557	3.6685		0.8712	2.9721		0.9165	2.8189	
	(8.99E-02)	(1.33E-03)		(5.79E-04)	(9.34E-04)		(3.96E-04)	(8.81E-04)	
Mean*	0.9989	0.9758	0.9978	1.0274	1.0284	1.0275	1.0252	1.0290	1.0241
	(3.66E-03)	(5.89E-05)	(4.25E-04)	(5.92E-04)	(8.50E-05)	(4.10E-04)	(6.63E-04)	(9.22E-05)	(4.14E-04)
Gini**	0.3043	0.2880	0.3035	0.3296	0.3098	0.3322	0.3298	0.3174	0.3299
	(1.12E-03)	(4.36E-06)	(1.94E-04)	(2.45E-05)	(6.07E-06)	(1.67E-04)	(3.01E-05)	(6.73E-06)	(1.73E-04)
Theil	0.1532	0.1332	0.1521	0.1920	0.1548	0.1842	0.1970	0.1626	0.1815
	(1.19E-03)	(4.12E-06)	(2.05E-04)	(3.23E-05)	(6.23E-06)	(1.97E-04)	(4.42E-05)	(7.09E-06)	(2.02E-04)
Pietra	0.2178	0.2060	0.2173	0.2337	0.2221	0.2385	0.2321	0.2276	0.2368
	(8.25E-04)	(3.19E-06)	(1.43E-04)	(1.79E-05)	(4.48E-06)	(1.24E-04)	(2.17E-05)	(4.98E-06)	(1.28E-04)

