



**Pedro Filipe  
Pessoa Macedo**

**Contributos para a teoria de máxima entropia na  
estimação de modelos mal-postos**

**Contributions to the theory of maximum entropy  
estimation for ill-posed models**





**Pedro Filipe  
Pessoa Macedo**

**Contributos para a teoria de máxima entropia na  
estimação de modelos mal-postos**

**Contributions to the theory of maximum entropy  
estimation for ill-posed models**

Tese apresentada à Universidade de Aveiro para cumprimento dos requisitos necessários à obtenção do grau de Doutor em Matemática, realizada sob a orientação científica do Doutor Manuel González Scotto, Professor Auxiliar com Agregação do Departamento de Matemática da Universidade de Aveiro, e da Doutora Elvira Maria de Sousa Silva, Professora Associada com Agregação da Faculdade de Economia da Universidade do Porto.

Apoio financeiro da FCT - Fundação para a Ciência e a Tecnologia através da bolsa de doutoramento com referência SFRH/BD/40821/2007.

Apoio financeiro do FEDER através do COMPETE – Programa Operacional Fatores de Competitividade e de fundos nacionais através do Centro de Investigação e Desenvolvimento em Matemática e Aplicações (Universidade de Aveiro) e da FCT, no âmbito do projeto PEst-C/MAT/UI4106/2011 com o número COMPETE FCOMP-01-0124-FEDER-022690.



Ao Francisco, à Carolina e à Sílvia.



## **o júri**

presidente

**Prof. Doutor Aníbal Manuel de Oliveira Duarte**  
Professor Catedrático da Universidade de Aveiro

vogais

**Prof. Doutor António Manuel Pacheco Pires**  
Professor Catedrático do Instituto Superior Técnico da Universidade Técnica de Lisboa

**Prof.<sup>a</sup> Doutora Elvira Maria de Sousa Silva**  
Professora Associada com Agregação da Faculdade de Economia da Universidade do Porto  
(Coorientadora)

**Prof.<sup>a</sup> Doutora Diana Elisabeta Aldea Mendes**  
Professora Associada do Instituto Universitário de Lisboa

**Prof. Doutor Manuel González Scotto**  
Professor Auxiliar com Agregação da Universidade de Aveiro (Orientador)

**Prof.<sup>a</sup> Doutora Andreia Teixeira Marques Dionísio**  
Professora Auxiliar da Escola de Ciências Sociais da Universidade de Évora

**Prof.<sup>a</sup> Doutora Maria Manuela Souto de Miranda**  
Professora Auxiliar da Universidade de Aveiro



## **agradecimentos**

Ao Professor Manuel Scotto e à Professora Elvira Silva pela dedicada orientação científica. Obrigadíssimo.

À Fundação para a Ciência e a Tecnologia pela bolsa de doutoramento.

À Universidade de Aveiro, ao Departamento de Matemática e ao Centro de Investigação e Desenvolvimento em Matemática e Aplicações pelos diversos apoios concedidos. Um especial agradecimento ao Professor João Santos e ao Professor Luís Castro.

Aos meus colegas do Departamento de Matemática e do Departamento de Economia, Gestão e Engenharia Industrial, em particular os que comigo privaram nos últimos anos, pela camaradagem.

À minha mulher e aos meus filhos pelo apoio incondicional.



## palavras-chave

Máxima entropia, colinearidade, *outliers*, pequenas amostras, regressão linear, regressão robusta, regressão *ridge*, parâmetro *ridge*, electrodinâmica quântica, eficiência técnica, fronteiras de produção com estados contingentes.

## resumo

As técnicas estatísticas são fundamentais em ciência e a análise de regressão linear é, quiçá, uma das metodologias mais usadas. É bem conhecido da literatura que, sob determinadas condições, a regressão linear é uma ferramenta estatística poderosíssima. Infelizmente, na prática, algumas dessas condições raramente são satisfeitas e os modelos de regressão tornam-se mal-postos, inviabilizando, assim, a aplicação dos tradicionais métodos de estimação.

Este trabalho apresenta algumas contribuições para a teoria de máxima entropia na estimação de modelos mal-postos, em particular na estimação de modelos de regressão linear com pequenas amostras, afetados por colinearidade e *outliers*. A investigação é desenvolvida em três vertentes, nomeadamente na estimação de eficiência técnica com fronteiras de produção condicionadas a estados contingentes, na estimação do parâmetro *ridge* em regressão *ridge* e, por último, em novos desenvolvimentos na estimação com máxima entropia.

Na estimação de eficiência técnica com fronteiras de produção condicionadas a estados contingentes, o trabalho desenvolvido evidencia um melhor desempenho dos estimadores de máxima entropia em relação ao estimador de máxima verosimilhança. Este bom desempenho é notório em modelos com poucas observações por estado e em modelos com um grande número de estados, os quais são comumente afetados por colinearidade. Espera-se que a utilização de estimadores de máxima entropia contribua para o tão desejado aumento de trabalho empírico com estas fronteiras de produção.

Em regressão *ridge* o maior desafio é a estimação do parâmetro *ridge*. Embora existam inúmeros procedimentos disponíveis na literatura, a verdade é que não existe nenhum que supere todos os outros. Neste trabalho é proposto um novo estimador do parâmetro *ridge*, que combina a análise do traço *ridge* e a estimação com máxima entropia. Os resultados obtidos nos estudos de simulação sugerem que este novo estimador é um dos melhores procedimentos existentes na literatura para a estimação do parâmetro *ridge*.

O estimador de máxima entropia de Leuven é baseado no método dos mínimos quadrados, na entropia de Shannon e em conceitos da eletrodinâmica quântica. Este estimador suplanta a principal crítica apontada ao estimador de máxima entropia generalizada, uma vez que prescinde dos suportes para os parâmetros e erros do modelo de regressão. Neste trabalho são apresentadas novas contribuições para a teoria de máxima entropia na estimação de modelos mal-postos, tendo por base o estimador de máxima entropia de Leuven, a teoria da informação e a regressão robusta. Os estimadores desenvolvidos revelam um bom desempenho em modelos de regressão linear com pequenas amostras, afetados por colinearidade e *outliers*.

Por último, são apresentados alguns códigos computacionais para estimação com máxima entropia, contribuindo, deste modo, para um aumento dos escassos recursos computacionais atualmente disponíveis.



**keywords**

Maximum entropy, collinearity, outliers, small samples sizes, robust regression, linear regression, ridge regression, ridge parameter, quantum electrodynamics, technical efficiency, state-contingent production frontiers.

**abstract**

Statistical techniques are essential in most areas of science being linear regression one of the most widely used. It is well-known that under fairly conditions linear regression is a powerful statistical tool. Unfortunately, some of these conditions are usually not satisfied in practice and the regression models become ill-posed, which means that the application of traditional estimation methods may lead to non-unique or highly unstable solutions.

This work is mainly focused on the maximum entropy estimation of ill-posed models, in particular the estimation of regression models with small samples sizes affected by collinearity and outliers. The research is developed in three directions, namely the estimation of technical efficiency with state-contingent production frontiers, the estimation of the ridge parameter in ridge regression, and some developments in maximum entropy estimation.

In the estimation of technical efficiency with state-contingent production frontiers, this work reveals that the maximum entropy estimators outperform the maximum likelihood estimator in most of the cases analyzed, namely in models with few observations in some states of nature and models with a large number of states of nature, which usually represent models affected by collinearity. The maximum entropy estimators are expected to make an important contribution to the increase of empirical work with state-contingent production frontiers.

The main challenge in ridge regression is the selection of the ridge parameter. There is a huge number of methods to estimate the ridge parameter and no single method emerges in the literature as the best overall. In this work, a new method to select the ridge parameter in ridge regression is presented. The simulation study reveals that, in the case of regression models with small samples sizes affected by collinearity, the new estimator is probably one of the best ridge parameter estimators available in the literature on ridge regression.

Founded on the Shannon entropy, the ordinary least squares estimator and some concepts from quantum electrodynamics, the maximum entropy Leuven estimator overcomes the main weakness of the generalized maximum entropy estimator, avoiding exogenous information that is usually not available. Based on the maximum entropy Leuven estimator, information theory and robust regression, new developments on the theory of maximum entropy estimation are provided in this work. The simulation studies and the empirical applications reveal that the new estimators are a good choice in the estimation of linear regression models with small samples sizes affected by collinearity and outliers.

Finally, a contribution to the increase of computational resources on the maximum entropy estimation is also accomplished in this work.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Motivation and objectives . . . . .	1
1.2	Structure of the thesis and main achievements . . . . .	7
<b>2</b>	<b>Background and state of the art</b>	<b>13</b>
2.1	Entropy . . . . .	13
2.1.1	Entropy concepts . . . . .	13
2.1.2	Maximum entropy principle . . . . .	23
2.2	Estimators . . . . .	30
2.2.1	The generalized maximum entropy estimator . . . . .	30
2.2.2	The generalized cross-entropy estimator . . . . .	35
2.2.3	Higher-order entropy estimators . . . . .	37
2.3	Regression diagnostics: collinearity and outliers . . . . .	39
2.4	Technical efficiency . . . . .	43
<b>3</b>	<b>Technical efficiency with state-contingent production frontiers</b>	<b>49</b>
3.1	Introduction . . . . .	49
3.2	State-contingent production with maximum entropy estimators . . . . .	51
3.2.1	A state-contingent production frontier model . . . . .	51
3.2.2	Maximum entropy estimators . . . . .	53
3.3	Simulation study . . . . .	58

3.4	Conclusions . . . . .	65
<b>4</b>	<b>The choice of the ridge parameter in ridge regression</b>	<b>69</b>
4.1	Introduction . . . . .	69
4.2	The ridge regression estimator . . . . .	72
4.3	The Ridge-GME estimator . . . . .	74
4.4	Simulation study . . . . .	76
4.5	Numerical example . . . . .	82
4.6	Conclusions . . . . .	85
<b>5</b>	<b>Some developments in the maximum entropy estimation</b>	<b>87</b>
5.1	Maximum entropy robust regression group estimators . . . . .	87
5.1.1	Introduction . . . . .	87
5.1.2	The maximum entropy Leuven estimator . . . . .	89
5.1.3	The MERG estimators . . . . .	92
5.1.4	Some properties of the MERG estimators . . . . .	95
5.1.5	Simulation study: collinearity and outliers . . . . .	98
5.1.6	Examples and additional simulation studies . . . . .	102
5.1.6.1	HDI and Portland cement models . . . . .	102
5.1.6.2	Additional simulation studies . . . . .	106
5.1.7	Conclusions . . . . .	108
5.2	An extension of the maximum entropy robust regression group estimators . .	108
5.2.1	Introduction . . . . .	108
5.2.2	The MERGE estimators . . . . .	109
5.2.3	Improvements for MERGE estimators . . . . .	112
5.2.3.1	Cross-entropy formalism . . . . .	112
5.2.3.2	Parameter inequality restrictions . . . . .	113
5.2.4	Simulation study . . . . .	114

5.2.5	Conclusions . . . . .	117
<b>6</b>	<b>Concluding remarks</b>	<b>119</b>
6.1	Final conclusions . . . . .	119
6.2	Future work . . . . .	122
	<b>References</b>	<b>125</b>
	<b>Index</b>	<b>141</b>
	<b>Appendices</b>	<b>144</b>
	Appendix A – Shannon entropy as a measure of information . . . . .	145
	Appendix B – Two new measures of technical inefficiency with directional technology distance functions . . . . .	149
	Appendix C – MATLAB codes . . . . .	155



# List of Figures

2.1	Shannon entropy. . . . .	16
2.2	Rényi entropy. . . . .	19
2.3	Tsallis entropy. . . . .	21
2.4	Estimated ME distributions for the die problem. . . . .	29
4.1	Support intervals and ridge interval in the Ridge-GME estimator for an arbitrary ridge trace. . . . .	75
4.2	Ridge trace for the Portland cement model (non-standardized coefficients). . . . .	84
4.3	Ridge trace for the Portland cement model (standardized coefficients). . . . .	84
4.4	Selection of the Ridge-GME estimate for the Portland cement model. . . . .	85
5.1	The analogy with quantum electrodynamics. . . . .	90
5.2	OLS regression lines in the HDI model. . . . .	103
A.1	Shannon entropy as a measure of information. . . . .	145
B.1	Technical inefficiency measures based on the mean and median of the data. . . . .	151
B.2	Different measures of technical inefficiency. . . . .	153



# List of Tables

2.1	Estimated ME distributions for the die problem. . . . .	28
3.1	MSEL and DMTE for the different estimators (Model 1 and Model 2). . . . .	61
3.2	MSEL and DMTE for the different estimators (Model 3 and Model 4). . . . .	62
3.3	MSEL and DMTE for the different estimators (Model 5 and Model 6). . . . .	63
3.4	MSEL and DMTE for the different estimators (Model 7 and Model 8). . . . .	64
3.5	MSEL and DMTE for the different estimators (Model 9 and Model 10). . . . .	65
4.1	MSEL for OLS and different ridge estimators ( $N = 10$ ). . . . .	79
4.2	MSEL for OLS and different ridge estimators ( $N = 20$ ). . . . .	80
4.3	MSEL for OLS and different ridge estimators ( $N = 50$ ). . . . .	81
4.4	MSEL for OLS and different ridge estimators ( $N = 100$ ). . . . .	82
4.5	MSE for different estimators in the Portland cement model. . . . .	85
5.1	MERG estimators. . . . .	93
5.2	MSEL in the simulation study with outliers and collinearity. . . . .	101
5.3	Estimates for $\beta_0$ and $\beta_1$ in the HDI model. . . . .	103
5.4	Estimates in the original Portland cement model. . . . .	104
5.5	Parameter supports for GME estimators in the Portland cement model. . . . .	105
5.6	Estimates in the Portland cement model with intercept. . . . .	106
5.7	MSEL in the simulation study with outliers. . . . .	107

5.8	MSEL in the simulation study with collinearity. . . . .	107
5.9	MERGE estimators. . . . .	110
5.10	MSEL for the estimators in the simulation study ( $N = 10$ ). . . . .	115
5.11	MSEL for the estimators in the simulation study ( $N = 30$ ). . . . .	116
B.1	Most popular directional vectors. . . . .	150
B.2	Results from Figure B.1. . . . .	151
B.3	Results from Figure B.2b. . . . .	153

# List of Abbreviations

BMM Bayesian method of moments

CONSTR optimization function for MATLAB

COVRATIO outlier diagnostic procedure

DEA data envelopment analysis

DFBETAS outlier diagnostic procedure

DFFITS outlier diagnostic procedure

DMTE difference between the true and the estimated mean of technical efficiency

FMINCON optimization function for MATLAB

GAMS software (<http://www.gams.com>)

GAUSS software (<http://www.aptech.com>)

GCE generalized cross-entropy

GCV generalized cross-validation

GDP gross domestic product

GEL generalized empirical likelihood

GLS generalized least squares

GME generalized maximum entropy

GME- $\alpha$  higher-order generalized maximum entropy

- GMM generalized method of moments
- HDI human development index
- HK ridge parameter estimator from Hoerl and Kennard [76]
- HKB ridge parameter estimator from Hoerl et al. [78]
- IEE information and entropy econometrics
- IRLS iteratively reweighted least squares
- KM4 fourth ridge parameter estimator from Muniz and Kibria [120]
- KM5 fifth ridge parameter estimator from Muniz and Kibria [120]
- KM6 sixth ridge parameter estimator from Muniz and Kibria [120]
- KS ridge parameter estimator from Khalaf and Shukur [90]
- LAD least absolute deviations
- LIMDEP software (<http://www.limdep.com>)
- LMS least median of squares
- LTS least trimmed squares
- MATLAB software (<http://www.mathworks.com>)
- ME maximum entropy
- MEL maximum entropy Leuven
- MERG maximum entropy robust regression group
- MERG(E) MERG and MERGE estimators
- MERGE maximum entropy robust regression group extended
- ML maximum likelihood
- MMEL modular maximum entropy Leuven

MSE mean squared error

MSEL mean squared error loss

OLS ordinary least squares

PPP US\$ purchasing power parity United States dollar

QR QR decomposition (orthogonal-triangular decomposition)

Ridge-GME ridge parameter estimator (combines the ridge trace and the GME estimator)

RR-MM robust ridge regression estimator based on repeated M-estimation

SFA stochastic frontier analysis

SIMPS optimization function for MATLAB

VIF variance inflation factors



# General Notation

$\mathbf{A}$  matrix (a letter in bold uppercase)

$\mathbf{A}'$  transpose of matrix  $\mathbf{A}$

$\mathbf{A}^{-1}$  inverse of matrix  $\mathbf{A}$

$\mathbf{a}$  column vector (a letter in bold lowercase)

$\mathbf{a}'$  transpose of the column vector  $\mathbf{a}$  (a row vector)

$\mathbf{x}_k$   $k$ th column of matrix  $\mathbf{X}$

$\mathbf{1}_N$  column vector of ones with dimension  $(N \times 1)$

$\mathbf{I}_N$  identity matrix with dimension  $(N \times N)$

$\hat{\boldsymbol{\beta}}$  estimator (or estimate) of the unknown parameter vector  $\boldsymbol{\beta}$

$\|\cdot\|$  Euclidean norm

$\text{cond}_2$  2-norm condition number

$\otimes$  Kronecker product

$\odot$  element-by-element Hadamard product



# Chapter 1

## Introduction

$$“S = k \log w”$$

Entropy formula in the epitaph of Boltzmann’s grave in Vienna.

The motivation and the objectives of this work, as well as the structure of the thesis and the main achievements, are discussed in this introduction. A reference to the publications and communications produced during the work is provided at the end of the chapter.

### 1.1 Motivation and objectives

Statistical techniques are essential in most areas of science being linear regression one of the most widely used. It is well-known that under fairly conditions linear regression is a powerful statistical tool. Unfortunately, some of these conditions are seldom satisfied in practice. This idea is well expressed in Golan et al. [69, p. 3]:

“[...] because of limited, partial data or insufficient information, many econometric problems fall in the ill-posed, underdetermined category. *Convenient assumptions*, representing information we do not possess, are typically used to convert ill-posed problems into seemingly well-posed statistical models [...]. However, this approach often leads to erroneous interpretations and treatments. In fact, in applied mathematics, statistics and econometrics, ill-posed inverse problems may be the rule rather than the exception.”

The main motivation of this work comes from this unfortunate reality. In a traditional linear regression framework in which the main interest is to recover the unknown parameter vector from a known matrix with explanatory variables and a known vector of noisy observations, a regression model is, in general, ill-posed if it does not satisfy the required conditions of classical statistical estimation methods.<sup>1</sup> In such cases, the application of traditional estimation methods might lead to obtain non-unique solutions and/or solutions that may be highly unstable, i.e., very sensitive to small perturbations in the original data.<sup>2</sup> Ill-posedness is a broad concept. However, the ill-posedness of a model typically arises from the limited information available, usually from small samples sizes, incomplete data or when the number of the unknown parameters exceeds the number of observations (under-determined model); and/or from an experiment that can be badly designed (e.g., lead to models with aggregated or missing data), or an experiment that can simply results in a model affected by collinearity and/or outliers; see, among others, Golan [65], Golan et al. [69] and O’Sullivan [126], and the references therein.

At this point it seems reasonable to ask how to make the best possible predictions with such ill-posed problems. An attractive approach which plays a central role in this work is the maximum entropy (ME) principle due to Edwin Jaynes; e.g., Jaynes [81, 82, 83, 84, 85, 86]. The ME principle, as well as others methods available in the literature, such as the generalized maximum entropy (GME), the higher-order generalized maximum entropy (GME- $\alpha$ ), the generalized cross-entropy (GCE), the generalized method of moments (GMM), the Bayesian method of moments (BMM), the generalized empirical likelihood (GEL) and, in general, all the methods directly or indirectly related to the information and entropy econometrics (IEE) research field, are designed to extract information from limited and noisy data using minimal statements on the data generation process; see Golan [64, 65, 66] and the references therein for a review.<sup>3</sup> These methods are established on three general assumptions:

**Assumption 1.1.** *Not everything about the model is known.*

**Assumption 1.2.** *Only minimal a priori assumptions should be assigned.*

---

<sup>1</sup>As noted by O’Sullivan [126], there is an inverse problem whenever inferences are made from partial or incomplete information, and thus statistical estimation is an inverse problem. The inverse problems that are not suitable to the classical statistical estimation methods are defined as ill-posed. In this work, for simplicity, an ill-posed inverse model/problem is just denoted as ill-posed model/problem.

<sup>2</sup>This definition is usually attributed to Jacques Hadamard at the beginning of the XX century.

<sup>3</sup>Golan [65, p. 9] presents an interesting graph with some historical topics on IEE.

**Assumption 1.3.** *The solution should reflect only the available information.*

These are logical assumptions for any estimation method to deal with ill-posed models. The IEE literature is mainly concerned with the estimation of ill-posed models and the study of information measures with a particular emphasis on economic problems. As the name itself suggests, entropy, information theory and ME are central in IEE; section 2.1 presents an overview on some interpretations of entropy from its advent in classical thermodynamics to current applications in science, as well as the ME principle proposed by Edwin Jaynes.

This thesis is mainly focused on the ME estimation of ill-posed models, in particular the ME estimation of linear regression models with small samples sizes affected by collinearity, a problem that hamper the empirical work with regression analysis, and it is probably the most frequent characteristic of the ill-posed models in real-world problems. The presence of outliers in the linear regression model is also discussed, though briefly, in this thesis, particularly when associated with collinearity; in section 2.2 a brief review of some ME estimators is presented.

Why are these topics so relevant in regression analysis and why their presence lead to ill-posed models? Regression models with small samples sizes may compromise the usefulness of classical statistical methods, as well as statistical inference. If there are cases where it is possible to collect additional information (normally requiring more time and cost), there are other cases where such additional information simply does not exist! In either case, it is necessary to make the best possible predictions with such limited information.<sup>4</sup> Collinearity is the term usually used in the literature to represent a near-linear relationship between two or more regressors. Collinearity is responsible for inflating the variance associated with the regression coefficients estimates, and, in general, may affect the signs of the estimates, as well as statistical inference. Outliers are atypical observations being often influential observations that can produce a large impact on the ordinary least squares (OLS) parameter estimates. In short, small samples, collinearity and outliers may lead, although due to different reasons, to absurd results in regression analysis, since the solutions may be undefined or may be highly unstable; in section 2.3 some diagnostic procedures and strategies for dealing with collinearity and outliers are briefly reviewed.

---

<sup>4</sup>Note that there are not precise definitions of a small sample and a large sample. These definitions vary across different areas of science and depend on several factors. However, in empirical work, the boundary between small and large samples usually lies between 30 and 50 observations.

In this research work, the ME estimation of regression models with small samples sizes, collinearity and outliers is investigated from both a theoretical and applied points of view. By using the links between information theory, ME and statistical inference this research is developed in three directions, namely: (a) the estimation of technical efficiency with state-contingent production frontiers; (b) the estimation of the ridge parameter in ridge regression, and (c) some potential developments in the ME estimation.

In a single input-output production technology, technical efficiency can be defined as the ability to minimize the quantity of input used in the production of a given quantity of an output, or the ability to maximize the quantity of output produced with a given quantity of an input. Technical efficiency can be computed comparing the observed output and the potential output of a production unit (e.g., a firm). Thus, technical efficiency analysis is a fundamental tool to measure the performance of the production activity. There is a wide range of methodologies to measure technical efficiency and the choice of a specific approach is always controversial, since different choices lead to different results; e.g., Kalirajan and Shand [89] and Kumbhakar and Lovell [96]. Section 2.4 presents a brief review on technical efficiency analysis.

In the last decade, after the work of Chambers and Quiggin [23], an increasing interest with the state-contingent production frontiers has emerged in the production literature, which contributed, at least partially, to decrease the controversy in the production analysis under uncertainty. This interest, as noted by Quiggin and Chambers [134], is due to the fact that uncertainty in economics is best interpreted in a state-contingent framework. However, this increasing (theoretical) interest has not yet been reflected in an increase of empirical work with this approach. Why? The answer is straightforward: the empirical models with state-contingent production frontiers are usually ill-posed. In particular, these empirical models involve small samples sizes and are affected by (severe) collinearity.

*Thus, how to increase the empirical work with state-contingent production frontiers? In particular, how to estimate technical efficiency with state-contingent production frontiers under difficult empirical conditions (ill-posed models)?*

The **first objective of this work** is to develop a fairly general extension of the production model proposed by O'Donnell et al. [125] that can be employed in real-world empirical applications, and to develop all the procedures to use the GME, GME- $\alpha$  and GCE estima-

tors to assess technical efficiency with state-contingent production frontiers under difficult empirical conditions. The main goal is to make a contribution to the empirical literature on state-contingent production frontiers, probably the most complete approach to model uncertainty in economics; see Chapter 3 for details.

Ridge regression discussed by Hoerl and Kennard [76] is a very popular estimation methodology to handle collinearity without removing variables from the regression model. The importance of this methodology is discussed by McDonald [112] that analyzed the number of publications related to ridge regression in the *Technometrics*, the *Journal of the American Statistical Association*, the *Communications in Statistics – Theory and Methods*, and the *Communications in Statistics – Simulation and Computation*. Approximately 300 articles related to ridge regression has been published in these four scientific journals since the seventies.<sup>5</sup> In the presence of collinearity, traditional estimators such as the OLS estimator perform poorly since the variances of the parameter estimates can be substantially large. By adding a small non-negative constant (often referred to as the ridge parameter) to the diagonal of the correlation matrix of the explanatory variables, it is possible to reduce the variance of the OLS estimator through the introduction of some bias into the regression model.

The challenge in ridge regression is the selection of the ridge parameter. This choice is usually made by the inspection of the ridge trace (a subjective choice) or by a formal method (depending on some parameters that must be estimated from the data). Moreover, there is a huge number of methods to estimate the ridge parameter (several dozens) and no single method emerges in the literature as the best overall.

*Thus, how to select the ridge parameter? Is it possible to find a method that reduces the subjectivity in the selection of the ridge parameter and/or does not depend crucially on other parameters that must be estimated from the data?*

The **second objective of this work** is to introduce a new estimator for the ridge parameter, combining the analysis of the ridge trace with the ME estimation, namely the GME estimator. The main goal is to create one of the best ridge parameter estimators in the literature of ridge regression, in particular concerning regression models with small samples sizes; see Chapter 4 for details.

---

<sup>5</sup>This number of publications results from a recent update made by us in May, 2012, based on the study of McDonald [112].

The GME estimator developed by Golan et al. [69] is described in subsection 2.2.1. The GME estimator has acquired special importance in the toolkit of econometric techniques, by allowing econometric formulations free of restrictive and unnecessary assumptions. Moreover, the GME estimator is useful in linear regression models with small samples sizes, in which the design matrix is ill-conditioned and/or the number of unknown parameters exceeds the number of observations. However, despite these advantages, many statisticians reject the GME estimator. The main weakness of the GME estimator is that support intervals (i.e., exogenous information that may not be always available) for the parameters and error vectors are needed. Those supports are defined as closed and bounded intervals in which each parameter or error is restricted to lie.

Because of this (possible) difficulty in the definition of support intervals, Paris [127] develops the maximum entropy Leuven (MEL) estimator based on some ideas from the theory of light (quantum electrodynamics) of Feynman [57], the Shannon entropy measure and the OLS estimator. Paris [127, 128] shows that the MEL estimator can rival with the GME estimator in linear regression models affected by collinearity, without requiring exogenous information as in the GME estimator.

*Is it possible to improve the MEL estimator? Is it possible to generalize this estimator using information theory and robust regression?*

The **third objective of this work** is to generalize the MEL estimator using other entropy measures and different methods employed in robust regression literature. The idea is to explore the advantages obtained by merging different entropy measures and robust estimators in order to improve the performance of the MEL estimator, namely in the estimation of linear regression models with small samples sizes affected by collinearity and/or outliers. Moreover, since there are some doubts whether the analogy with the theory of light (quantum electrodynamics) used in the MEL estimator is valid in different regression models, other approaches should be investigated. Discussing some directions for future research in IEE, Golan [66] raises a question about the possibility of making the theory easier for application by the practitioners. The third objective also attempts to address this question, by developing new methodologies that are simpler and easier to apply; see Chapter 5 for details.

Finally, concerning computational resources for the ME estimation, there are only a few (and sometimes limited) options available in commercial software for the estimation of linear

regression models. Thus, in most of the cases, researchers and practitioners need to develop their own codes using different optimization tools.<sup>6</sup> Naturally, the lack of computational resources for the ME estimation does not help the diffusion of these estimation techniques among practitioners.

*Is it possible to develop some friendly ME codes designed to users that are not familiar with the ME estimation, using a popular and widely disseminated software?*

A **final objective of this work** is to develop some user friendly codes for the ME estimation in MATLAB, and thus to make a contribution to the increase of computational resources for this type of estimation; see Appendix C for details.

## 1.2 Structure of the thesis and main achievements

This thesis contains six chapters, including the introduction and the concluding remarks. Chapter 2 includes background results and the state of the art on different topics covered in this work. Although the literature review is the main focus of Chapter 2, some original work (e.g., examples, discussions and proofs) is also presented within this chapter. Chapters 3, 4 and 5 present the main contributions of this work, namely the estimation of technical efficiency with state-contingent production frontiers, the estimation of the ridge parameter in ridge regression, and some developments in the ME estimation. Finally, an illustration of the Shannon entropy as a measure of information, a short original research on measures of technical inefficiency with directional distance functions, and some new MATLAB codes for ME estimation are provided in the appendices.

Chapter 3 presents original work concerning the estimation of technical efficiency with state-contingent production frontiers under difficult empirical conditions, combining the GCE, GME and GME- $\alpha$  estimators. The contributions of this chapter are

- an extension of the production model proposed by O'Donnell et al. [125] that can be employed in real-world empirical applications;
- the procedures to use the GME, GME- $\alpha$  and GCE estimators to assess technical efficiency with state-contingent production frontier models, namely

---

<sup>6</sup>Some computational codes are available in <http://www.american.edu/cas/economics/info-metrics>.

- a new proposal to define the supports for the inefficiency error;
- different possibilities to define the supports with the GCE estimator;
- the possibility to include different orders of entropy with the GME- $\alpha$  estimators in the two-error component rather than using the same value for both;
- the evidence that the GME, GME- $\alpha$  and GCE estimators are powerful alternatives to the maximum likelihood (ML) estimator in the estimation of state-contingent production frontiers under severe empirical conditions, namely in
  - models with few observations in some states of nature (that strongly restrict the use of traditional estimators);
  - models with collinearity and severe collinearity problems.

Although the theory of state-contingent production is well-established, the empirical implementation of this approach is still in an infancy stage. The ME estimators are expected to make an important contribution to the increase of empirical work with state-contingent production frontiers in the near future. This is an important contribution since, for many authors, the state-contingent approach is the most complete procedure in the production literature that should be used to evaluate technical efficiency in the context of production uncertainty.

Chapter 4 presents original work in the ridge regression with an application of the GME estimator in the estimation of the ridge parameter. The idea is to introduce a new estimator for the ridge parameter, which efficiently combines the ridge trace and the GME estimator. The contributions of this chapter are

- the development of a new estimator, denoted as the Ridge-GME estimator, that combines the analysis of the ridge trace with the GME estimator;
- the discussion and comparison of the performance of the Ridge-GME estimator with several traditional competitors in a Monte Carlo simulation study and an empirical application to the well-known Portland cement data set.

The simulation study reveals that, in the case of regression models with small samples sizes affected by collinearity, the Ridge-GME estimator is probably one of the best ridge parameter

estimators available in the literature on ridge regression. This finding is very important for ridge regression users. It is important to note that the main challenge in the ridge regression is the selection of the ridge parameter and there is in the literature a huge number of methods to estimate this parameter! According to that finding, the Ridge-GME estimator can be recommended to practitioners and should belong to the restricted group of ridge parameter estimators that may be considered in any ridge regression analysis with small samples.

In Chapter 5, section 5.1, a third set of contributions of this thesis is presented, which may be considered an upgrade of the GME estimator or, more generally, a new approach to the ME estimation. In fact, a first step is already made by Paris [127] with the MEL estimator, based on the Shannon entropy, some concepts from the theory of light (quantum electrodynamics) and the OLS estimator. The contributions of this section are

- the introduction of the maximum entropy robust regression group (MERG) estimators, that represent a generalization of the MEL estimator, and the discussion of
  - the structure of the MERG estimators, which includes the Shannon, Rényi and Tsallis entropies, the OLS estimator and different estimators based on robust regression, namely the least trimmed squares (LTS), the least absolute deviations (LAD), and the least median of squares (LMS) estimators;
  - some properties, namely scale invariance, consistency and asymptotic normality for some MERG estimators;
- the evaluation and comparison of the performance of the MERG estimators with several traditional estimators in different simulation studies and in models with real data.

The MERG estimators may be a good choice in the estimation of linear regression models with small samples sizes affected not only by outliers and collinearity simultaneously, but also in models only affected by collinearity or outliers separately. The MERG estimators are easy to compute and, mostly important, no relevant prior information is needed to implement them. These two features are probably the most important ones of the MERG estimators.

Additional original work is provided in Chapter 5, section 5.2: an extension of the MERG estimators, denoted as MERGE estimators. This acronym is the initials of the words *maximum entropy robust regression group extended*, but it also reflects the objective of merging

different estimators in a new class with high performance in linear regression models with small samples sizes affected with collinearity and outliers. The contributions of this section are

- the introduction of the MERGE estimators, that are an extension of the MERG estimators developed in section 5.1, and the discussion of
  - the structure of the MERGE estimators which avoids the analogy with quantum electrodynamics and include supports for the parameters as in the GME estimator;
  - some potential improvements for the MERGE estimators, namely the possibility to impose parameter inequality restrictions through the parameter support matrix (as made in the GME estimator) and the use of the cross-entropy formalism;
- the evaluation and comparison of the performance of the MERGE estimators with the MERG estimators and a recent powerful estimator for the combined collinearity-outliers problem in linear regression.

This extension allows to include supports for the parameters as in the GME estimator since there are regression models where the supports for the parameters are known and provided by the theory (e.g., in economics estimating the marginal propensity to consume), or by the experience of the researchers. The simulation study reveals a good performance of the MERGE estimators in linear regression models with small samples sizes affected by collinearity and outliers.

Finally, Appendix A presents an illustration of the Shannon entropy as a measure of information in the context of two simple games with roulette wheels; Appendix B provides a short research on technical inefficiency with directional technology distance functions; and Appendix C includes MATLAB codes with some estimators presented in the thesis. In summary, the contributions in the appendices are

- two new measures of technical inefficiency with directional distance functions;
- new MATLAB codes for the ME estimation.

It is important to note that directional technology distance functions provide a complete representation of a firm's production technology. Moreover, a directional distance function itself

provides a natural technical inefficiency measure; see Chambers et al. [27]. The MATLAB codes in Appendix C are based on the ones used in this thesis, but they are particularly designed to users that are not familiar with the ME estimation. These and other codes, as well as new updates and additional information on the ME estimators will soon be available in <http://www.ua.pt/mat>. In order to disseminate the ME estimators to a wider audience, some of those codes for GAMS and Microsoft Excel will also be available in the same website.

As a final remark, five papers, based on this work, were prepared for publication (four in international scientific journals and one in a conference proceeding): Macedo et al. [103, 104, 105, 106]; one paper was recently submitted. Additionally, eleven presentations were given at national and international conferences, namely in the 27th and 28th European Meetings of Statisticians (two talks; one poster), the 17th, 18th, 19th and 20th Conferences of the Portuguese Statistical Society (five talks; one poster), the first Research Day in the University of Aveiro (one poster), and the 58th World Statistics Congress of the International Statistical Institute (one talk).



## Chapter 2

# Background and state of the art

“[...] the principle of maximum entropy is not an Oracle telling which predictions *must* be right; it is a rule for inductive reasoning that tells us which predictions *are most strongly indicated by our present information.*”

Jaynes [86, p. 369].