Sample size: 22,592,900 *Census estimate: 1.029 **Census estimate: 0.3330

Table D.14: ESTIMATED DISTRIBUTION FUNCTIONS 2015 INCOME

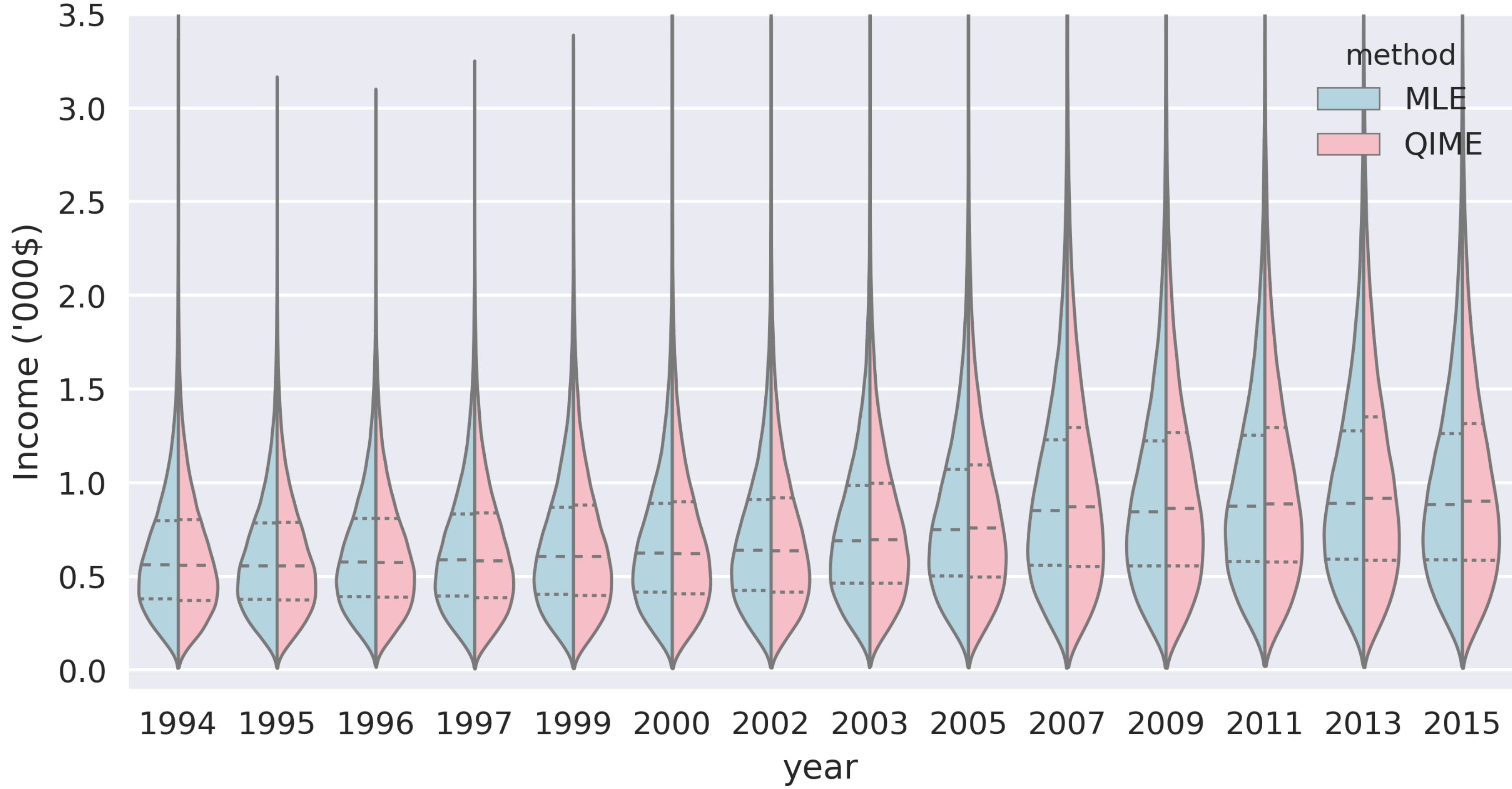
	QME			IME			QIME		
	GB2	G	LN	GB2	G	LN	GB2	G	LN
p, μ	113.7727		-0.1570	4.5473		-0.1608	1.1297		-0.1673
	(4.53E+01)		(1.15E-04)	(4.47E-02)		(1.21E-04)	(2.76E-03)		(1.23E-04)
q, σ^2	62.7674		0.2964	5.1257		0.3407	1.3951		0.3499
	(1.84E+01)		(1.04E-04)	(5.01E-02)		(1.02E-04)	(3.57E-03)		(1.04E-04)
a	0.2894	3.6701		1.1583	3.1771		2.5889	3.0439	
	(4.81E-02)	(1.22E-03)		(6.19E-03)	(9.10E-04)		(4.25E-03)	(8.63E-04)	
b, λ	0.1083	3.7825		0.9538	3.1484		0.9538	3.0156	
	(7.64E-02)	(1.36E-03)		(1.69E-03)	(9.74E-04)		(4.95E-04)	(9.27E-04)	
Mean*	0.9944	0.9703	0.9912	1.0094	1.0091	1.0096	1.0088	1.0094	1.0076
	(5.36E-02)	(5.53E-05)	(4.23E-04)	(4.18E-04)	(7.61E-05)	(4.18E-04)	(5.42E-04)	(8.13E-05)	(4.11E-04)
Gini**	0.3019	0.2847	0.2997	0.3199	0.3044	0.3202	0.3214	0.3104	0.3242
	(2.77E-03)	(4.09E-06)	(2.00E-04)	(6.32E-05)	(5.55E-06)	(1.87E-04)	(2.25E-05)	(6.04E-06)	(1.81E-04)
Theil	0.1512	0.1301	0.1482	0.1722	0.1492	0.1703	0.1812	0.1554	0.1749
	(2.93E-03)	(3.82E-06)	(2.07E-04)	(7.25E-05)	(5.58E-06)	(2.11E-04)	(2.78E-05)	(6.21E-06)	(2.07E-04)
Pietra	0.2161	0.2036	0.2145	0.2284	0.2180	0.2296	0.2268	0.2225	0.2326
	(2.03E-03)	(3.00E-06)	(1.47E-04)	(4.65E-05)	(4.09E-06)	(1.38E-04)	(1.62E-05)	(4.46E-06)	(1.33E-04)