In this chapter, an overview on the main topics covered in the thesis is presented, with particular attention to entropy and ME estimation.<sup>1</sup> In section 2.1, some interpretations of entropy and the ME principle are discussed. Section 2.2 presents a brief review of some ME estimators, namely the GME, GCE, and GME- $\alpha$  estimators. The two remaining sections briefly review some diagnostic procedures and the strategies for dealing with collinearity and outliers, as well as some topics on technical efficiency analysis.

## 2.1 Entropy

### 2.1.1 Entropy concepts

The notion of *entropy* appears in the foundations of thermodynamics in the XIX century. This concept is introduced by Clausius [30], expressing a relationship between heat and tem-

---

<sup>1</sup>This thesis covers different topics from different areas, such as physics, econometrics, economics, statistics and mathematics, which causes a natural problem of inclusion. An additional effort is made so that this work is self-contained. Naturally, a complete background and state of the art on these topics can only be achieved by reviewing some seminal and relevant work which is mentioned throughout the thesis, where detailed definitions, properties, proofs, and some other historical details can be found.

perature in a physical system. Later, based on the work by Maxwell [110, 111] in the kinetic theory of gases, Ludwig Boltzmann, Josiah Gibbs and Max Plank are the main architects of statistical mechanics; see, for example, Gibbs [62]. From the work produced by these three authors, reflected in a large number of books and articles,<sup>2</sup> the following formulas of entropy emerge in the literature:

$$S = -k \sum_{i=1}^w p_i \log p_i \quad (2.1)$$

and

$$S = k \log w. \quad (2.2)$$

In the above expressions,  $S$  is the entropy,  $k$  is the Boltzmann constant,  $p_i$  is the probability of microstate  $i$  and  $w$  is the number of microstates for a given macrostate of the system. If the probabilities,  $p_i$ , are equal, equation (2.2) is obtained from (2.1). In equation (2.2), the entropy of a macrostate increases when the number of microstates increases, i.e., when the number of possible configurations of the atoms increases.

The entropy is also connected with the second law of thermodynamics, which generally states that the entropy of isolated systems always increases in order to achieve a maximum at the equilibrium. This law expresses an unmistakable reality of nature through the one way it provides for spontaneous processes. Due to this natural irreversibility imposed by the second law of thermodynamics, a famous “demon” appears in the literature: the Maxwell’s demon. This demon is an imaginary being that is able to contradict the second law of thermodynamics and accomplishes the impossible task of reducing the entropy in an isolated system.<sup>3</sup> Note that even with the small importance that this issue has been debated over the years, it has never been abandoned; e.g., Raizen [135].

The interpretation of entropy is still controversial nowadays. This is evident when a simple search is made on some scientific journals and several dozens of works are found that directly or indirectly debate this issue. For example, Styer [163, p. 1090] states that

“Of all the difficult concepts of classical physics – concepts like acceleration, energy, electric field, and time – the most difficult is entropy. Even von Neumann claimed that “nobody really knows what entropy is anyway.” [...] The

---

<sup>2</sup>Some of these works can be downloaded from <http://www.archive.org>. In the website of the School of Mathematics and Statistics in the University of St Andrews, Scotland, some biographies can be found.

<sup>3</sup>In <http://nautilus.fis.uc.pt/molecularium/pt/entropia/index.html>, an interesting game with the Maxwell’s demon can be found.

metaphoric images invoked for entropy include “disorder,” “randomness,” “smoothness,” “dispersion,” and “homogeneity.” In a posthumous fragment, Gibbs mentioned “entropy as mixed-up-ness.”

The terms “complexity” and “unavailable energy” are also usually used. Styer [163] suggests the use of both “disorder” and “freedom” to define entropy. This discussion is naturally beyond the scope of this thesis. The literature concerning entropy, thermodynamics and statistical mechanics is massive; see, among many others, Ben-Naim [11], Dugdale [42], Jaynes [86], Sethna [148], Styer [163, 164] and the references therein.

The concept of entropy presented above is strictly confined to physics. The Shannon entropy measure presented next have, in general, a different interpretation and represents a broader concept of entropy. It is interesting to note how Shannon [149, p. 10] introduces the problem:

“Suppose we have a set of possible events whose probabilities of occurrence [...] are known but that is all we know concerning which event will occur. Can we find a measure of how much “choice” is involved in the selection of the event or of how uncertain we are of the outcome?”

In the paper *A mathematical theory of communication*, Shannon [149] begins by defining three properties that such measure should satisfy.<sup>4</sup> These properties are usually the minimal axioms required for a consistent measure of the “amount of uncertainty”; see Jaynes [86, Chapter 11]. Consider  $H(p_1, p_2, \dots, p_K)$  as the measure to find, and  $p_1, p_2, \dots, p_K$  the probabilities of occurrence from a set of possible events.

**Axiom 2.1.**  $H(p_1, p_2, \dots, p_K)$  should be a continuous function of the  $p_k$ ,  $k = 1, 2, \dots, K$ .

**Axiom 2.2.** If all the  $p_k$ ,  $k = 1, 2, \dots, K$ , are equal, then  $H(p_1, p_2, \dots, p_K)$  should be a monotonic increasing function of  $K$ .

**Axiom 2.3.** If a choice is split into two successive choices, the original  $H(p_1, p_2, \dots, p_K)$  should be equal to the weighted sum of the individual values of  $H(p_1, p_2, \dots, p_K)$ .

---

<sup>4</sup>It is considered here only the case of discrete random variables. The entropy of a continuous distribution can be defined in a similar way considering probability density functions. The continuous case shares most of the properties of the discrete case, but not all of them; see Shannon [149, pp. 35–38] for further details.

Shannon [149] and Jaynes [86, Chapter 11] demonstrate that the only  $H(p_1, p_2, \dots, p_K)$  that satisfies Axioms 2.1–2.3 is given by equation (2.3), presented below.

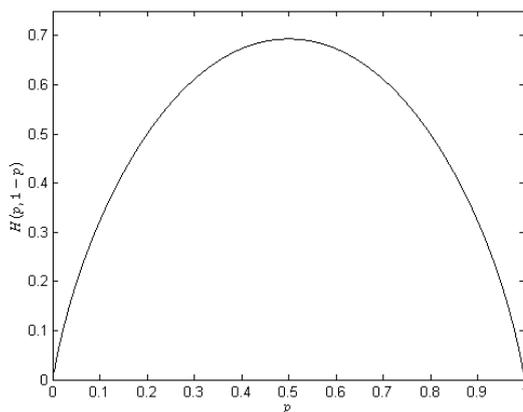
**Definition 2.1.** *The Shannon entropy measure<sup>5</sup> is given by*

$$H(p_1, p_2, \dots, p_K) = -c \sum_{k=1}^K p_k \ln p_k, \quad (2.3)$$

where  $c$  is a positive constant and  $p_k$ ,  $k = 1, 2, \dots, K$ , are the probabilities of occurrence from a set of possible events.

Figure 2.1 illustrates the Shannon entropy measure in the case of two possible outcomes with probabilities  $p$  and  $1 - p$ , where  $H(p, 1 - p) = -p \ln p - (1 - p) \ln(1 - p)$ , considering  $c = 1$ . Note that  $0 \ln 0 = 0$  is considered in Definition 2.1 if some  $p_k = 0$ , since  $\lim_{x \rightarrow 0} x \ln x = 0$  and the continuity of  $H(p_1, p_2, \dots, p_K)$  imposed by Axiom 2.1 is satisfied. It is important to note that the Shannon entropy represents an average logarithm of the probabilities  $p_k$ , and the events with the low or high probability have a small contribution to the entropy value; e.g., Golan and Perloff [68] and Holste et al. [79].

**Figure 2.1:** Shannon entropy.



Since the choice of  $c$  is just a matter of convenience and merely amounts to a choice of an unit of measure, it is usually considered that  $c = 1$ . Although, it is used the natural logarithm in Definition 2.1, it is possible to take any logarithm with any base greater than

<sup>5</sup>The notation is slightly changed here, but the expression remains the same. See Shannon [149, p. 11].

one (the constant  $c$  reflects these changes). As noted by Shannon [149, p. 1], the choice of the base corresponds to the choice of the units in which the information is measured; see the example in Appendix A for details.

With Claude Shannon the concept of entropy acquires a new meaning as a measure of information or uncertainty. In addition to Shannon, others pioneers in information theory deserve a mention here, namely Hartley [74], Nyquist [122, 123] and Wiener [176].

Shannon [149, pp. 11–13] presents six properties supporting the choice of (2.3) as a reasonable measure of information; see also Jaynes [86] and Khinchin [91]. Among these properties, there is one extremely important: for a given  $K$ ,  $H(p_1, p_2, \dots, p_K)$  reaches a maximum when all the  $p_k$  are equal. This is intuitively the most uncertain situation and, in this case,  $H(p_1, p_2, \dots, p_K) = c \ln K$ .

There are several stories in the literature on the choice of the name for the measure presented in Definition 2.1. It seems that Claude Shannon did not know which name should give to the new measure of information and John von Neumann was the one who suggested the name entropy, since there was already a similar expression used in statistical mechanics; see (2.1) presented previously. This story is told in Tribus and McIrvine [172, p. 180]:

“In 1961 one of us (Tribus) asked Shannon what he had thought about when he had finally confirmed his famous measure. Shannon replied: “My greatest concern was what to call it. I thought of calling it ‘information,’ but the word was overly used, so I decided to call it ‘uncertainty.’ When I discussed it with John von Neumann, he had a better idea. Von Neumann told me, ‘You should call it entropy, for two reasons. In the first place your uncertainty function has been used in statistical mechanics under that name, so it already has a name. In the second place, and more important, no one knows what entropy really is, so in a debate you will always have the advantage.’ ””

In Appendix A, the Shannon entropy as a measure of information is illustrated in the discussion of two simple games with roulette wheels. Several others interesting examples that illustrate the Shannon entropy as a measure of information can be found, among many others, in Brillouin [17] and Yaglom and Yaglom [178]. The Shannon entropy has acquired a remarkable importance as a measure of information that goes beyond the telecommunications

field where it was initially developed. Nowadays, the Shannon entropy measure is found in many areas of science.

Two of the most well-known generalizations of the Shannon entropy measure are presented next: the Rényi entropy and the Tsallis entropy.

Rényi [139, 140] generalizes the notion of random variable and defines an incomplete random variable, an incomplete probability distribution and a complete conditional distribution of the incomplete random variable. Note that if  $X$  is an incomplete random variable with values  $x_k$  and associated probabilities  $p_k > 0$ ,  $k = 1, 2, \dots, K$ , then  $\sum_{k=1}^K p_k \leq 1$ , and not necessarily  $\sum_{k=1}^K p_k = 1$ .

Rényi [140, pp. 570–574] defines the gain of information, denoted by  $I(Q\|P)$ , and presents six postulates that this measure must satisfy.  $I(Q\|P)$  represents the gain of information obtained when an incomplete distribution  $P = (p_1, p_2, \dots, p_K)$ , with  $p_k > 0, \forall k$ , of an incomplete random variable  $X$  is substituted by an incomplete distribution  $Q = (q_1, q_2, \dots, q_K)$ . Assuming that  $I(Q\|P)$  satisfies the six postulates already mentioned, Theorem 2.1 defines the gain of information (Rényi [140, p. 574]).

**Theorem 2.1.** (*Measure of order  $\alpha$  of the gain of information*). *There exists a real number  $\alpha \neq 1$  such that*

$$I(Q\|P) = I_\alpha(Q\|P) = \frac{1}{\alpha - 1} \ln \left( \frac{1}{\sum_{k=1}^K q_k} \sum_{k=1}^K \frac{(q_k)^\alpha}{(p_k)^{\alpha-1}} \right) \quad (2.4)$$

or

$$I(Q\|P) = I_1(Q\|P) = \frac{1}{\sum_{k=1}^K q_k} \sum_{k=1}^K q_k \ln \frac{q_k}{p_k}. \quad (2.5)$$

*Proof.* See Rényi [140, pp. 574–578]. □

Based on Theorem 2.1, the Rényi entropy can be established for a complete probability distribution (a distribution of an ordinary random variable).

**Definition 2.2.** The Rényi entropy measure<sup>6</sup> of order  $\alpha$  for a complete probability distribution is given by

$$H_\alpha^R(p_1, p_2, \dots, p_K) = \frac{1}{1 - \alpha} \ln \sum_{k=1}^K (p_k)^\alpha, \quad (2.6)$$

where  $\alpha \neq 1$  is a real number and  $p_k$  are the probabilities from the complete distribution.

An important relation between the Rényi entropy and the Shannon entropy is presented below.

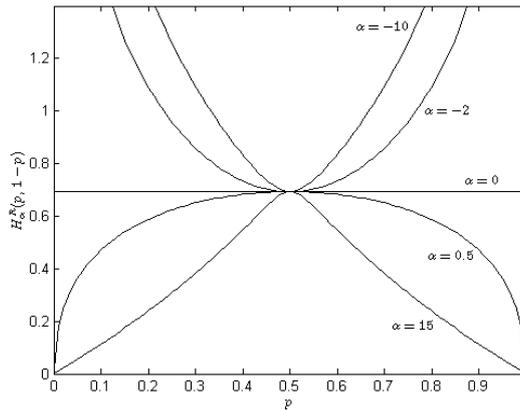
**Proposition 2.1.** The Rényi entropy reduces to the Shannon entropy when  $\alpha \rightarrow 1$ .

*Proof.* Using L'Hôpital's rule, it can be established that

$$\lim_{\alpha \rightarrow 1} \frac{\ln \sum_{k=1}^K (p_k)^\alpha}{1 - \alpha} = \lim_{\alpha \rightarrow 1} \frac{\frac{1}{\sum_{k=1}^K (p_k)^\alpha} \sum_{k=1}^K (p_k)^\alpha \ln p_k}{-1} = - \sum_{k=1}^K p_k \ln p_k. \quad (2.7)$$

□

**Figure 2.2:** Rényi entropy.



The Rényi entropy satisfies some of the properties of the Shannon entropy, namely standard additivity, non-negativity and both reach an extreme value when all probabilities are equal; see Curado and Tsallis [34], Rényi [140] and Tavares [167]. However, one distinctive

<sup>6</sup>The notation is slightly changed here, but the expression remains the same. See Rényi [140, p. 579].

characteristic of the Rényi entropy, when compared with the Shannon entropy, is the dependence on the real number  $\alpha$ , implying that the Rényi entropy is not always concave, as illustrated in Figure 2.2 for a complete distribution  $P = (p, 1 - p)$ .

As mentioned by Rényi [140, p. 581], the entropy measure presented in Definition 2.2 should be considered as a true measure of information only when  $\alpha$  is positive.<sup>7</sup> Furthermore, the Rényi entropy is defined as an average of probabilities  $p_k$  raised to powers of  $\alpha$ , rather than the average logarithm defined by the Shannon entropy. The value of  $\alpha > 1$  defines the relative contribution of event  $k$  to the entropy value and thus events with higher probability contribute more to the value of the entropy than the lower probability events; see, for example, Golan and Perloff [68] and Holste et al. [79].

Later, Tsallis [173] develops another entropy measure, that is defined next.

**Definition 2.3.** *The Tsallis entropy measure<sup>8</sup> is given by*

$$H_\alpha^T(p_1, p_2, \dots, p_K) = \frac{c}{\alpha - 1} \left( 1 - \sum_{k=1}^K (p_k)^\alpha \right), \quad (2.8)$$

where  $\alpha \neq 1$  is a real number,  $c$  is a positive constant (usually assumed  $c = 1$ ) and  $p_k$  are the probabilities of occurrence from a set of possible events (possible microscopic configurations of a system in the original context).

Figure 2.3 illustrates the Tsallis entropy measure presented in Definition 2.3 for  $K = 2$  and some values of  $\alpha$ . Considering only the case of  $\alpha > 1$  as in the Rényi entropy, the events with higher probability contribute more to the value of the Tsallis entropy than the lower probability events.

**Proposition 2.2.** *The Tsallis entropy reduces to the Shannon entropy when  $\alpha \rightarrow 1$ .*

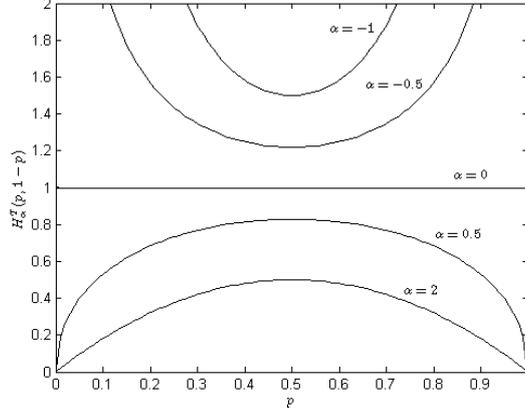
*Proof.* It can be easily established that

$$\lim_{\alpha \rightarrow 1} \frac{c}{\alpha - 1} \left( 1 - \sum_{k=1}^K (p_k)^\alpha \right) = c \lim_{\alpha \rightarrow 1} \frac{\sum_{k=1}^K p_k (1 - (p_k)^{\alpha-1})}{\alpha - 1} = -c \sum_{k=1}^K p_k \ln p_k, \quad (2.9)$$

which is the Shannon entropy measure; see Tavares [167, p. 62] for a detailed discussion.  $\square$

<sup>7</sup>For convenience, it is always considered, throughout this work, that  $\alpha > 1$ .

<sup>8</sup>The notation is slightly changed here, but the expression remains the same. See Tsallis [173, p. 479].

**Figure 2.3:** Tsallis entropy.

The Tsallis entropy shares some properties with the Shannon and Rényi entropies, namely it is non-negative and it reaches an extreme value when all the probabilities are equal. However, the standard additivity property satisfied by the Shannon and Rényi entropies is violated by the Tsallis entropy. One of the most important features of the Tsallis entropy is the pseudoadditivity; e.g., Abe [1], Curado and Tsallis [34], Santos [147] and Tsallis [173].

**Theorem 2.2.** (*Pseudoadditivity*). For two independent random variables  $A$  and  $B$ ,

$$H_{\alpha}^T(A, B) = H_{\alpha}^T(A) + H_{\alpha}^T(B) + (1 - \alpha)H_{\alpha}^T(A)H_{\alpha}^T(B). \quad (2.10)$$

*Proof.* With no loss of generality, it is assumed, as usually, that  $c = 1$ . Considering

$$H_{\alpha}^T(A) = \frac{1}{\alpha - 1} \left( 1 - \sum_{k_A=1}^{K_A} (p_{k_A})^{\alpha} \right) \quad \text{and} \quad H_{\alpha}^T(B) = \frac{1}{\alpha - 1} \left( 1 - \sum_{k_B=1}^{K_B} (p_{k_B})^{\alpha} \right), \quad (2.11)$$

since  $A$  and  $B$  are independent random variables, then,

$$H_{\alpha}^T(A, B) = \frac{1}{\alpha - 1} \left( 1 - \sum_{k_A=1}^{K_A} \sum_{k_B=1}^{K_B} (p_{k_A} p_{k_B})^{\alpha} \right). \quad (2.12)$$

Thus, it follows that

$$\begin{aligned}
H_\alpha^T(A, B) &= \frac{1}{\alpha - 1} \left( 1 - \sum_{k_A=1}^{K_A} (p_{k_A})^\alpha \right) + \frac{1}{\alpha - 1} \left( 1 - \sum_{k_B=1}^{K_B} (p_{k_B})^\alpha \right) - \\
&\quad - \frac{1}{\alpha - 1} \left( 1 - \sum_{k_A=1}^{K_A} (p_{k_A})^\alpha - \sum_{k_B=1}^{K_B} (p_{k_B})^\alpha + \sum_{k_A=1}^{K_A} (p_{k_A})^\alpha \sum_{k_B=1}^{K_B} (p_{k_B})^\alpha \right) \\
&= H_\alpha^T(A) + H_\alpha^T(B) + (1 - \alpha) H_\alpha^T(A) H_\alpha^T(B).
\end{aligned} \tag{2.13}$$

□

For  $\alpha \neq 1$ , the Tsallis entropy is a nonextensive measure where  $\alpha$  determines the degree of nonextensivity of a system; see, for example, Di Sisto et al. [37], Plastino and Plastino [131], Santos [147] and Suyari [165] for further details. Nonextensive statistical mechanics is nowadays a very impressive research field. The latest research conducted by the group of Prof. Constantino Tsallis, as well as a list with some thousands of published work related to nonextensive statistical mechanics can be accessed in the website of *Centro Brasileiro de Pesquisas Físicas*.<sup>9</sup>

It is also important to note that, in addition to the fact that the Rényi and Tsallis entropies are related to the Shannon entropy, the Rényi and Tsallis entropies are also related to each other (e.g., Curado and Tsallis [34]).

**Proposition 2.3.** *The Rényi entropy is related to the Tsallis entropy as follows:*

$$H_\alpha^R(p_1, p_2, \dots, p_K) = \frac{1}{1 - \alpha} \ln \left( 1 + (1 - \alpha) H_\alpha^T(p_1, p_2, \dots, p_K) \right).$$

*Proof.* Assuming  $c = 1$ , the relation trivially holds by substitution. □

It is important to note that the Shannon, Rényi and Tsallis entropies are not the only entropy measures in the literature. For example, Taneja [166] presents 25 different entropy expressions, where 24 of them have the Shannon entropy as a limit or a particular case. Taneja [166, p. 410] presents an interesting “entropy graph” that indicates how the 24 entropies are related to the Shannon entropy measure.

The literature on entropy and information theory is huge, both theoretical and applied; see, among many others, Ash [8], Brillouin [17], Dionísio et al. [38, 39], Galleani and Garello

<sup>9</sup>Information is available in <http://portal.cbpf.br>. See also <http://tsallis.cat.cbpf.br/TEMUCO.pdf>.

[59], Jaynes [86], Khinchin [91], Mana [107], Merhav [115], Rastegin [138], Saboia et al. [145], Vila et al. [174] and Yaglom and Yaglom [178]. Tavares [167] is an excellent reference in Portuguese language concerning mathematical issues of the Shannon, Rényi and Tsallis entropies, as well as the foundations of entropy.

### 2.1.2 Maximum entropy principle

Jaynes [86, p. 365] states that the maximum entropy (ME) principle is a simple and straightforward idea. This statement is explored in this subsection. Consider a pure linear inverse model defined as usually.

**Definition 2.4.** *A pure linear inverse model is stated as*

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta}, \quad (2.14)$$

where  $\mathbf{y}$  denotes a known  $(N \times 1)$  vector of observations,  $\boldsymbol{\beta}$  is a  $(K \times 1)$  vector of unknown parameters, and  $\mathbf{X}$  is a known  $(N \times K)$  matrix.

Assuming an exact relationship between the dependent variable and the independent variables, there is no error term in (2.14). Following Golan et al. [69], an ill-posed model is specified by considering that  $\mathbf{X}$  is a non-invertible matrix with  $N < K$ , and  $\boldsymbol{\beta} = \mathbf{p}$  is a vector of probabilities such that  $\sum_{k=1}^K p_k = 1$  and  $0 < p_k < 1$ , for  $k = 1, 2, \dots, K$ . From all the probability distributions that satisfy model (2.14), how can an unambiguous estimate of  $\mathbf{p}$  be chosen? The ME principle proposed by Jaynes [81, 82] provides an answer by choosing the distribution of probabilities that maximizes the Shannon entropy measure. Jaynes [81, p. 623] is clear:

“[...] in making inferences on the basis of partial information we must use that probability distribution which has maximum entropy subject to whatever is known. This is the only unbiased assignment we can make [...]”

A question arises: why the maximization of the Shannon entropy? In order to justify this, an approach based on the Wallis derivation<sup>10</sup> is considered since it leads to the maximization

---

<sup>10</sup>Suggestion made in 1962 by Graham Wallis to Edwin Jaynes. See Jaynes [86, p. 351].

of (2.3) without conditions and, maybe more important, the need for an interpretation of a measure of uncertainty; see Jaynes [86, p. 351] for further details.

Consider an experiment<sup>11</sup> with  $K$  possible outcomes that is repeated in  $N$  trials, and  $N_1, N_2, \dots, N_K$  the number of times that each outcome  $k$  occurs in the experiment, such that  $\sum_{k=1}^K N_k = N$  and  $N_k \geq 0$ . The number of ways a particular set of frequencies, say  $N_k = Np_k$ , can be realized is given by

$$W = \frac{N!}{N_1!N_2!\dots N_K!}, \quad (2.15)$$

known as the multinomial coefficient. Thus, the most probable set of frequencies (i.e., the set of frequencies that occurs in the greatest number of ways) must be chosen in order to maximize  $W$  or a monotonic function of  $W$ , such as

$$\ln W = \ln N! - \sum_{k=1}^K \ln N_k!. \quad (2.16)$$

As  $N \rightarrow \infty$ , it follows that  $N_k/N \rightarrow p_k$ , and using the Stirling's approximation, it can be found

$$\begin{aligned} \ln W &\approx N \ln N - N - \left( \sum_{k=1}^K N_k \ln N_k - \sum_{k=1}^K N_k \right) \\ &\approx N \ln N - \sum_{k=1}^K N_k \ln N_k \\ &\approx N \ln N - \sum_{k=1}^K N p_k \ln N p_k. \end{aligned} \quad (2.17)$$

Since

$$\sum_{k=1}^K N p_k \ln N p_k = \sum_{k=1}^K N_k \ln N + \sum_{k=1}^K N p_k \ln p_k, \quad (2.18)$$

it follows that

$$\ln W \approx -N \sum_{k=1}^K p_k \ln p_k, \quad (2.19)$$

which means that  $N^{-1} \ln W \approx -\sum_{k=1}^K p_k \ln p_k$ .<sup>12</sup> The set of frequencies that occurs in the greatest number of ways is just the one that maximizes the Shannon entropy measure. Following Jaynes [81, 86] and Golan et al. [69], by maximizing (2.3) subject to the limited available

<sup>11</sup>It is followed here a similar development to the one presented by Golan et al. [69, pp. 8–9]. The discussion of the Wallis derivation is provided in Jaynes [86, p. 351].

<sup>12</sup>Some simulations with marbles in cells illustrating this relationship are provided by Prof. Arieh Ben-Naim. Information is available in <http://www.ariiebennaim.com/books/discover.html>.

data in model (2.14), the most probable set of  $p_k$  that is consistent with the information available is obtained.<sup>13</sup> The importance of measure (2.3) in the ME principle is recognized by Jaynes [81, p. 622]:

“The great advance provided by information theory lies in the discovery that there is a unique, unambiguous criterion for the “amount of uncertainty” represented by a discrete probability distribution [...]”

The ME formalism is defined next using a matricial form.

**Definition 2.5.** *For a pure linear inverse model as stated in Definition 2.4, where  $\boldsymbol{\beta} = \mathbf{p}$  is a vector of probabilities, the ME formalism is defined as*

$$\operatorname{argmax}_{\mathbf{p}} \{ -\mathbf{p}' \ln \mathbf{p} \}, \quad (2.20)$$

*subject to the model (or data consistency) constraint,  $\mathbf{X}\mathbf{p} = \mathbf{y}$ , and the additivity (or normalization) constraint,  $\mathbf{1}'\mathbf{p} = 1$ , where  $\mathbf{1}$  is a  $(K \times 1)$  vector of ones, and  $\mathbf{p} > \mathbf{0}$  is a  $(K \times 1)$  vector of probabilities.*

The ME principle provides a tool to make the best prediction (i.e., the one that is the most strongly indicated) from the available information (and only this, since the introduction of any other subjective information is not recommended). If the entropy function in (2.20) is maximized without the model constraint, a solution from a uniform distribution is obtained. In this case, it is interesting to note that the ME principle can be seen as an extension of the Bernoulli’s principle of insufficient reason; see Jaynes [81, p. 623].

The analytical solution of the maximization problem specified in Definition 2.5 can be obtained using the traditional Lagrange multipliers method. Using the matricial form the Lagrangian function is given by

$$L(\mathbf{p}, \boldsymbol{\lambda}, \mu) = -\mathbf{p}' \ln \mathbf{p} + \boldsymbol{\lambda}'(\mathbf{y} - \mathbf{X}\mathbf{p}) + \mu(1 - \mathbf{1}'\mathbf{p}), \quad (2.21)$$

with the first-order optimality conditions

$$\frac{\partial L(\cdot)}{\partial \mathbf{p}} = -\ln \mathbf{p} - \mathbf{1} - \mathbf{X}'\boldsymbol{\lambda} - \mu\mathbf{1} = \mathbf{0}, \quad (2.22)$$

---

<sup>13</sup>See Jaynes [86, Chapter 11] for a complete overview about the ME principle.

$$\frac{\partial L(\cdot)}{\partial \boldsymbol{\lambda}} = \mathbf{y} - \mathbf{X}\mathbf{p} = \mathbf{0}, \quad (2.23)$$

$$\frac{\partial L(\cdot)}{\partial \mu} = 1 - \mathbf{1}'\mathbf{p} = 0. \quad (2.24)$$

Solving for  $p_k$  in terms of  $\boldsymbol{\lambda}$ , it follows that

$$\hat{p}_k = \frac{\exp(-\mathbf{x}'_k \hat{\boldsymbol{\lambda}})}{\sum_{k=1}^K \exp(-\mathbf{x}'_k \hat{\boldsymbol{\lambda}})}, \quad (2.25)$$

where  $\mathbf{x}_k$  is a  $(N \times 1)$  vector corresponding to the  $k$ th column of  $\mathbf{X}$  and  $\hat{\boldsymbol{\lambda}}$  is a  $(N \times 1)$  vector of estimated Lagrange multipliers on the model constraint. Equivalently, the solution can also be presented by

$$\hat{p}_k = \frac{\exp\left(-\sum_{n=1}^N x_{nk} \hat{\lambda}_n\right)}{\sum_{k=1}^K \exp\left(-\sum_{n=1}^N x_{nk} \hat{\lambda}_n\right)}. \quad (2.26)$$

Given the Lagrangian function and the first-order optimality conditions, the Hessian matrix is given by

$$\nabla^2(\mathbf{p}) = \begin{bmatrix} -\frac{1}{p_1} & 0 & \cdots & 0 \\ 0 & -\frac{1}{p_2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & -\frac{1}{p_K} \end{bmatrix}, \quad (2.27)$$

implying it is negative definite for  $0 < p_k < 1$  and a unique solution (a global maximum) of the ME formalism is ensured. The maximization problem in Definition 2.5 does not have a closed-form solution which means that the ME solution must be found with numerical optimization procedures. Finally, it is important to note that the Lagrange multipliers on the model constraint reflect the information contribution of each constraint to the objective function. For example, if some  $\lambda_n$  is zero this implies that the corresponding constraint has no ‘‘informational value’’ and does not reduce the maximum entropy value, i.e., the level of uncertainty; see Golan et al. [69, p. 26] for further details.

To illustrate the ME formalism a simple example is presented. Suppose that an experiment with three possible outcomes, 1, 2 and 3, is repeated for a large number  $N$  of times and the only available information from this  $N$  independent trials is the average of the outcomes, say

$y$ . For example, by assuming that  $y = 2.5$  what is the expected probability of outcome 1 in the  $N + 1$  trial of this experiment?

There are three unknowns ( $p_1, p_2$  and  $p_3$ ) and two constraints ( $p_1 + p_2 + p_3 = 1$  and  $p_1 + 2p_2 + 3p_3 = y$ ). Following the ME formalism from Definition 2.5, the solution is given by the probabilities that maximize

$$H(p_1, p_2, p_3) = -p_1 \ln p_1 - p_2 \ln p_2 - p_3 \ln p_3 \quad (2.28)$$

subject to

$$p_1 + 2p_2 + 3p_3 = 2.5 \quad (2.29)$$

and

$$p_1 + p_2 + p_3 = 1. \quad (2.30)$$

Since  $p_1$  is the only variable that matters, one can simply maximize the following entropy function

$$H(p_1) = -p_1 \ln p_1 - (-2p_1 + 0.5) \ln(-2p_1 + 0.5) - (p_1 + 0.5) \ln(p_1 + 0.5). \quad (2.31)$$

It follows that  $p_1 \approx 0.12$ . Suppose now that the average of outcomes from a large number  $N$  is  $y = 2$ . What is now the expected probability of outcome 1? It is expected that  $p_1 = 1/3$ , since  $y = 2$  is the mean of a discrete uniform distribution  $(1, 3)$ .

To explain this issue in more detail, the previous example is extended to the die problem presented in Golan et al. [69, p. 12], which is based, in turn, on the problems discussed by Jaynes [83, 86]. Knowing that the average outcome from a large number  $N$  of independent rolls of a die is  $y$ , the aim is to estimate the probability vector  $\mathbf{p} = (p_1, p_2, \dots, p_6)$ . Only with this information at hand (the average of the results), the ME principle can be applied to select the probability vector  $\mathbf{p}$  that maximizes

$$H(\mathbf{p}) = H(p_1, p_2, \dots, p_6) = - \sum_{k=1}^6 p_k \ln p_k \quad (2.32)$$

subject to the model constraint

$$\sum_{k=1}^6 k p_k = y \quad (2.33)$$

and the additivity constraint

$$\sum_{k=1}^6 p_k = 1. \quad (2.34)$$

**Table 2.1:** Estimated ME distributions for the die problem.

$y$	$H(\hat{\mathbf{p}})$	$\hat{\lambda}$	$\hat{p}_1$	$\hat{p}_2$	$\hat{p}_3$	$\hat{p}_4$	$\hat{p}_5$	$\hat{p}_6$
1.5	0.9534	1.0865	0.6637	0.2239	0.0755	0.0255	0.0086	0.0029
2.0	1.3675	0.6295	0.4781	0.2548	0.1356	0.0724	0.0385	0.0205
2.5	1.6136	0.3711	0.3475	0.2398	0.1654	0.1142	0.0788	0.0544
3.0	1.7485	0.1746	0.2468	0.2072	0.1740	0.1461	0.1227	0.1031
3.5	1.7918	$\approx 0$	0.1667	0.1667	0.1667	0.1667	0.1667	0.1667
4.0	1.7485	-0.1746	0.1031	0.1227	0.1461	0.1740	0.2072	0.2468
4.5	1.6136	-0.3711	0.0544	0.0788	0.1142	0.1654	0.2398	0.3475
5.0	1.3675	-0.6295	0.0205	0.0385	0.0724	0.1356	0.2548	0.4781
5.5	0.9534	-1.0865	0.0029	0.0086	0.0255	0.0755	0.2239	0.6637

Since  $H(\mathbf{p})$  is strictly concave in the interior of the additivity constraint set and the intersection of the model and the additivity constraint sets is non-empty for  $y \in (1, 6)$ , there is a unique solution to this ME problem. In Table 2.1 and Figure 2.4 the estimated ME distributions<sup>14</sup> for different values of  $y$  are presented. As expected, the maximum value of the objective function (2.32) occurs when  $y = 3.5$  and the estimated ME distribution is a uniform distribution. Moreover, the smallest value for  $\hat{\lambda}$  ( $-0.2328 \times 10^{-16} \approx 0$ ) is found when  $y = 3.5$ , which means that the constraint (2.33) with  $y = 3.5$  represents the case of complete ignorance, i.e., the constraint (2.33) with  $y = 3.5$  does not contribute to reduce the level of uncertainty. An interesting discussion concerning the original Jaynes' die problem can be found in Grendár Jr. and Grendár [72].