Sample size: 23,224,200 *Census estimate: 1.009 **Census estimate: 0.3230

Appendix E

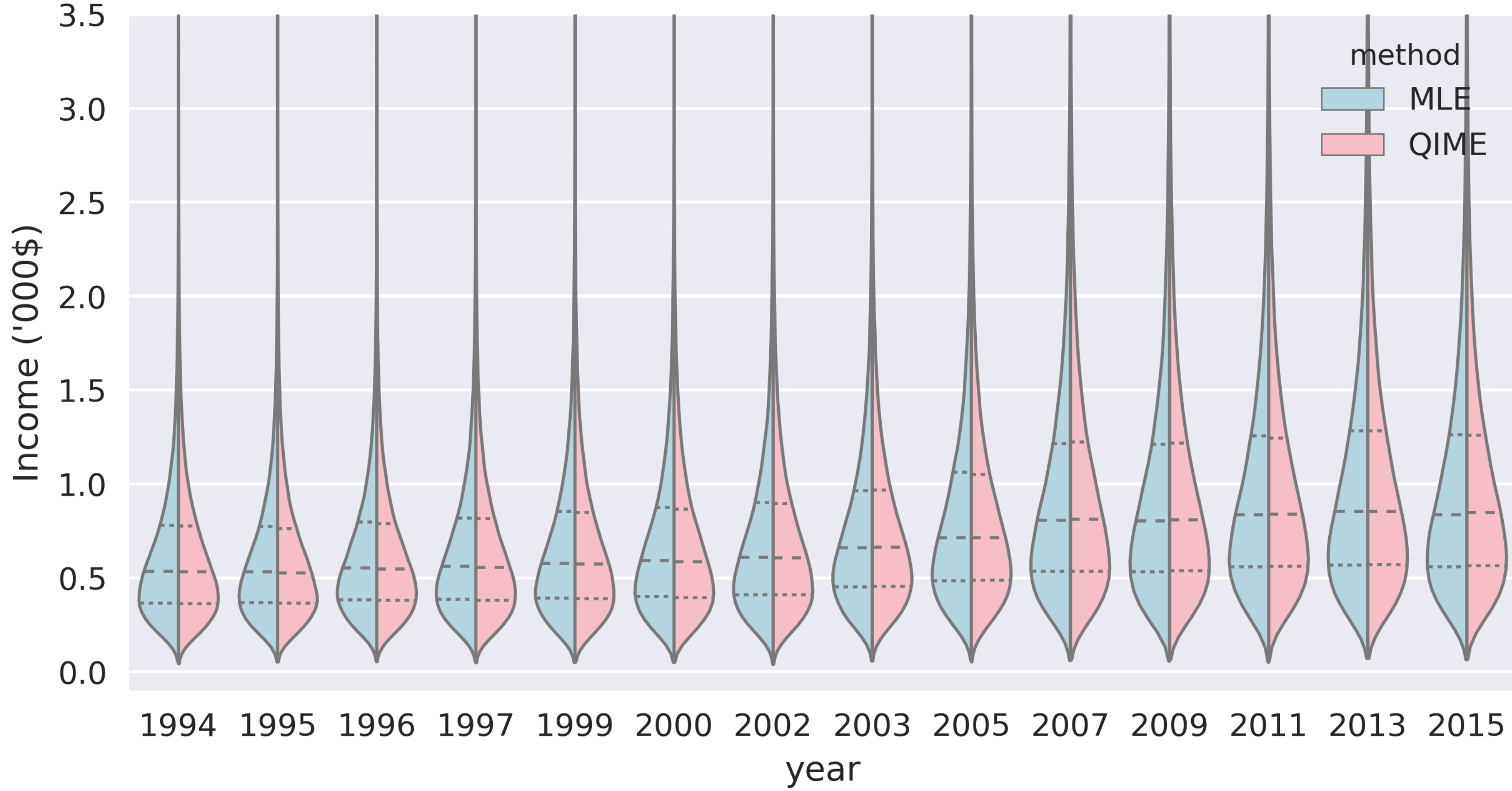
Figures

Figure E-1: VIOLIN PLOT OF GAMMA INCOME DISTRIBUTIONS, Australia, 1994-1995 to 2015-2016