The formal solution is easily derived for the die problem. For  $k = 1, 2, \dots, 6$ , the Lagrangian function is given by

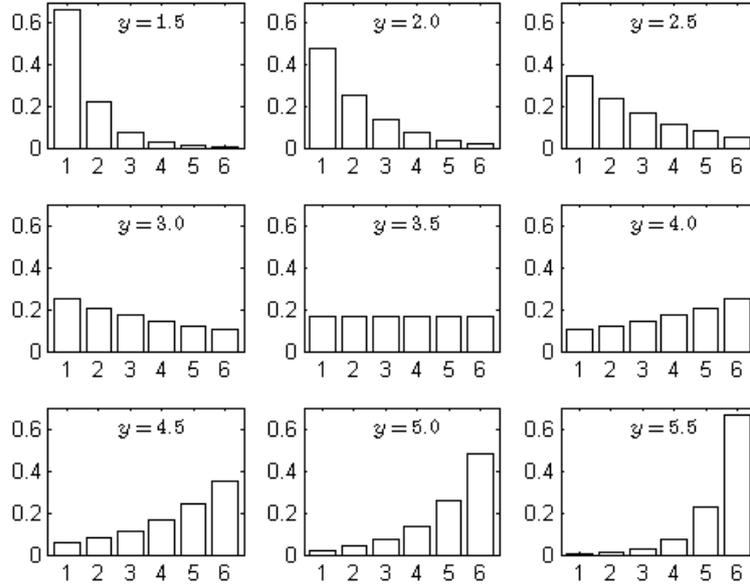
$$L(p_k, \lambda, \mu) = - \sum_{k=1}^6 p_k \ln p_k + \lambda \left( y - \sum_{k=1}^6 k p_k \right) + \mu \left( 1 - \sum_{k=1}^6 p_k \right), \quad (2.35)$$

with the first-order optimality conditions

$$\frac{\partial L(\cdot)}{\partial p_k} = - \ln p_k - 1 - k\lambda - \mu = 0, \quad (2.36)$$

$$\frac{\partial L(\cdot)}{\partial \lambda} = y - \sum_{k=1}^6 k p_k = 0, \quad (2.37)$$

<sup>14</sup>A MATLAB code, presented in Appendix C, can be used in this type of ME problems.

**Figure 2.4:** Estimated ME distributions for the die problem.

$$\frac{\partial L(\cdot)}{\partial \mu} = 1 - \sum_{k=1}^6 p_k = 0. \quad (2.38)$$

From (2.36), it follows, say, for  $k = 1$ ,

$$\mu = -\ln p_1 - 1 - \lambda, \quad (2.39)$$

and, by substitution,

$$p_k = p_1 e^{-(k-1)\lambda}, \quad (2.40)$$

for  $k = 2, 3, \dots, 6$ . Substituting each  $p_k$ ,  $k = 2, 3, \dots, 6$ , in (2.40) into (2.38), the value of  $p_1$  is determined. Then, substituting  $p_1$  into (2.40), the value of each  $p_k$ ,  $k = 2, 3, \dots, 6$ , is obtained. The formal solution to the die problem is then

$$\hat{p}_k = \frac{e^{-k\hat{\lambda}}}{\sum_{k=1}^6 e^{-k\hat{\lambda}}}, \quad (2.41)$$

which is, naturally, the formal solution (2.26) for  $N = 1$  and  $x_k = k$ , for  $k = 1, 2, \dots, 6$ .

At this point, and by looking to the ME principle described above, the next discussion from Golan [65, p. 60] provides an excellent synthesis:

“Two basic questions keep coming up in the literature: Is the ME principle “too simple?” and does the ME principle “produce something from nothing?” The answer to the above questions is contained in the simple explanation that, under the ME principle, only the relevant information is used, while all irrelevant details are eliminated from the calculations by an averaging process that averages over them. Therefore, it does not produce “something” from “nothing” but rather it only makes use of the available observed information where that information enters as constraints in the optimization.”

The principle of ME is often used for solving ill-posed problems, for example, in physics, informatics, linguistics, biology, medicine, communication engineering, statistics and economics. Some examples of applications of the ME principle can be found in Dionísio et al. [40], Golan and Dose [67], Miller and Horn [116], Park and Bera [130], Poletini [132], Preckel [133] and Vinod and López-de-Lacalle [175], among many others.

The work of Kullback [93, 94], Kullback and Leibler [95] and Lindley [100] were fundamental to connect the areas of ME and information theory with statistical inference. Furthermore, in the last years many authors, among which Bera and Biliás [13], Csiszár [33], Donoho et al. [41], Gamboa and Gassiat [60], Golan [63], Golan et al. [69], Jaynes [83, 84, 85, 86],<sup>15</sup> Judge and Mittelhammer [87], Levine and Tribus [99], Maasoumi [102], Shore and Johnson [154], Skilling [158, 159], Soofi [160], Soofi and Retzer [161] and Zellner [180, 181, 182] have developed efforts to make the ME more understandable (and also much less controversial) to a wider audience; see Golan [64, 65, 66] and the references therein for a guide in IEE. Recent developments can be followed at Info-Metrics Institute, at the American University, Washington.<sup>16</sup>

## 2.2 Estimators

### 2.2.1 The generalized maximum entropy estimator

The general linear regression model is used throughout this work. It is formalized next.

---

<sup>15</sup>In the website of the Washington University in St. Louis, it can be found other works of Edwin Jaynes. Information is available in <http://bayes.wustl.edu>.

<sup>16</sup>Information is available in <http://www.american.edu/cas/economics/info-metrics>.

**Definition 2.6.** *The general linear regression model is stated as*

$$\mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{u}, \quad (2.42)$$

where  $\mathbf{y}$  denotes a  $(N \times 1)$  vector of noisy observations,  $\boldsymbol{\beta}$  is a  $(K \times 1)$  vector of unknown parameters,  $\mathbf{X}$  is a known  $(N \times K)$  matrix of explanatory variables<sup>17</sup> and  $\mathbf{u}$  is a  $(N \times 1)$  vector of random disturbances (errors), usually assumed to have a conditional expected value of zero and representing spherical disturbances, i.e.,  $E[\mathbf{u}|\mathbf{X}] = \mathbf{0}$  and  $E[\mathbf{u}\mathbf{u}'|\mathbf{X}] = \sigma^2\mathbf{I}$ , where  $\mathbf{I}$  is a  $(N \times N)$  identity matrix and  $\sigma^2$  is the error variance.

As noted by Golan et al. [69], statistical data are frequently limited and affected by collinearity implying that the associated statistical models may be ill-posed, unless simplifying assumptions/procedures are imposed to generate seemingly well-posed statistical models, that can be estimated with traditional statistical tools. Giving heed to this problem, Golan et al. [69] generalized the ME formalism specified in Definition 2.5 to linear inverse problems with noise, expressed in Definition 2.6. The idea is to treat each  $\beta_k$  as a discrete random variable with a compact support and  $2 \leq M < \infty$  possible outcomes, and each  $u_n$  as a finite and discrete random variable with  $2 \leq J < \infty$  possible outcomes. Assuming that both the unknown parameters and the unknown error terms may be bounded *a priori*, the linear model in Definition 2.6 can be presented as

$$\mathbf{y} = \mathbf{X}\mathbf{Z}\mathbf{p} + \mathbf{V}\mathbf{w}, \quad (2.43)$$

where

$$\boldsymbol{\beta} = \mathbf{Z}\mathbf{p} = \begin{bmatrix} z'_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & z'_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & z'_K \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_K \end{bmatrix}, \quad (2.44)$$

with  $\mathbf{Z}$  a  $(K \times KM)$  matrix of support values and  $\mathbf{p}$  a  $(KM \times 1)$  vector of unknown weights (probabilities), and

$$\mathbf{u} = \mathbf{V}\mathbf{w} = \begin{bmatrix} v'_1 & \mathbf{0} & \cdots & \mathbf{0} \\ \mathbf{0} & v'_2 & \cdots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & v'_N \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{bmatrix}, \quad (2.45)$$

---

<sup>17</sup>This matrix usually contains a column of ones corresponding to the constant term. When this column does not exist (e.g., the matrix is centered and scaled), the dimension  $K$  is adjusted accordingly.

with  $\mathbf{V}$  a  $(N \times NJ)$  matrix of support values and  $\mathbf{w}$  a  $(NJ \times 1)$  vector of unknown weights (probabilities). By using the ME formalism, Golan et al. [69] propose the generalized maximum entropy (GME) estimator to select the unknown  $\mathbf{p}$  and  $\mathbf{w}$  vectors that maximize

$$H(\mathbf{p}, \mathbf{w}) = - \sum_{k=1}^K \sum_{m=1}^M p_{km} \ln p_{km} - \sum_{n=1}^N \sum_{j=1}^J w_{nj} \ln w_{nj} \quad (2.46)$$

subject to the model constraint

$$\sum_{k=1}^K \sum_{m=1}^M x_{nk} z_{km} p_{km} + \sum_{j=1}^J v_{nj} w_{nj} = y_n, \quad (2.47)$$

for  $n = 1, 2, \dots, N$ ; and the two additivity constraints,

$$\sum_{m=1}^M p_{km} = 1, \quad (2.48)$$

for  $k = 1, 2, \dots, K$ ; and

$$\sum_{j=1}^J w_{nj} = 1, \quad (2.49)$$

for  $n = 1, 2, \dots, N$ . The GME estimator is defined next using the matricial form.

**Definition 2.7.** *For the linear regression model specified in Definition 2.6, the GME estimator is given by*

$$\operatorname{argmax}_{\mathbf{p}, \mathbf{w}} \{ -\mathbf{p}' \ln \mathbf{p} - \mathbf{w}' \ln \mathbf{w} \} \quad (2.50)$$

subject to the model constraint

$$\mathbf{y} = \mathbf{XZ}\mathbf{p} + \mathbf{V}\mathbf{w}, \quad (2.51)$$

and the additivity constraints for  $\mathbf{p}$  and  $\mathbf{w}$ , respectively,

$$\begin{aligned} \mathbf{1}_K &= (\mathbf{I}_K \otimes \mathbf{1}'_M) \mathbf{p}, \\ \mathbf{1}_N &= (\mathbf{I}_N \otimes \mathbf{1}'_J) \mathbf{w}, \end{aligned} \quad (2.52)$$

where  $\otimes$  represents the Kronecker product,  $\mathbf{1}$  is a column vector of ones with a specific dimension,  $\mathbf{I}$  is an identity matrix with a specific dimension and, as defined in (2.44) and (2.45),  $\mathbf{Z}$  and  $\mathbf{V}$  are the matrices of supports, and  $\mathbf{p} > \mathbf{0}$  and  $\mathbf{w} > \mathbf{0}$  are probability vectors to be estimated.

The GME estimator generates the optimal probability vectors  $\hat{\mathbf{p}}$  and  $\hat{\mathbf{w}}$  that can be used to form point estimates of the unknown parameters and the unknown random errors through

the reparameterizations (2.44) and (2.45), respectively.<sup>18</sup> As noted by Golan et al. [69], since the objective function (2.50) is strictly concave in the interior of the additivity constraint set, a unique solution for the GME estimator is guaranteed if the intersection of the model and the additivity constraint sets is non-empty.

To formalize the GME solution, the Lagrangian function and the first-order optimality conditions are presented. The Lagrangian function,  $L(\mathbf{p}, \mathbf{w}, \boldsymbol{\lambda}, \boldsymbol{\mu}, \boldsymbol{\nu})$ , is given by

$$L(\cdot) = -\mathbf{p}' \ln \mathbf{p} - \mathbf{w}' \ln \mathbf{w} + \boldsymbol{\lambda}'(\mathbf{y} - \mathbf{XZp} - \mathbf{Vw}) + \boldsymbol{\mu}'(\mathbf{1}_K - (\mathbf{I}_K \otimes \mathbf{1}'_M)\mathbf{p}) + \boldsymbol{\nu}'(\mathbf{1}_N - (\mathbf{I}_N \otimes \mathbf{1}'_J)\mathbf{w}), \quad (2.53)$$

and the first-order optimality conditions are

$$\frac{\partial L(\cdot)}{\partial \mathbf{p}} = -\ln \mathbf{p} - \mathbf{1}_{KM} - \mathbf{Z}'\mathbf{X}'\boldsymbol{\lambda} - (\mathbf{I}_K \otimes \mathbf{1}_M)\boldsymbol{\mu} = \mathbf{0}, \quad (2.54)$$

$$\frac{\partial L(\cdot)}{\partial \mathbf{w}} = -\ln \mathbf{w} - \mathbf{1}_{NJ} - \mathbf{V}'\boldsymbol{\lambda} - (\mathbf{I}_N \otimes \mathbf{1}_J)\boldsymbol{\nu} = \mathbf{0}, \quad (2.55)$$

$$\frac{\partial L(\cdot)}{\partial \boldsymbol{\lambda}} = \mathbf{y} - \mathbf{XZp} - \mathbf{Vw} = \mathbf{0}, \quad (2.56)$$

$$\frac{\partial L(\cdot)}{\partial \boldsymbol{\mu}} = \mathbf{1}_K - (\mathbf{I}_K \otimes \mathbf{1}'_M)\mathbf{p} = \mathbf{0}, \quad (2.57)$$

$$\frac{\partial L(\cdot)}{\partial \boldsymbol{\nu}} = \mathbf{1}_N - (\mathbf{I}_N \otimes \mathbf{1}'_J)\mathbf{w} = \mathbf{0}. \quad (2.58)$$

Although in a higher dimension, the solution of the GME estimator is obtained analogously as the solution of the ME estimator. The formal solution of the GME estimator for  $\mathbf{p}$  is given by

$$\hat{p}_{km} = \frac{\exp\left(-z_{km}\mathbf{x}'_k\hat{\boldsymbol{\lambda}}\right)}{\sum_{m=1}^M \exp\left(-z_{km}\mathbf{x}'_k\hat{\boldsymbol{\lambda}}\right)}, \quad (2.59)$$

where  $\mathbf{x}_k$  is a  $(N \times 1)$  vector corresponding to the  $k$ th column of  $\mathbf{X}$  and  $\hat{\boldsymbol{\lambda}}$  is a  $(N \times 1)$  vector of estimated Lagrange multipliers on the constraint (2.51). The solution (2.59) can be presented alternatively as

$$\hat{p}_{km} = \frac{\exp\left(-z_{km} \sum_{n=1}^N x_{nk}\hat{\lambda}_n\right)}{\sum_{m=1}^M \exp\left(-z_{km} \sum_{n=1}^N x_{nk}\hat{\lambda}_n\right)}. \quad (2.60)$$

<sup>18</sup>A user friendly MATLAB code for the GME estimator is presented in Appendix C.

Being the GME estimator a generalization of the ME estimator, the similarity of these two equivalent expressions with (2.25) and (2.26) is not surprising. Finally, the formal solution for  $\mathbf{w}$  is given by

$$\hat{w}_{nj} = \frac{\exp(-v_{nj}\hat{\lambda}_n)}{\sum_{j=1}^J \exp(-v_{nj}\hat{\lambda}_n)}. \quad (2.61)$$

The GME estimator contributed to the development of the ME econometrics literature in the recent years. In view of the fact that the ill-posed real-world problems seem to be the rule rather than the exception in applied mathematics and statistics, the GME estimator has acquired special importance in the toolkit of statistical techniques, by allowing statistical formulations free of restrictive and unnecessary assumptions. In particular, this estimator is widely used in linear regression models in which (a) the design matrix is ill-conditioned (collinearity), (b) the number of unknown parameters exceeds the number of observations, and (c) in regression models with small samples sizes.

As mentioned previously, the supports in matrices  $\mathbf{Z}$  and  $\mathbf{V}$  are defined as being closed and bounded intervals within which each parameter or error is restricted to lie, implying that researchers need to provide exogenous information (which, unfortunately, it is not always available). This is considered the main weakness of the GME estimator; see, for example, Caputo and Paris [22] and Paris [128] for further details. Golan et al. [69] discuss these issues in the case of minimal prior information: for the unknown parameters, the authors recommend the use of wide bounds (this is naturally subjective) for the supports in  $\mathbf{Z}$ , without extreme risk consequences; for the unknown errors, the authors suggest the use of the three-sigma rule with a sample scale parameter.<sup>19</sup> Several simulation studies are provided by the authors to illustrate the stability of the GME estimates under different scenarios.

The number of points  $M$  and  $J$  in the supports is less controversial. Based on the experiments conducted by Golan et al. [69],  $M = 5$  and  $J = 3$  are usually used in the literature, since there is likely no significant improvement in the estimation with more points in supports. Naturally, as the number of points in the supports increases, the computational effort also increases.

---

<sup>19</sup>The supports in matrix  $\mathbf{V}$  are all equal with this procedure. However, there is always the possibility to choose different supports for each error, as illustrated in the model constraint (2.47).

Some properties of the GME estimator, such as consistency and asymptotic normality, are discussed in detail in Golan and Perloff [68] and Golan et al. [69, pp. 96–109]. Basic statistics for inference, including normalized entropy measures, asymptotic covariance and some statistical tests, are presented, for example, in Golan [65] and Golan et al. [69]. For the GME estimator, the asymptotic covariance matrix is given by

$$\text{Var}(\widehat{\boldsymbol{\beta}}) = \sigma^2(\mathbf{X}'\mathbf{X})^{-1}, \quad (2.62)$$

where an estimate of  $\sigma^2$  is given by

$$\widehat{\sigma}^2 = \frac{\frac{1}{N} \sum_{n=1}^N \widehat{\lambda}_n^2}{\left( \frac{1}{N} \sum_{n=1}^N (\widehat{\sigma}_{vn}^2)^{-1} \right)^2}, \quad (2.63)$$

and

$$\widehat{\sigma}_{vn}^2 = \sum_{j=1}^J v_{nj}^2 \widehat{w}_{nj} - \left( \sum_{j=1}^J v_{nj} \widehat{w}_{nj} \right)^2. \quad (2.64)$$

Since the GME estimator is frequently used in problems with limited information, researchers are usually more concerned with properties and basic statistics for inference in small samples sizes. Golan et al. [69, p. 108] use the asymptotic normality property to approximate the distribution of the GME estimator in small samples. Alternatively, the bootstrap method can be, in practice, a valuable tool to characterize the GME estimates; e.g., Campbell and Hill [21]. For details on bootstrap methods, see, for example, Davison and Hinkley [35] and Efron [43]. See also Amado and Pires [6] and Singh [157] for robust versions of bootstrap.

Some applications of the GME estimator can be found in Campbell et al. [20], Campbell and Hill [21], Ferreira et al. [55], Fraser [58], Lansink et al. [97], Lence and Miller [98], Paris and Howitt [129], Shen and Perloff [151], Tonini and Jongeneel [169], among many others.<sup>20</sup>

### 2.2.2 The generalized cross-entropy estimator

Considering the pure linear inverse problem specified in Definition 2.4, where  $\boldsymbol{\beta} = \mathbf{p}$  is a vector of probabilities, the cross-entropy formalism uses a vector of prior information about

<sup>20</sup>As already mentioned in section 2.1, recent work and developments can be accessed at the Info-Metrics Institute. Information is available in <http://www.american.edu/cas/economics/info-metrics>.

the unknown probabilities to obtain an estimate of  $\mathbf{p}$  which satisfies the constraints and is the one closest to the vector of prior information, i.e., the entropy distance between the data and the prior information is minimized, subject to the model and the additivity constraints.

**Definition 2.8.** *If the prior information about the unknown  $\mathbf{p}$  is defined in the form of a prior distribution of probabilities, a  $(K \times 1)$  vector  $\mathbf{q}$ , the cross-entropy solution of the pure linear inverse problem specified in Definition 2.4 is given by the minimization of*

$$\begin{aligned} H(\mathbf{p}, \mathbf{q}) &= \mathbf{p}' \ln(\mathbf{p}/\mathbf{q}) \\ &= \mathbf{p}' \ln \mathbf{p} - \mathbf{p}' \ln \mathbf{q}, \end{aligned} \tag{2.65}$$

subject to the model constraint,  $\mathbf{X}\mathbf{p} = \mathbf{y}$ , and the additivity constraint,  $\mathbf{1}'\mathbf{p} = 1$ .

In the cross-entropy formalism, the estimate  $\hat{\mathbf{p}}$  is obtained through the information provided by the data and the prior distribution of probabilities. Moreover, if  $\mathbf{q}$  represents a uniform distribution, the cross-entropy solution coincides with the ME solution in (2.26); see, for example, Golan et al. [69] and Shore and Johnson [154] for further details.

This approach can be generalized to the model specified in Definition 2.6.

**Definition 2.9.** *For the linear regression model specified in Definition 2.6, if  $\mathbf{q}_1$  is a  $(KM \times 1)$  vector with prior information about the unknown parameters  $\boldsymbol{\beta}$ , and  $\mathbf{q}_2$  is a  $(NJ \times 1)$  vector with prior information about the unknown errors  $\mathbf{u}$ , the generalized cross-entropy (GCE) estimator select the vectors  $\mathbf{p}$  and  $\mathbf{w}$  that minimize*

$$\begin{aligned} H(\mathbf{p}, \mathbf{w}, \mathbf{q}_1, \mathbf{q}_2) &= \mathbf{p}' \ln(\mathbf{p}/\mathbf{q}_1) + \mathbf{w}' \ln(\mathbf{w}/\mathbf{q}_2) \\ &= \mathbf{p}' \ln \mathbf{p} - \mathbf{p}' \ln \mathbf{q}_1 + \mathbf{w}' \ln \mathbf{w} - \mathbf{w}' \ln \mathbf{q}_2, \end{aligned} \tag{2.66}$$

subject to the model constraint

$$\mathbf{y} = \mathbf{XZ}\mathbf{p} + \mathbf{V}\mathbf{w}, \tag{2.67}$$

and the additivity constraints for  $\mathbf{p}$  and  $\mathbf{w}$ ,

$$\begin{aligned} \mathbf{1}_K &= (\mathbf{I}_K \otimes \mathbf{1}'_M)\mathbf{p}, \\ \mathbf{1}_N &= (\mathbf{I}_N \otimes \mathbf{1}'_J)\mathbf{w}, \end{aligned} \tag{2.68}$$

where  $\otimes$  represents the Kronecker product. (See also the GME estimator in Definition 2.7.)

Golan et al. [69, p. 90–92] present a detailed derivation of the GCE estimator and show that the Hessian matrix  $\nabla^2(\mathbf{p}, \mathbf{w})$  is positive definite for  $\mathbf{0} < \mathbf{p}, \mathbf{w} < \mathbf{1}$ , which guarantees

that the GCE solution is a unique global minimum. Note that this finding also applies to the GME estimator, since this estimator is a particular case of the GCE when the prior information (vectors  $\mathbf{q}_1$  and  $\mathbf{q}_2$ ) is expressed as a uniform distribution. Further details on the GCE estimator can be found in Golan [65] and Golan et al. [69].

### 2.2.3 Higher-order entropy estimators

Golan and Perloff [68] introduce two more general estimation methods based on the GME estimator: the GME- $\alpha$  estimators. These estimators are a generalization of the GME estimator, where the Shannon entropy in the objective function (2.50) is replaced by the Rényi (see Definition 2.2) or Tsallis entropies (see Definition 2.3).

**Definition 2.10.** *Considering the linear regression model specified in Definition 2.6 and  $c = 1$  in Definition 2.3, the GME- $\alpha$  estimator with the Tsallis entropy selects  $\mathbf{p}$  and  $\mathbf{w}$  that maximize*

$$\frac{1}{\alpha_1 - 1} \left( 1 - \sum_{k=1}^K \sum_{m=1}^M (p_{km})^{\alpha_1} \right) + \frac{1}{\alpha_2 - 1} \left( 1 - \sum_{n=1}^N \sum_{j=1}^J (w_{nj})^{\alpha_2} \right), \quad (2.69)$$

subject to the model constraint

$$y_n = \sum_{k=1}^K \sum_{m=1}^M x_{nk} z_{km} p_{km} + \sum_{j=1}^J v_{nj} w_{nj}, \quad (2.70)$$

and the additivity constraints,

$$\sum_{m=1}^M p_{km} = 1 \quad \text{and} \quad \sum_{j=1}^J w_{nj} = 1, \quad (2.71)$$

for all  $k$  and  $n$ , respectively. The parameters  $\alpha_1$  and  $\alpha_2$  are the order of the Tsallis entropy.

**Definition 2.11.** *Considering the linear regression model specified in Definition 2.6, the GME- $\alpha$  estimator with the Rényi entropy selects the vectors  $\mathbf{p}$  and  $\mathbf{w}$  that maximize*

$$\frac{1}{1 - \alpha_1} \ln \sum_{k=1}^K \sum_{m=1}^M (p_{km})^{\alpha_1} + \frac{1}{1 - \alpha_2} \ln \sum_{n=1}^N \sum_{j=1}^J (w_{nj})^{\alpha_2}, \quad (2.72)$$

subject to the model constraint

$$y_n = \sum_{k=1}^K \sum_{m=1}^M x_{nk} z_{km} p_{km} + \sum_{j=1}^J v_{nj} w_{nj}, \quad (2.73)$$

and the additivity constraints,

$$\sum_{m=1}^M p_{km} = 1 \quad \text{and} \quad \sum_{j=1}^J w_{nj} = 1, \quad (2.74)$$

for all  $k$  and  $n$ , respectively. The parameters  $\alpha_1$  and  $\alpha_2$  are the order of the Rényi entropy.

Following the same procedure as for the GME estimator, it can be shown, from the first- and second-order conditions of each method, that the Hessian matrix is negative definite implying that the solution, if it exists, is unique; see Golan and Perloff [68] for details.

Based on the axiomatic approaches to the classical ME principle discussed by Csiszár [33], Shore and Johnson [154] and Skilling [159], Golan and Perloff [68, pp. 202–203] define five axioms that represent a set of desirable properties for a consistent method of inference from a finite data set. These five consistency axioms can be stated informally as:

**Axiom 2.4.** *The solution should be unique.*

**Axiom 2.5.** *The choice of a coordinate system should not matter.*

**Axiom 2.6.** *The solution should remain the same if no additional information is available.*

**Axiom 2.7.** *The information contained in one subset of the data should not affect the solution obtained from the other subset if these two subsets are independent.*

**Axiom 2.8.** *The same solution should be obtained using information of independent systems separately in terms of their different densities functions or together in terms of their joint density function.*

Golan and Perloff [68] show that, under the traditional convexity assumptions, the GME estimator is the only one that satisfies Axioms 2.4–2.8, whereas the GME- $\alpha$  with the Tsallis entropy violates Axiom 2.8 and the GME- $\alpha$  with the Rényi entropy violates Axiom 2.7. The violation of Axiom 2.8 by the GME- $\alpha$  with the Tsallis entropy is expected, taking into account Theorem 2.2. Although not clearly stated in the proof of Golan and Perloff [68], the GME- $\alpha$  estimator with the Rényi entropy violates Axiom 2.7 because the Rényi entropy does not satisfy the Shannon additivity property; e.g., Curado and Tsallis [34].

Although each estimator violates one of the desirable properties, the GME- $\alpha$  estimators can be useful in models with small samples sizes and models affected by collinearity or outliers.

In a set of experiments conducted by Golan and Perloff [68], some GME- $\alpha$  estimators (for different values of  $\alpha$ ) have a lower mean squared error (MSE) than the GME estimator. Undoubtedly, the GME- $\alpha$  estimators deserve further investigation.

## 2.3 Regression diagnostics: collinearity and outliers

Collinearity and outliers hamper the empirical work with regression analysis. Despite the methodologies available in the literature to detect and deal with both collinearity and outliers, these problems are sometimes ignored in practice or frequently circumvented by simply removing variables that cause collinearity (can these variables be dispensed from the model?) and/or removing outliers (can these values be removed?).<sup>21</sup>

Collinearity is the term usually used in the literature to represent a near-linear relationship between two or more regressors. Other terms are also used to represent near exact dependencies among regressors namely multicollinearity or ill conditioning; e.g., Belsley et al. [10, p. 85]. Collinearity inflates the variance associated with the estimated regression coefficients and, thus, it may affect the signs of the estimated coefficients, as well as statistical inference.<sup>22</sup> Large variances in the OLS estimates can be easily explained using the linear regression model (2.42) with two regressors. Considering that the  $\mathbf{X}$  matrix is centered and scaled to have unit length, the correlation matrix is

$$\mathbf{X}'\mathbf{X} = \begin{bmatrix} 1 & r_{12} \\ r_{12} & 1 \end{bmatrix}, \quad (2.75)$$

where  $r_{12}$  is the correlation coefficient between the two explanatory variables.<sup>23</sup> It is important to note that the analysis of the correlation matrix can indeed detect the presence of collinearity, yet low correlation coefficients do not mean the absence of collinearity, since this matrix only identifies linear dependency between pairs of regressors. It is always possible that three or more regressors may generate collinearity and any pair of regressors taken alone have a large correlation coefficient.

The inverse of the correlation matrix (2.75) is given by

---

<sup>21</sup>Considering that all the variables are correctly added to the regression model and the outliers are truly values that belong to the regression model, the “elimination strategy” should be avoided.

<sup>22</sup>Collinearity also causes computational problems (e.g., numerical instability in matrix inversion).

<sup>23</sup>As noted by Montgomery et al. [119], these elements are usually called simple correlations between the regressors, although the term “correlation” may not be appropriate unless the regressors are random variables.

$$(\mathbf{X}'\mathbf{X})^{-1} = \begin{bmatrix} 1 & -r_{12} \\ \frac{1-r_{12}^2}{1-r_{12}^2} & \frac{-r_{12}}{1-r_{12}^2} \\ \frac{-r_{12}}{1-r_{12}^2} & \frac{1}{1-r_{12}^2} \end{bmatrix}, \quad (2.76)$$

where the elements of the main diagonal are the variance inflation factors (VIF), sometimes used to detect the presence of collinearity. Thus, in this case,

$$\text{Var}(\hat{\beta}_1) = \text{Var}(\hat{\beta}_2) = \frac{\sigma^2}{1-r_{12}^2}, \quad (2.77)$$

which means that when the correlation coefficient approaches one, reflecting an increasing collinearity between the two regressors, the variances of the parameter estimates can be substantially large. This explanation can be generalized to the case of more than two regressors, where the elements of the main diagonal of the  $(\mathbf{X}'\mathbf{X})^{-1}$  matrix are given by

$$\text{VIF}_k = \frac{1}{1-R_k^2}, \quad (2.78)$$

where  $R_k^2$  is the multiple correlation coefficient from the regression of each explanatory variable on the remaining explanatory variables.

As mentioned above, the VIF are usually used to detect collinearity and available in the majority of the statistical software. There are, however, other approaches, where the most important ones are based on the numerical analysis literature.<sup>24</sup> These approaches are based on the singular value decomposition.

**Definition 2.12.** *The singular value decomposition of a  $(N \times K)$  matrix  $\mathbf{X}$  is given by*

$$\mathbf{X} = \mathbf{U}\mathbf{S}\mathbf{V}' \quad (2.79)$$

where the  $(N \times K)$  matrix  $\mathbf{U}$  and the  $(K \times K)$  matrix  $\mathbf{V}$  are orthogonal,  $\mathbf{U}'\mathbf{U} = \mathbf{I}$  and  $\mathbf{V}'\mathbf{V} = \mathbf{I}$ , and  $\mathbf{S}$  is a  $(K \times K)$  diagonal matrix containing the singular values of  $\mathbf{X}$ .

Using the notion of a generalized inverse and Definition 2.12, the well-known condition number<sup>25</sup> can be extended to any rectangular matrix.

**Definition 2.13.** *The 2-norm condition number of a  $(N \times K)$  matrix  $\mathbf{X}$ , denoted as  $\text{cond}_2 \mathbf{X}$ , is given by the ratio of the largest singular value of  $\mathbf{X}$  to the smallest singular value.*

<sup>24</sup>This should not be surprising since collinearity is a problem of the data.

<sup>25</sup>In linear systems of equations,  $\mathbf{A}\mathbf{x} = \mathbf{b}$ , with  $\mathbf{A}$  a square and non-singular matrix, the condition number, defined as  $\|\mathbf{A}\|\|\mathbf{A}^{-1}\|$ , measures the sensitivity of the solution  $\mathbf{x}$  to small perturbations in  $\mathbf{A}$  or  $\mathbf{b}$ .

The 2-norm condition number is always a value greater than or equal to one (it approaches one in a matrix with orthonormal columns). Naturally, not all collinearity problems are potentially harmful. The higher the value of the 2-norm condition number, the more harmful is collinearity. Other approaches based on Definitions 2.12 and 2.13, such as the condition indices or the variance decomposition proportions, are also used in the literature; e.g., Belsley et al. [10], Montgomery et al. [119] and Ryan [144]. In Chapters 3, 4 and 5, the harmfulness of collinearity is evaluated by the 2-norm condition number and by the comparison with the results from the OLS or the ML estimators.

Once collinearity is detected, the next question is obviously: how to deal with collinearity? Unfortunately, the answer is not straightforward. Dealing with collinearity (and/or outliers) is often considered a mixture of art and science! Elimination of regressors (or data points) identified with collinearity, collection of new regressors, specification of new models (eliminating or including interaction effects), the use of principal component analysis, and the use of shrinkage estimators<sup>26</sup> (methods of regularization) are some of the strategies, among others, usually used in the literature. While the use of shrinkage estimators is generally well accepted, the choice of a shrinkage estimator is controversial; e.g., Hastie et al. [75] and Zou and Hastie [185]. In this work, in addition to the ME estimators, particular attention is given to ridge regression; see Chapter 4 for details.

Outliers are atypical observations being often influential observations that can produce a large impact on the OLS parameter estimates. Naturally, not all outliers are harmful. There are cases where the outliers do not affect the OLS estimation and are an important source of information concerning the problem under study; e.g., Belsley et al. [10], Maronna et al. [109] and Rousseeuw and Leroy [142].

Outlier detection is a broad topic. The usual outlier diagnostics are based on the analysis of the residuals from regression methods, and the analysis of the OLS estimates based on the original sample and different subsamples, where each observation (or groups of observations) is (are) sequentially ignored from the original sample. Traditional diagnostics available in statistical software are the Cook's squared distance, difference in betas (DFBETAS), difference in prediction (DFFITs) and the precision of the estimation using the covariance matrix (COVRATIO), among others.

---

<sup>26</sup>Note that the GME, GCE and GME- $\alpha$  estimators can be viewed as shrinkage estimators; e.g., Golan [65].

The Cook's squared distance<sup>27</sup> is usually interpreted as a measure of the difference between the estimated parameter vector using the original sample and the estimated parameter vectors using different subsamples, where each observation is sequentially omitted (usually, observations with a value of the Cook's squared distance greater than one should be investigated). DFFITS is closely related to the Cook's squared distance, measuring the impact on prediction from the elimination of a particular observation from the estimation (usually, observations with absolute values of DFFITS greater than  $2\sqrt{K/N}$  should be investigated). DFBETAS measures the changes in the  $\hat{\beta}_k$  regression coefficients when a specific observation is omitted (usually, observations with absolute values of DFBETAS greater than  $2/\sqrt{N}$  should be investigated). Finally, the COVRATIO can be interpreted as a measure of the precision of the estimation when a specific observation is omitted (usually, for  $N > 3K$ , observations with absolute values of  $(\text{COVRATIO}-1)$  greater than  $3K/N$  should be investigated).<sup>28</sup>

If influential observations are found in the sample, the next step is how to deal with them. Excluding the elimination of those observations, the solution is provided by the robust regression literature; e.g., Huber and Ronchetti [80] and Maronna et al. [109]. Given the large number of methods in the literature, this brief review is limited to the three robust methods applied in Chapter 5: the least absolute deviations (LAD), the least trimmed squares (LTS) and the least median of squares (LMS) estimators. Considering model (2.42), it is well-known that the OLS method generates the estimates of  $\beta_k$  by minimizing the sum of the squared residuals. LAD, LMS and LTS estimators can be defined similarly; e.g., Rousseeuw [141].

**Definition 2.14.** *For the linear regression model specified in Definition 2.6, LAD, LMS and LTS estimators are determined, respectively, by the minimization of*

$$\sum_{n=1}^N |r_n|, \quad \text{med}_n r_n^2, \quad \text{and} \quad \sum_{n=1}^h r_{(n:N)}^2, \quad (2.80)$$

where  $r_n$  represents the  $n$ th residual,  $r_{(1:N)}^2 \leq r_{(2:N)}^2 \leq \dots \leq r_{(N:N)}^2$  represents the ordered squared residuals, and  $h = [(1 - \rho)N] + 1$  with  $\rho$  being a trimming proportion and  $[\cdot]$  the integer portion of  $(1 - \rho)N$ .