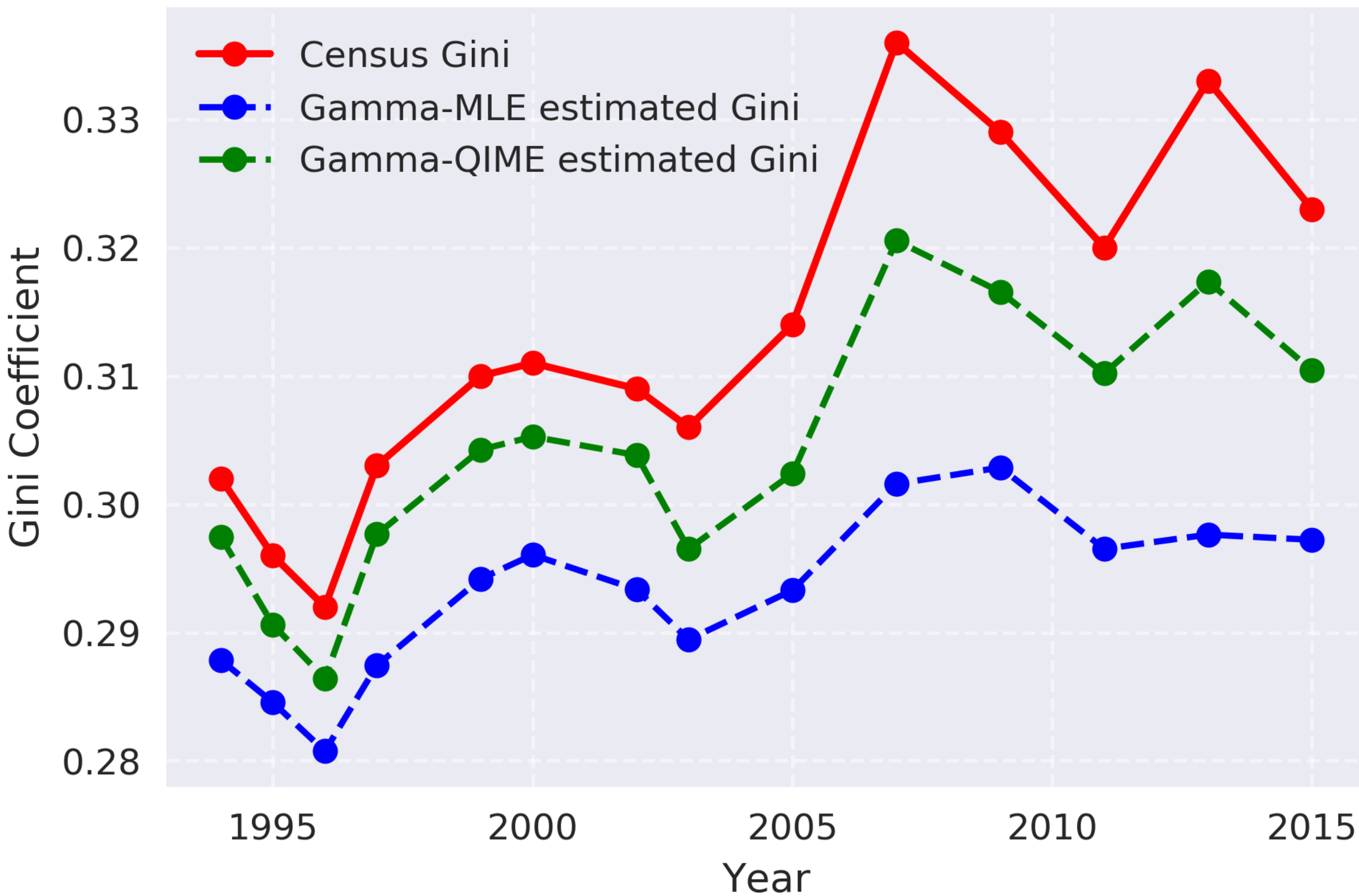
Notes: A violin plot is drawn to analyse the Australian income dynamics over 14 periods using MLE and QIME with Gamma as a model of income distribution. The symmetric density graphs in all periods show excellent agreement between two methods. The picture also reveals a consistent upward shift of overall income level over time. Horizontal dotted lines indicate the estimated first quartile, the median, and the third quartile.

Figure E-2: VIOLIN PLOT OF LOG-NORMAL INCOME DISTRIBUTIONS, 1994 to 2015



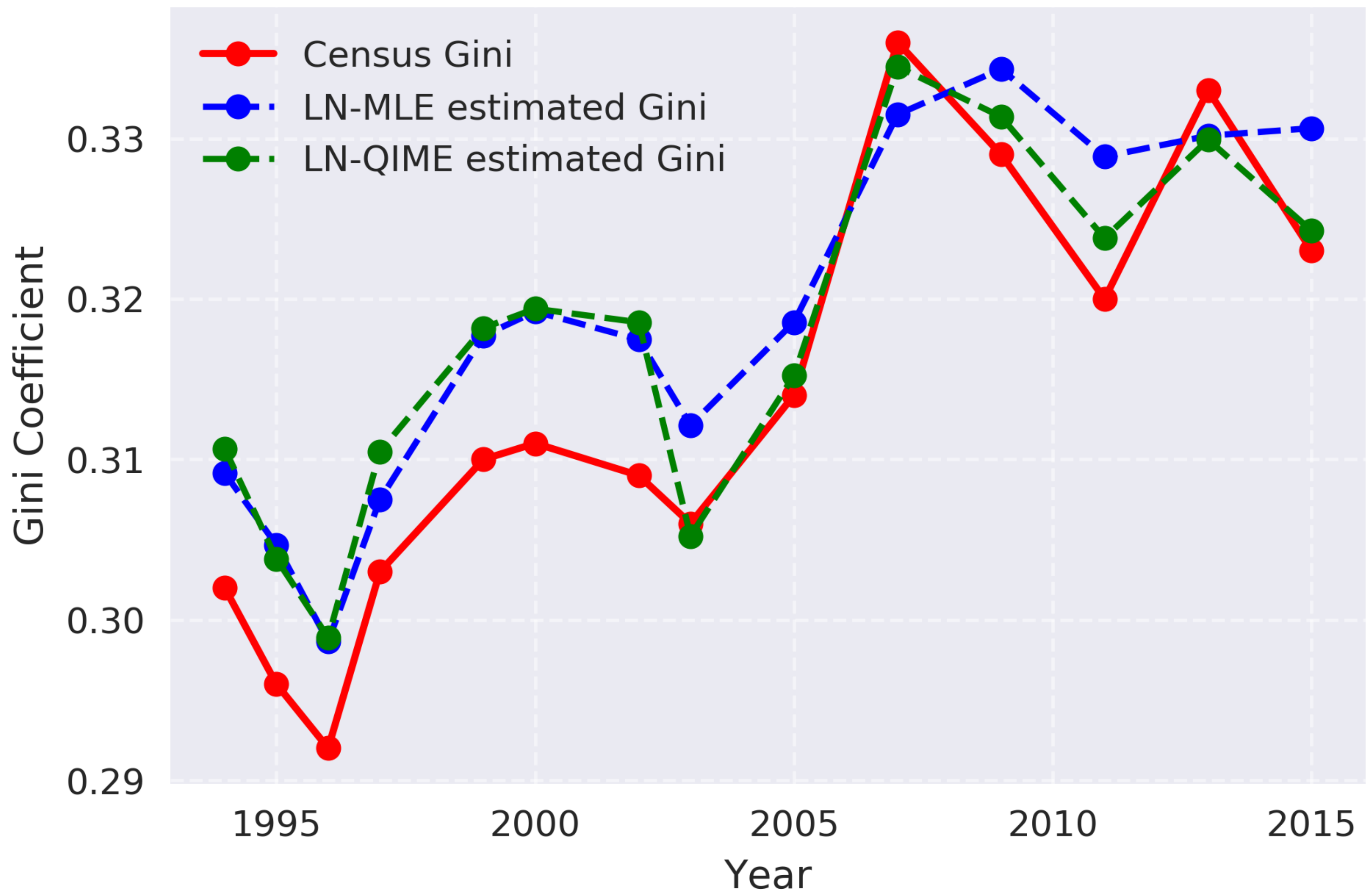
Notes: A violin plot are drawn to analyse the Australian income dynamics over 14 periods using MLE and QIME with Log-normal as a model of income distribution. The symmetric density graphs in all periods show excellent agreement between two methods. The picture also reveals a consistent upward shift of overall income level over time. Horizontal dotted lines indicate the estimated first quartile, the median, and the third quartile.