---

<sup>27</sup>There are other possible interpretations of the Cook's squared distance, namely as the effect on fitted values resulting from the elimination of observations. This measure can be also generalized to groups of observations, i.e., groups of observations that are sequentially eliminated from the estimation (many-outlier detection procedure).

<sup>28</sup>Note that all of these cutoff values (mentioned in brackets) are the usual suggested values. Yet other cutoff values can be found in the literature; e.g., Belsley et al. [10], Rousseeuw and Leroy [142] and Ryan [144].

The LAD estimator is robust to outliers in the  $y$ -direction (outliers in the response variable), but it is sensitive to outliers in the  $x$ -direction (outliers in the explanatory variables). The breakdown point of the LAD estimator is 0%, exactly the same as for the OLS estimator, which means that the presence of a single outlier is sufficient for this estimator to produce arbitrary estimates. Both the LMS and the LTS estimators are robust to outliers in the  $x$ - and  $y$ -direction, and both of them have a breakdown point of 50% (the highest possible value).<sup>29</sup> Other properties, such as equivariance, asymptotic and finite-sample efficiency, and convergence rates, can be found in Maronna et al. [109] and Rousseeuw and Leroy [142].

Finally, it is important to note that the interest on the LAD, LMS and LTS estimators in this work is not on their stand-alone applications, but in their use in a general framework with maximum entropy components; see Chapter 5 for details.

## 2.4 Technical efficiency

Technical efficiency measures the ability to minimize the quantity of input(s) used in the production of output(s) given the production technology (input technical efficiency), or the ability to maximize the quantity of output(s) produced with a quantity of input(s) given the production technology (output technical efficiency). Generally speaking, output technical efficiency can be measured as the ratio of the output produced by a production unit (e.g., a firm) to its potential output; see, for example, Kumbhakar and Lovell [96, pp. 42–50] for a detailed discussion. Different methodologies to measure technical efficiency are available in the economic literature; e.g., Kalirajan and Shand [89] and Kumbhakar and Lovell [96].

In what follows, a production technology is formally defined and the notion of output technical efficiency is illustrated for the case of a single output. The two most common methods used in the efficiency literature are then presented. Finally, the state-contingent approach to production under uncertainty is briefly discussed.

**Definition 2.15.** *A production technology is represented by*

$$T = \{(\mathbf{x}, \mathbf{y}) : \mathbf{x} \text{ can produce } \mathbf{y}\}, \quad (2.81)$$

---

<sup>29</sup>For the LTS estimator, the breakdown point of 50% depends on  $h$ ; e.g., Rousseeuw and Leroy [142, p. 132]. For instance, note that, if  $h = N$ , the breakdown point is 0%.

where  $\mathbf{x} = (x_1, x_2, \dots, x_N) \in \mathbb{R}_+^N$  is a non-negative vector of inputs,  $\mathbf{y} = (y_1, y_2, \dots, y_M) \in \mathbb{R}_+^M$  is a non-negative vector of outputs, and  $T \subseteq \mathbb{R}_+^N \times \mathbb{R}_+^M$  is the production possibilities set.

Some regularity conditions for the production technology can be found in Appendix B.

**Definition 2.16.** An input set of the production technology  $T$  is given by

$$L(\mathbf{y}) = \{\mathbf{x} : (\mathbf{x}, \mathbf{y}) \in T\}, \quad (2.82)$$

that represents the set of the input vectors that can produce at least the output vector  $\mathbf{y} \in \mathbb{R}_+^M$ .

**Definition 2.17.** An output set of the production technology  $T$  is given by

$$P(\mathbf{x}) = \{\mathbf{y} : (\mathbf{x}, \mathbf{y}) \in T\}, \quad (2.83)$$

that represents the set of the output vectors that can be produced with the input vector  $\mathbf{x} \in \mathbb{R}_+^N$ .

Given a production technology, represented by  $T$ ,  $L(\mathbf{y})$  or  $P(\mathbf{x})$ , the definition of a production frontier is presented next for the case of a single output,  $y$ .

**Definition 2.18.** A production frontier, in the case of a single output,  $y$ , is given by

$$f(\mathbf{x}) = \max \{y : \mathbf{x} \in L(y)\} = \max \{y : y \in P(\mathbf{x})\} = \max \{y : (\mathbf{x}, y) \in T\}, \quad (2.84)$$

where  $L(y)$  is the input set,  $P(\mathbf{x})$  is the output set and  $T$  is the production possibilities set.

Properties of the production frontier can be found in Kumbhakar and Lovell [96, p. 26]. Output technical efficiency can be measured as the distance of the (observed) output produced by a production unit to the production frontier. Shephard [152, 153] develops the output distance function that provides a measure of output technical efficiency. For an overview of Shephard distance functions, see Färe and Primont [50].

**Definition 2.19.** An output distance function is given by

$$D(\mathbf{x}, y) = \min \left\{ \mu : \frac{y}{\mu} \in P(\mathbf{x}) \right\}, \quad (2.85)$$

where  $P(\mathbf{x})$  is the output set and  $D(\mathbf{x}, y) \leq 1$ .

If  $D(\mathbf{x}, y) = 1$ , the production unit is output technically efficient. If  $D(\mathbf{x}, y) < 1$ , the production unit is output technically inefficient. An inefficient production unit produces

only  $y$  given  $\mathbf{x}$ , whereas its potential output is  $y/D(\mathbf{x}, y)$ . Properties of the output distance function, as well as the definition of an input distance function, can be found in Kumbhakar and Lovell [96, pp. 28–32].

**Proposition 2.4.** *In a single output scenario, the production frontier specified in Definition 2.18 is related to the output distance function specified in Definition 2.19 through*

$$D(\mathbf{x}, y) = \frac{y}{f(\mathbf{x})} \leq 1. \quad (2.86)$$

Proposition 2.4 can be used as a measure of output technical efficiency; e.g., Debreu [36] and Farrell [54]. For other measures of technical efficiency, see Briec [16], Chambers et al. [26, 27], Färe and Lovell [49], Färe et al. [51, 52] and Zieschang [184].<sup>30</sup>

Several methods to estimate technical efficiency are available in the efficiency literature. These methods are usually distinguished as parametric or non-parametric, and stochastic or non-stochastic. The data envelopment analysis (DEA) and the stochastic frontier analysis (SFA) are the most dominant methods in the literature.

The DEA method, proposed by Charnes et al. [28], is based on the previous work of Afriat [2], Boles [14], Bressler [15], Farrell [54], among others. DEA uses linear programming to construct a non-parametric piece-wise linear production frontier, using different return to scales, and the possibility of multiple inputs and multiple outputs. Since DEA does not account for noise, all deviations from the production frontier are estimated as technical inefficiency. The estimation of output technical efficiency is presented in the context of a DEA model considering variable returns to scale; see Coelli et al. [31, p. 158].

**Definition 2.20.** *DEA output technical efficiency estimates can be generated by*

$$\max_{\theta, \lambda} \theta \quad (2.87)$$

subject to

$$\begin{aligned} -\theta \mathbf{y}_n + \mathbf{Y} \boldsymbol{\lambda} &\geq \mathbf{0}, \\ \mathbf{x}_n - \mathbf{X} \boldsymbol{\lambda} &\geq \mathbf{0}, \\ (\mathbf{N} \mathbf{1})' \boldsymbol{\lambda} &= 1, \\ \boldsymbol{\lambda} &\geq \mathbf{0}, \end{aligned} \quad (2.88)$$

---

<sup>30</sup>Two new measures of technical inefficiency are proposed in Appendix B.

where  $N$  is the number of producers ( $n = 1, 2, \dots, N$ ),  $\mathbf{X}$  is the matrix of inputs,  $\mathbf{Y}$  is the matrix of outputs,  $\mathbf{x}_n$  and  $\mathbf{y}_n$  are column vectors from  $\mathbf{X}$  and  $\mathbf{Y}$ , respectively, for the  $n$ th producer,  $1/\theta$  is a firm-specific technical efficiency estimate,  $0 < 1/\theta \leq 1$ , and  $\boldsymbol{\lambda}$  is the intensity vector.

It is important to note that the problem (2.87)–(2.88) is solved  $N$  times, i.e., once for each producer. Given its popularity, there is a wide range of software to estimate technical efficiency with DEA, making it easily accessible to practitioners. The literature on DEA is massive; see, for example, Charnes et al. [28] and Cooper et al. [32] for a brief review.

Aigner et al. [3], Battese and Corra [9] and Meeusen and van den Broeck [114] were the pioneers of the SFA methodology. The model proposed by these authors is defined next.

**Definition 2.21.** *The general stochastic frontier model is given by*

$$\ln(y_n) = f(\mathbf{x}_n, \boldsymbol{\beta}) + v_n - u_n, \quad (2.89)$$

where  $N$  is the number of producers ( $n = 1, 2, \dots, N$ ),  $f(\cdot)$  is the production frontier,  $y_n$  is the scalar output for producer  $n$ ,  $\mathbf{x}_n$  is a row vector with logarithms of inputs,  $\boldsymbol{\beta}$  is a column vector of parameters,  $v$  is a random variable representing noise (measurement errors and/or random shocks) and  $u \geq 0$  is a one-sided random variable representing technical inefficiency (it is assumed that  $v$  and  $u$  are independently distributed of each other).

The random variable  $v$  is usually assumed to have a normal distribution,  $N(0, \sigma_v^2)$ , and  $u$  is defined through different distributions, such as exponential, non-negative half normal or gamma distributions.

**Definition 2.22.** *From model (2.89) and Proposition 2.4, an output-oriented measure of technical efficiency is given by*

$$TE_n = \frac{y_n}{\exp(f(\mathbf{x}_n, \boldsymbol{\beta}) + v_n)} = \frac{\exp(f(\mathbf{x}_n, \boldsymbol{\beta}) + v_n - u_n)}{\exp(f(\mathbf{x}_n, \boldsymbol{\beta}) + v_n)} = \exp(-u_n), \quad (2.90)$$

that represents the ratio of the observed output to the potential output for the  $n$ th producer.

The potential output, mentioned in Definition 2.22, is defined by the stochastic production frontier:  $\exp(f(\mathbf{x}_n, \boldsymbol{\beta}) + v_n)$ . Naturally,  $TE_n$  assumes a value between zero and one.

Model (2.89) is usually estimated with the maximum likelihood (ML) estimator. Kumbhakar and Lovell [96, pp. 74–90] present all the estimation procedures with the ML estimator for different distributional assumptions required for the two-error components. Indeed, this is the main criticism in SFA in particular the choice of the distribution for the  $u$  error component, since different distributional assumptions can lead to different estimates of technical efficiency. The main advantage of SFA is the structure of the composed error, which separates the impacts on production outside the producer’s control (strikes, bad weather, luck) from technical efficiency.

Given the popularity of SFA there is a wide range of software available to estimate technical efficiency with SFA. The literature on SFA is huge; see, for example, Kumbhakar and Lovell [96] for an interesting overview.

The state-contingent approach to modeling production under uncertainty goes back to the work of Arrow and Debreu [7], but the theory has been considerably developed recently by Chambers and Quiggin [23]. In the state-contingent framework, uncertainty in production is described through a set of possible states of nature, where producers can allocate different inputs (in different states of nature) to manage the uncertainty caused by “nature”. This approach assumes that the producer choices can lead to different results in different states of nature. This is an important advantage of state-contingent production frontier models over the usual stochastic production frontier models, as the one in Definition 2.21. Furthermore, the few empirical work developed within the state-contingent framework indicates that the usual SFA can provide biased estimates of technical efficiency; e.g., Chambers and Quiggin [23], Nauges et al. [121], Quiggin and Chambers [134] and Rasmussen [136].

The discussion on technical efficiency provided in this section is limited to the minimum requirements for the understanding of Chapter 3. A complete overview of this topic is beyond the scope of this thesis.



## Chapter 3

# Technical efficiency with state-contingent production frontiers

“[...] in economics a truly well posed problem is virtually unknown.”

Comment from Edwin Jaynes mentioned in the authors' preface of Golan et al. [69].

In this chapter, the performance of the GME, GCE and GME- $\alpha$  estimators to assess output technical efficiency with state-contingent production frontiers under difficult empirical conditions is illustrated. Moreover, a fairly general extension of the model proposed by O'Donnell et al. [125] is introduced. Furthermore, the procedures to accommodate those ME estimators are also presented, in particular a new proposal to define the supports for the inefficiency error component. Finally, a simulation study shows that those ME estimators are powerful alternatives to the ML estimator.

### 3.1 Introduction

In the last decade, the work of Chambers and Quiggin [23] inspired a remarkable research in the production economics literature. The theory of state-contingent production is not

new, but it has been considerably developed in the recent years; e.g., Chambers and Quiggin [24, 25], Chavas [29], O'Donnell and Griffiths [124], O'Donnell et al. [125] and Rasmussen [136]. The reason that may explain this increasing interest is pointed out by Quiggin and Chambers [134, p. 167]:

“Almost every problem in economics involves uncertainty and, in almost every case, uncertainty is best interpreted in a state-contingent framework.”

The state-contingent approach allows a more realistic representation of economic production problems under uncertainty, since producers can allocate different inputs to different states of nature and manage in this way uncertainty. This approach provides higher estimates of technical efficiency when compared with the traditional stochastic frontier analysis (SFA), which may result in different economic policy implications.

The state-contingent framework allows the decomposition of the deviations from the production frontier into statistical noise (e.g., errors from the assumed functional forms or errors by not considering all the possible states of nature), inefficiency of the producer and risk (identification of output shortfalls due to adverse states of nature); e.g., O'Donnell et al. [125]. Thus, differences in outputs may not be due to the inefficiency of the producer, but merely result from an adverse state of nature. This recognition is fundamental!

Although the theory of state-contingent production is well established, the empirical implementation of this approach is still in an infancy stage. The information *per* state of nature is usually not available, since there is no tradition in collecting the production information associated with the states of nature. However, it is very unlikely that this difficulty will remain in the near future. The players involved in the production process benefit from the adoption of the state-contingent approach. On the one hand, policy agents know that only with good information on the production activity, it is possible to define the best policies that can contribute to its improvement. On the other hand, by collecting<sup>1</sup> detailed information, producers know that the productivity and efficiency evaluation can be more realistic and fair.

The empirical work with state-contingent production frontiers faces two main difficulties: the number of states of nature may be large, leading to the possibility of under-determined

---

<sup>1</sup>For example, in agricultural production the uncertainty is mainly due to the weather conditions. Even with a simple weather station, it is possible to obtain data for each farmer and for any desired period.

models; and with the increasing number of states of nature, it is very likely to have few observations in some states of nature, as well as collinearity problems. Generally speaking, the models in empirical work with state-contingent production frontiers are usually ill-posed. Because of these and other difficulties, there is an urgent need to develop robust estimation techniques in this context; e.g., Chavas [29] and O'Donnell et al. [125].

To the best of the authors' knowledge, the GCE and the GME- $\alpha$  estimators have not yet been used in the estimation of state-contingent production frontiers. However, there have been few attempts employing the GME estimator in this context. For example, using a simulation study, Rasmussen and Karantininis [137] compare the GME with the generalized least squares (GLS) estimator in the estimation of state-contingent production functions, although the technical efficiency is not considered. Rasmussen and Karantininis [137] conclude that the GME estimator may be useful but more studies are needed to evaluate the performance of this estimator in the state-contingent framework.

The main purpose of this chapter is to evaluate the performance of the GME, GME- $\alpha$  and GCE estimators to assess output technical efficiency with state-contingent production frontiers under difficult empirical conditions. This work is the first contribution to the estimation of technical efficiency with state-contingent production frontiers that combines the GME, GME- $\alpha$  and GCE estimators. Section 3.2 presents a fairly general extension of the model proposed by O'Donnell et al. [125] as well as all the procedures to use the GME, GME- $\alpha$  and GCE estimators. The simulation study is discussed in section 3.3.

## 3.2 State-contingent production with maximum entropy estimators

### 3.2.1 A state-contingent production frontier model

O'Donnell et al. [125] specify a production technology defined by the following Cobb-Douglas function

$$\ln q_s = b^{-1}(\ln x_s - \ln a_s), \quad (3.1)$$

where  $q_s$  is the observed output in state  $s$  ( $s = 1, 2, \dots, S$ ),  $b \geq 1$  is a parameter that accounts for the possibility of output substitution between states,  $x_s$  is the amount of input allocated

to state  $s$  and  $a_s > 0$  are *ex post* realizations of an unobservable random variable under nature's control. Alternatively,  $a_s$  can also be thought of as technical parameters that are specific to the production of output in state  $s$ . A detailed explanation of  $b$  and  $a_s$  can be found in O'Donnell et al. [125].

Following Nauges et al. [121], the production technology in (3.1) can be reformulated to accommodate exogenous variables and/or non-state-allocable inputs as follows

$$\ln q_s = b^{-1}(\ln x_s - (\ln a_s + \ln f(\mathbf{z}))), \quad (3.2)$$

where  $\mathbf{z}$  is a  $(1 \times K)$  vector of exogenous variables and/or non-state-allocable inputs and  $a_s$  is replaced by  $a_s f(\mathbf{z})$ . Assuming, for simplicity, that the component  $\ln f(\mathbf{z})$  is linear, the state-contingent production frontier model in a stochastic framework is given by

$$\ln q_n = b^{-1} \sum_{s=1}^S d_s (\ln x_s - \ln a_s) - b^{-1} \sum_{k=1}^K \alpha_k z_k + v_n - u_n, \quad (3.3)$$

where  $q_n$  is the output of producer  $n$ ,  $n = 1, 2, \dots, N$ ,  $d_s$  is a dummy variable that assumes the value 1 when the state  $s$  is chosen and 0 otherwise,  $K$  represents the number of exogenous variables (including non-state-allocable inputs),  $\alpha_k$  are parameters to be estimated,  $z_k$  are exogenous variables,  $v$  is a random variable representing noise and  $u \geq 0$  is a one-sided random variable representing technical inefficiency. The random variable  $v$  is usually assumed to be symmetric and distributed independently of  $u$ ; see Definition 2.21 and Kumbhakar and Lovell [96] for a detailed discussion of distributional assumptions in stochastic frontier models.