Figure E-3: CENSUS AND ESTIMATED GINI COEFFICIENTS USING GAMMA, 1994 to 2015



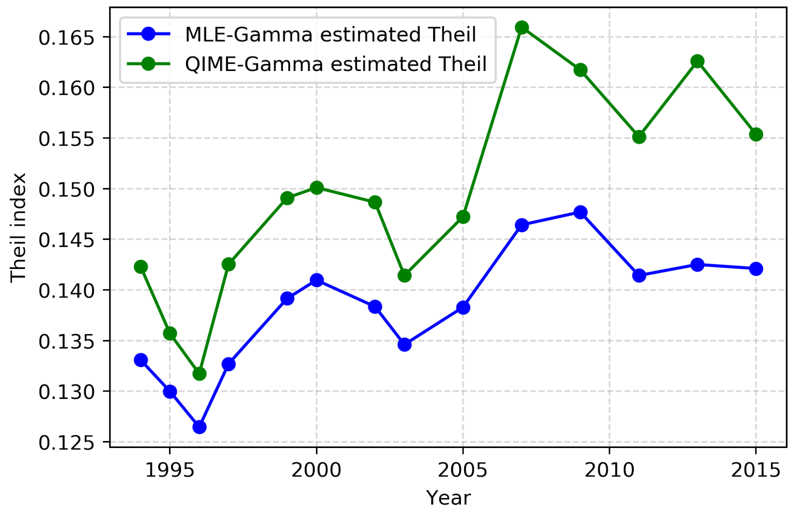
Notes: A time series plot is drawn to depict the movement of Gini index over time, extracted from MLE and QIME using Gamma distribution. Both estimated Gini series is able to trace the rising or falling movements but consistently underestimate the census series.

Figure E-4: CENSUS AND ESTIMATED GINI COEFFICIENTS USING LOG-NORMAL, Australia, 1994-1995 to 2015-2016



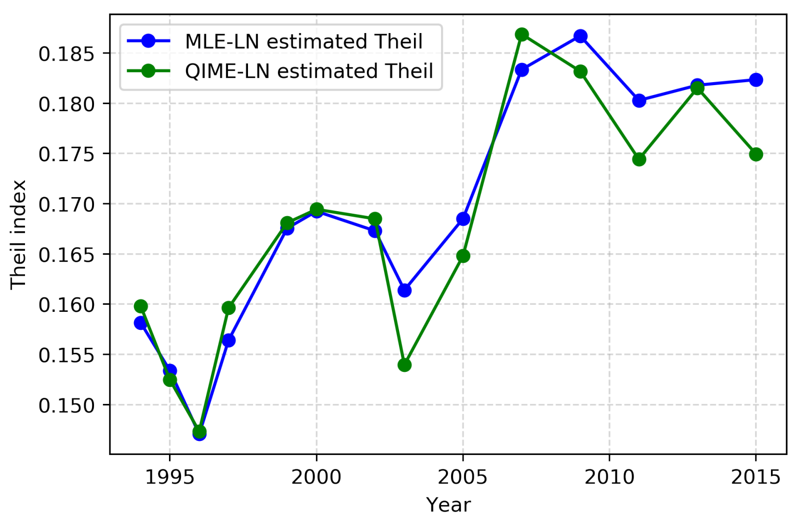
Notes: A time series plot is drawn to depict the movement of Gini index over time, extracted from MLE and QIME using Log-normal distribution. Both estimated Gini series trace the census series moderately closed.

Figure E-5: ESTIMATED THEIL COEFFICIENTS USING GAMMA, 1994 to 2015



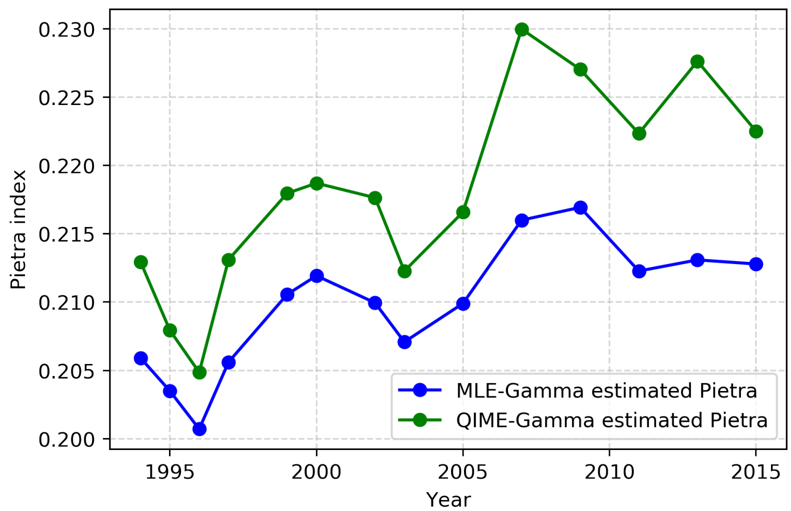
Notes: A time series plot is drawn to depict the movement of Theil index over time, extracted from MLE and QIME using Gamma distribution. Both estimated Theil series exhibit an increasing trend in inequality consistent with the Gini measure.

Figure E-6: ESTIMATED THEIL COEFFICIENTS USING LOG-NORMAL, 1994 to 2015



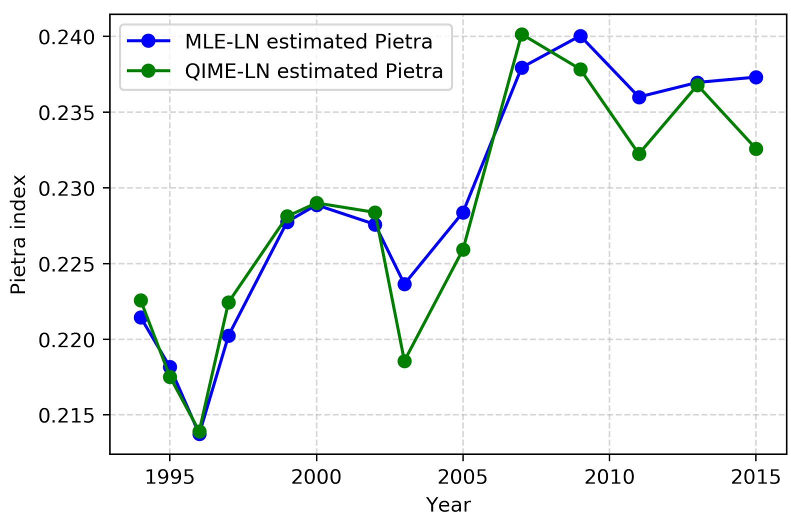
Notes: A time series plot is drawn to depict the movement of Theil index over time, extracted from MLE and QIME using Log-normal distribution. Both estimated Theil series exhibit an increasing trend in inequality consistent with the Gini measure.

Figure E-7: ESTIMATED PIETRA COEFFICIENTS USING GAMMA, 1994 to 2015



Notes: A time series plot is drawn to depict the movement of Pietra index over time, extracted from MLE and QIME using Gamma distribution. Both estimated Pietra series exhibit an increasing trend in inequality consistent with the Gini measure.

Figure E-8: ESTIMATED PIETRA COEFFICIENTS USING LOG-NORMAL, 1994 to 2015



Notes: A time series plot is drawn to depict the movement of Pietra index over time, extracted from MLE and QIME using Log-normal distribution. Both estimated Pietra series exhibit an increasing trend in inequality consistent with the Gini measure.

Bibliography

Alvaredo, F., L. Chancel, T. Piketty, E. Saez and G. Zucman (2018): *World inequality report 2018*. Belknap Press of Harvard University Press.

Arnold, B. C., N. Balakrishnan and H. N. Nagaraja (1992): *A first course in order statistics*, vol. 54. Siam.

Bahadur, R. R. (1966): "A note on quantiles in large samples," *The Annals of Mathematical Statistics*, 37(3), 577–580.

Bakker, A. and J. Creedy (1998): "Estimating the exponential family using grouped data: An application to the New Zealand income distribution," *New Zealand Economic Papers*, 32(1), 19–39.

Bakker, A. and J. Creedy (1999): "Macroeconomic variables and income inequality in New Zealand: an exploration using conditional mixture distributions," *New Zealand Economic Papers*, 33(2), 59–79.

Blitzstein, J. K. and J. Hwang (2014): *Introduction to probability*. Chapman and Hall/CRC.

Chotikapanich, D. (2008): *Modeling income distributions and Lorenz curves*, vol. 5. Springer Science & Business Media.

Chotikapanich, D. and W. E. Griffiths (2002): "Estimating Lorenz curves using a Dirichlet distribution," *Journal of Business & Economic Statistics*, 20(2), 290–295.

——— (2008): "Estimating income distributions using a mixture of gamma densities," in *Modeling Income Distributions and Lorenz Curves*, pp. 285–302. Springer.

Chotikapanich, D., W. E. Griffiths and D. P. Rao (2007): "Estimating and combining national income distributions using limited data," *Journal of Business & Economic Statistics*, 25(1), 97–109.

Chotikapanich, D., D. P. Rao and K. K. Tang (2007): "Estimating income inequality in China using grouped data and the generalized beta distribution," *Review of Income and Wealth*, 53(1), 127–147.

Cowell, F. (2011): *Measuring Inequality*, LSE perspectives in economic analysis. OUP Oxford.

Cramér, H. (2016): *Mathematical methods of statistics (PMS-9)*, vol. 9. Princeton university press.

Dagum, C. (1977): "New model of personal income-distribution-specification and estimation," *Economie appliquée*, 30(3), 413–437.

Dempster, A. P., N. M. Laird and D. B. Rubin (1977): "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the royal statistical society. Series B (methodological)*, pp. 1–38.

Do, C. B. and S. Batzoglou (2008): "What is the expectation maximization algorithm?," *Nature biotechnology*, 26(8), 897.

Dominicy, Y. and D. Veredas (2013): "The method of simulated quantiles," *Journal of Econometrics*, 172(2), 235–247.

Dorfman, R. (1979): "A formula for the Gini coefficient," *The review of economics and statistics*, pp. 146–149.

Fisk, P. R. (1961): "The graduation of income distributions," *Econometrica: journal of the Econometric Society*, pp. 171–185.

GIBRAT, R. (1931): *Les inégalités économiques: applications: aux inégalitês des richesses, à la concentration des entreprises, aux populations des villes, aux statistiques des familles, etc: d'une loi nouvelle: la loi de l'effet proportionnel*. Librairie du Recueil Sirey.

GINI, C. (1912): "Variability and mutability, contribution to the study of statistical distribution and relaitons," *Studi Economico-Giuricici della R.*

HANSEN, L. P. (1982): "Large sample properties of generalized method of moments estimators," *Econometrica: Journal of the Econometric Society*, pp. 1029–1054.

HANSEN, L. P., J. HEATON AND A. YARON (1996): "Finite-sample properties of some alternative GMM estimators," *Journal of Business & Economic Statistics*, 14(3), 262–280.

HANSEN, L. P. AND K. J. SINGLETON (1982): "Generalized instrumental variables estimation of nonlinear rational expectations models," *Econometrica: Journal of the Econometric Society*, pp. 1269–1286.

HAYASHI, F. (2000): "Econometrics. Princeton," *New Jersey, USA: Princeton University.*

HINKLEY, D. V. AND D. COX (1979): *Theoretical statistics*. Chapman and Hall/CRC.

KAKWANI, N. C. AND N. PODDER (1976): "Efficient estimation of the Lorenz curve and associated inequality measures from grouped observations," *Econometrica: Journal of the Econometric Society*, pp. 137–148.

KLEIBER, C. AND S. KOTZ (2003): *Statistical size distributions in economics and actuarial sciences*, vol. 470. John Wiley & Sons.

KOBAYASHI, G. AND K. KAKAMU (2016): "Approximate Bayesian Computation for Lorenz Curves from Grouped Data," *arXiv preprint arXiv:1607.02735.*

KOECHLIN, T. (2013): "The Rich Get Richer: Neoliberalism and Soaring Inequality in the United States," *Challenge*, 56(2), 5–30.

KROESE, D. P., J. C. CHAN ET AL. (2014): *Statistical modeling and computation*. Springer.

LORENZ, M. O. (1905): "Methods of measuring the concentration of wealth," *Publications of the American statistical association*, 9(70), 209–219.

LU, T.-T. AND S.-H. SHIOU (2002): "Inverses of 2× 2 block matrices," *Computers & Mathematics with Applications*, 43(1-2), 119–129.

LUBRANO, M. AND A. A. J. NDOYE (2016): "Income inequality decomposition using a finite mixture of log-normal distributions: A Bayesian approach," *Computational Statistics & Data Analysis*, 100, 830–846.

MAGNUS, W., F. OBERHETTINGER AND R. P. SONI (2013): *Formulas and theorems for the special functions of mathematical physics*, vol. 52. Springer Science & Business Media.

MANDELBROT, B. (1960): "The Pareto-Levy law and the distribution of income," *International Economic Review*, 1(2), 79–106.

MATSUYAMA, Y. (2003): "The a-EM algorithm: surrogate likelihood maximization using a-logarithmic information measures," *IEEE Transactions on Information Theory*, 49(3), 692–706.

MCDONALD, J. B. (1984): "Some generalized functions for the size distribution of income," *Econometrica: journal of the Econometric Society*, pp. 647–663.

MCDONALD, J. B. AND B. C. JENSEN (1979): "An Analysis of Some Properties of Alternative Measures of Income Inequality Based on the Gamma Distribution Function," *Journal of the American Statistical Association*, 74(368), 856–860.

McDonald, J. B. and M. R. Ransom (1979): "Functional forms, estimation techniques and the distribution of income," *Econometrica: Journal of the Econometric Society*, pp. 1513–1525.

McDonald, J. B. and Y. J. Xu (1995): "A generalization of the beta distribution with applications," *Journal of Econometrics*, 66(1-2), 133–152.

Paapaa, R. and H. K. Van Dijk (1998): "Distribution and mobility of wealth of nations," *European Economic Review*, 42(7), 1269–1293.

Pareto, V. and C. d. Politique (1897): "Vol. 2," *Rouge, Lausanne*.

Pen, J. (1974): *Income distribution*. Penguin (Non-Classics).

Piketty, T. (2014): *Capital in the Twenty-First Century*. Harvard University Press.

——— (2015): *The Economics of Inequality*. Harvard University Press.

Rao, C. R. (1992): "Information and the accuracy attainable in the estimation of statistical parameters," in *Breakthroughs in statistics*, pp. 235–247. Springer.

Reed, W. J. and F. Wu (2008): "New four-and five-parameter models for income distributions," in *Modeling Income Distributions and Lorenz Curves*, pp. 211–223. Springer.

Salem, A. B. and T. D. Mount (1974): "A convenient descriptive model of income distribution: the gamma density," *Econometrica: journal of the Econometric Society*, pp. 1115–1127.

Sarabia, J. M. and V. Jordá (2014): "Explicit expressions of the Pietra index for the generalized function for the size distribution of income," *Physica A: Statistical Mechanics and its Applications*, 416, 582–595.

Sargan, J. D. (1958): "The estimation of economic relationships using instrumental variables," *Econometrica: Journal of the Econometric Society*, pp. 393–415.

Serfling, R. J. (2009): *Approximation theorems of mathematical statistics*, vol. 162. John Wiley & Sons.

Sgouropoulos, N., Q. Yao and C. Yastremiz (2015): "Matching a distribution by matching quantiles estimation," *Journal of the American Statistical Association*, 110(510), 742–759.

Simon, H. A. (1955): "On a class of skew distribution functions," *Biometrika*, 42(3/4), 425–440.

——— (1957): "The compensation of executives," *Sociometry*, 20(1), 32–35.

Simon, H. A. and C. P. Bonini (1958): "The size distribution of business firms," *The American economic review*, 48(4), 607–617.

Singhi, S. and G. Maddala (1976): "A Function for the Size Distribution of Incomes. In 29," *Econometrica*, 44(5), 963–970.

Stigler, S. M. (1973): "The asymptotic distribution of the trimmed mean," *The Annals of Statistics*, pp. 472–477.

Strang, L. A. (1980): "Linear Algebra and Its Applications: Second Edition," .

Theil, H. (1967): *Economics and information theory*, Studies in mathematical and managerial economics. North-Holland Pub. Co.

Wooldridge, J. M. (2010): *Econometric analysis of cross section and panel data*. MIT press.

Wu, C. J. (1983): "On the convergence properties of the EM algorithm," *The Annals of statistics*, pp. 95–103.