The state-contingent production frontier model in (3.3) only allows a single state-allocable input,  $x$ . In order to include  $P$  state-allocable inputs in this state-contingent production frontier model, the production technology specified in (3.1) can be generalized as follows

$$\ln q_s = \sum_{p=1}^P b_{ps}^{-1} (\ln x_{ps} - \ln a_s). \quad (3.4)$$

In this case, the state-contingent production frontier model in (3.3) is rewritten as

$$\ln q_n = \sum_{s=1}^S \sum_{p=1}^P d_s b_{ps}^{-1} \ln x_{ps} - \sum_{s=1}^S \sum_{p=1}^P d_s b_{ps}^{-1} \ln a_s - \sum_{s=1}^S \sum_{p=1}^P b_{ps}^{-1} \sum_{k=1}^K \alpha_k z_k + v_n - u_n, \quad (3.5)$$

with  $0 < b_{ps}^{-1} \leq 1$ ,  $a_s > 0$  and  $0 < \sum_p b_{ps}^{-1} \leq P$ , for all  $s$  ( $s = 1, 2, \dots, S$ ).

The state-contingent production frontier model in (3.5) is used in the simulation study. It is important to note that in many real-world problems the production technology specified

in (3.4) may be restrictive, since it implies that the output in any particular state of nature  $s$  will be zero if any one of the  $P$  inputs is zero. In order to overcome this potential limitation in some real-world problems, the production technology defined in (3.4) can be extended by

$$q_s = \prod_{p=1}^P \left( \frac{x_{ps} - \delta_p x_{ps} + \delta_p x_p}{a_s} \right)^{b_{ps}^{-1}}, \quad (3.6)$$

with  $x_p = \sum_{s=1}^S x_{ps}$  and  $0 \leq \delta_p \leq 1$ , for all  $p$  ( $p = 1, 2, \dots, P$ ). When all  $\delta_p$  are equal to zero, the production technology (3.4) is obtained.<sup>2</sup> Some flexible functional forms (e.g., generalized quadratic or translog) for the production technology will be used in future work.

### 3.2.2 Maximum entropy estimators

As already discussed in section 2.2, the essence of the GME and the GME- $\alpha$  estimators is identical: the error component in a regression model is considered as another set of unknown parameters to be estimated simultaneously along with the unknown parameters of the model. Both sets of unknown parameters are defined in the form of probabilities that are estimated within bounded support spaces using the maximum entropy principle. The main difference between these estimators lies in the entropy measure used: while the GME estimator employs the Shannon entropy, the GME- $\alpha$  uses the Tsallis and Rényi entropies. This difference leads to different properties for these estimators; see subsection 2.2.3.

As noted by Golan and Perloff [68], the GME and GME- $\alpha$  coefficient estimates converge asymptotically as the sample size grows unbounded.<sup>3</sup> However, the GME and GME- $\alpha$  estimators can produce quite different results in small samples sizes schemes. In the simulation study performed in section 3.3, models with few observations in some states of nature (small samples sizes) are investigated, as well as models affected by collinearity (usually when the number of states of nature is very large and there are few observations in some states).

For simplicity, consider the state-contingent production frontier model in (3.5) defined by the general matricial form

$$\ln \mathbf{q} = f(\mathbf{X}; \boldsymbol{\beta}) + \mathbf{v} - \mathbf{u}. \quad (3.7)$$

---

<sup>2</sup>Naturally, the ME estimators can also be applied in this case. This model is not used in the simulation study due to some numerical instability. Additional research on this issue is required in future work.

<sup>3</sup>As the number of observations increases, the GME, GME- $\alpha$ , OLS and ML estimates converge asymptotically to the same solution.

In order to use the GME and GME- $\alpha$  estimators, the reparameterization of the  $(R \times 1)$  vector  $\boldsymbol{\beta}$  and the  $(N \times 1)$  vector  $\boldsymbol{v}$  follows the same procedure as in the traditional regression model. Each parameter is treated as a discrete random variable with a compact support and  $2 \leq M < \infty$  possible outcomes and each error is defined as a finite and discrete random variable with  $2 \leq J < \infty$  possible outcomes. Assuming that both the unknown parameters and the unknown error terms can be bounded *a priori*, the reparameterization is given by

$$\boldsymbol{\beta} = \mathbf{Z}\mathbf{p} = \begin{bmatrix} z'_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & z'_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & z'_R \end{bmatrix} \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ p_R \end{bmatrix}, \quad (3.8)$$

with  $\mathbf{Z}$  a  $(R \times RM)$  matrix of support points and  $\mathbf{p}$  a  $(RM \times 1)$  vector of unknown probabilities, and

$$\boldsymbol{v} = \mathbf{A}\mathbf{w} = \begin{bmatrix} a'_1 & \mathbf{0} & \dots & \mathbf{0} \\ \mathbf{0} & a'_2 & \dots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \dots & a'_N \end{bmatrix} \begin{bmatrix} w_1 \\ w_2 \\ \vdots \\ w_N \end{bmatrix}, \quad (3.9)$$

with  $\mathbf{A}$  a  $(N \times NJ)$  matrix of support points and  $\mathbf{w}$  a  $(NJ \times 1)$  vector of unknown probabilities.

This traditional approach can be extended to the vector  $\mathbf{u}$ . The reparameterization<sup>4</sup> is similar to the one conducted for the random variable representing statistical noise,  $\boldsymbol{v}$ , taking only into account that  $\mathbf{u}$  is a one-sided random variable which implies that the lower bound for the supports (with  $2 \leq L < \infty$  points) is zero for all error values (the full efficiency case, since  $\exp(0) = 1$ ).<sup>5</sup> Note also that the unknown error terms can be bounded *a priori*, since technical efficiency varies between 0 and 100%. However, since  $\lim_{x \rightarrow -\infty} \exp(x) = 0$ , a large value is needed to define an upper bound in the supports. For example,  $\exp(-5)$  represents a technical efficiency of 0.67%, near the lowest level of technical efficiency. The reparameterization of  $\mathbf{u}$  can be defined by  $\mathbf{u} = \mathbf{B}\boldsymbol{\rho}$ , with  $\mathbf{B}$  a  $(N \times NL)$  matrix of support points and  $\boldsymbol{\rho}$  a  $(NL \times 1)$  vector of unknown probabilities.

The maximum entropy principle is applied to estimate the unknown  $\mathbf{p}$ ,  $\mathbf{w}$  and  $\boldsymbol{\rho}$  vectors

<sup>4</sup>Campbell et al. [20] consider this reparameterization with SFA.

<sup>5</sup>Note that, given the model (3.7), the ratio of observed output to potential output is given by  $\exp(-\mathbf{u})$ , representing a measure of output technical efficiency; see Definition 2.22.

that maximize

$$H(\mathbf{p}, \mathbf{w}, \boldsymbol{\rho}) = \sum_{r=1}^R H_{\alpha_1}^e(\beta_r) + \sum_{n=1}^N H_{\alpha_2}^e(v_n) + \sum_{n=1}^N H_{\alpha_2}^e(u_n), \quad (3.10)$$

or, equivalently,

$$H(\mathbf{p}, \mathbf{w}, \boldsymbol{\rho}) = H_{\alpha_1}^e(\boldsymbol{\beta}) + H_{\alpha_2}^e(\mathbf{v}) + H_{\alpha_2}^e(\mathbf{u}), \quad (3.11)$$

subject to the model constraint and the additivity constraints, respectively,

$$\begin{aligned} \ln \mathbf{q} &= \mathbf{XZp} + \mathbf{Aw} - \mathbf{B}\boldsymbol{\rho}, \\ \mathbf{1}_R &= (\mathbf{I}_R \otimes \mathbf{1}'_M)\mathbf{p}, \\ \mathbf{1}_N &= (\mathbf{I}_N \otimes \mathbf{1}'_J)\mathbf{w}, \\ \mathbf{1}_N &= (\mathbf{I}_N \otimes \mathbf{1}'_L)\boldsymbol{\rho}, \end{aligned} \quad (3.12)$$

where  $\otimes$  represents the Kronecker product,  $H_{\alpha_1}^e(\cdot)$  and  $H_{\alpha_2}^e(\cdot)$  are entropy measures (Shannon, Rényi or Tsallis), and  $\alpha_1, \alpha_2$  are the order of the entropy measure applied when  $e =$  Rényi or Tsallis entropies. Note that the case  $\alpha_1 = \alpha_2 = 1$  is equivalent to  $e =$  Shannon entropy. The maximum entropy estimators generate the optimal probability vectors  $\hat{\mathbf{p}}$ ,  $\hat{\mathbf{w}}$  and  $\hat{\boldsymbol{\rho}}$  that are used to obtain point estimates of the unknown parameters and the unknown errors, using the reparameterizations defined previously.

The support matrices  $\mathbf{Z}$ ,  $\mathbf{A}$  and  $\mathbf{B}$  are defined by the researcher based on prior information. When such information does not exist, as already mentioned in subsection 2.2.1, Golan et al. [69] suggest the definition of wide bounds without extreme risk consequences. However, it is important to note that some parameter spaces appearing in the state-contingent production frontier model in (3.5) are known, which means that some supports are easily defined for the matrix  $\mathbf{Z}$ . Since the vector  $\mathbf{v}$  is a two-sided random variable representing noise, the supports in the matrix  $\mathbf{A}$  can be defined symmetrically and centered around zero. Given the lack of information concerning this error vector, the supports can be defined using the three-sigma rule with the empirical standard deviation of the noisy observations (the usual strategy when no prior information exists). The number of support points is usually five and three, respectively, for each support in  $\mathbf{Z}$  and  $\mathbf{A}$ .

As previously mentioned, in order to specify the matrix  $\mathbf{B}$ , the support spaces for each error can be defined by the interval  $[0, b]$ , where  $b$  is large enough to provide technical efficiency estimates near zero. Given the support points, it is possible to express prior beliefs

or information about the error component (e.g., skewness). It is important to note that the traditional distributional assumptions concerning the error inefficiency component (half normal, truncated normal, exponential and gamma distributions, among others) have been used in several empirical studies, since it is expected a particular behavior in the distribution of technical inefficiency estimates. In the discussion of the normal – half normal model, Kumbhakar and Lovell [96, p. 74] argue that the choice of the latter distribution (a non-negative half normal) for the inefficiency error component

“[...] is based on the plausible proposition that the modal value of technical inefficiency is zero, with increasing values of technical inefficiency becoming increasingly less likely.”

Similar beliefs still hold for the other traditional models. These distributional assumptions are not necessary with the maximum entropy estimators, but the same beliefs can be expressed in the model through the error supports.

In a state-contingent production frontier framework, a similar approach as the one developed and tested by Campbell et al. [20] with SFA can be used. This approach considers five points and assumes a negative skewness for the estimates of technical efficiency.<sup>6</sup> Campbell et al. [20] recognize that the choice of the supports is somewhat arbitrary and suggest the use of the mean of the DEA and SFA efficiency estimates to define the supports for the error inefficiency component.<sup>7</sup> To reduce the subjectivity in this support choice, and since the state-contingent approach provides higher estimates of technical efficiency (when compared with DEA or SFA methodologies), a simpler approach is suggested here in which the supports of matrix  $\mathbf{B}$  are defined as

$$\mathbf{b}'_n = [0, 0.01, 0.02, 0.03, -\ln(\text{DEA}_n)], \quad (3.13)$$

where  $\text{DEA}_n$  represents the DEA technical efficiency estimate for the production unit  $n$  ( $n = 1, 2, \dots, N$ ). Since all deviations from the DEA production frontier are due to inefficiency, the DEA method provides lower levels of efficiency than the SFA and the state-contingent approach; e.g., O'Donnell et al. [125]. Thus, DEA can be used to define an upper bound

<sup>6</sup>See Tonini and Pedre [170] for an empirical application of this approach.

<sup>7</sup>Campbell et al. [20] show that the GME estimator provides a potential alternative frontier estimation approach that combines the strengths of SFA and DEA, without making distributional assumptions on the inefficiency error component.

for the supports. Specific supports can be defined for each production unit (the desirable approach) or, for simplicity, equal supports can be defined for all observations, considering  $DEA_n$  as the lowest technical efficiency estimate in the  $N$  observations. Needless to say that other prior information can alternatively be used to define the lower bound for the efficiency values (for example, other results with different technical efficiency estimation methods or other results in the same production activity). In the worst case scenario, where no prior information exists, a value  $b$  can be selected such that  $\exp(-b)$  provides a reasonable small value for the expected technical efficiency estimates.

The GCE formulation is very appealing when prior information exists concerning the unknown parameters and the unknown errors, that can be expressed as a set of subjective probability distributions. The idea underlying the GCE is to minimize the entropy distance between the data (in the form of  $\mathbf{p}$ ,  $\mathbf{w}$  and  $\boldsymbol{\rho}$ ) and the prior information (in the form of  $\mathbf{q}_1$ ,  $\mathbf{q}_2$  and  $\mathbf{q}_3$ ). Considering only the Shannon entropy measure, the GCE formulation selects the vectors  $\mathbf{p}$ ,  $\mathbf{w}$  and  $\boldsymbol{\rho}$  that minimize

$$\begin{aligned} H(\mathbf{p}, \mathbf{q}_1, \mathbf{w}, \mathbf{q}_2, \boldsymbol{\rho}, \mathbf{q}_3) &= (\mathbf{p}' \ln \mathbf{p} - \mathbf{p}' \ln \mathbf{q}_1) + (\mathbf{w}' \ln \mathbf{w} - \mathbf{w}' \ln \mathbf{q}_2) + (\boldsymbol{\rho}' \ln \boldsymbol{\rho} - \boldsymbol{\rho}' \ln \mathbf{q}_3) \\ &= \mathbf{p}' \ln(\mathbf{p}/\mathbf{q}_1) + \mathbf{w}' \ln(\mathbf{w}/\mathbf{q}_2) + \boldsymbol{\rho}' \ln(\boldsymbol{\rho}/\mathbf{q}_3) \end{aligned} \quad (3.14)$$

subject to conditions (3.12). When these probability distributions expressed in the vectors  $\mathbf{q}_1$ ,  $\mathbf{q}_2$  and  $\mathbf{q}_3$  are uniform (non-informative), the GCE and GME solutions are equal. In the context of state-contingent production, only the vector  $\mathbf{q}_3$  is non-uniform following the prior beliefs mentioned previously. Thus, the model (3.14) is rewritten as the minimization of

$$H(\mathbf{p}, \mathbf{w}, \boldsymbol{\rho}, \mathbf{q}_3) = \mathbf{p}' \ln \mathbf{p} + \mathbf{w}' \ln \mathbf{w} + \boldsymbol{\rho}' \ln(\boldsymbol{\rho}/\mathbf{q}_3) \quad (3.15)$$

subject to conditions (3.12). For the GCE estimator, the supports in matrix  $\mathbf{B}$  can follow a similar structure as in (3.13), although with equally spaced points in the range  $(0, -\ln(DEA_n))$ . For example, suppose that  $DEA_n = 45\%$  for a given observation. Thus, the support for this observation can be defined by  $\mathbf{b}'_n = [0, 0.2, 0.4, 0.6, 0.8]$ . The set of subjective probability distribution may take the form

$$\mathbf{q}_3 = [0.40, 0.30, 0.15, 0.10, 0.05]' \quad (3.16)$$

for each observation, where the cross-entropy objective shrinks the posterior distribution in order to have more mass near zero. It is important to note that eventual wrong prior beliefs

expressed in  $\mathbf{q}_3$  do not have a significant impact on the estimates. As pointed out by Golan et al. [69, p. 142], this is an important feature of the GCE approach, since

“[...] incorrect prior information is effectively discounted by the GCE criterion if it does not agree with the sample.”

Since the model constraint must be satisfied for any interior solution, the estimates cannot be substantially different from the true values, even with wrong prior information expressed in  $\mathbf{q}_3$ . In other words, the GCE estimator is robust to incorrect prior information. The GCE formulation employing the Rényi and Tsallis entropies is left for future research.

### 3.3 Simulation study

In this section, a simulation study is presented to illustrate the performance of the GME, GCE and GME- $\alpha$  estimators to assess output technical efficiency with state-contingent production frontiers. Besides comparing these maximum entropy estimators, special interest is devoted to models in which the ML estimator should be avoided. Thus, particular attention is given to models with few observations in some states of nature and models with a large number of states of nature (which are models usually affected by collinearity), illustrating potential real problems when using the state-contingent production approach. Note that few observations *per* state restrict the use of traditional estimators and, even with a reasonable number of observations in each state, the collinearity problem arises with an increasing number of states of nature.

Before generating the artificial data sets based on the state-contingent production frontier model in (3.5), the number of states ( $S$ ), the number of state-allocable inputs ( $P$ ) and the number of exogenous variables ( $K$ ) are defined. The number of observations in each state is randomly generated between one and a maximum specified value. The 2-norm condition number (i.e., the ratio of the largest singular value of  $\mathbf{X}$  to the smallest singular value) is considered to evaluate the collinearity in the design matrix. As expected, models with a very high 2-norm condition number appeared frequently in this study. Since the 2-norm condition number increases with the number of (state- and non-state-allocable) inputs and considering that the number of inputs, in empirical studies, usually lies between four and six,  $P = 2$  and

$K = 3$  are, respectively the number of state-allocable inputs and the number of exogenous variables defined in this study, representing a very likely empirical scenario. Depending also on the number of observations *per* state, models with a 2-norm condition number greater than  $5 \times 10^3$  are very likely for  $P \geq 3$  and  $K \geq 4$ .  $S = 2, 3, 5$  and 10 are the possible number of states of nature considered in this study.

The matrix  $\mathbf{X}$  and the parameter vector  $\boldsymbol{\beta}$  defined in (3.7) are generated according to the state-contingent production frontier model in (3.5). The state-allocable inputs and the exogenous variables are generated randomly using the absolute values of normal distributions with means between one and ten, and standard deviations equal to one. The parameters are generated using uniform distributions, namely  $b_{ps}^{-1}$  is generated using  $U(10^{-3}, 1)$ ; for the constant parameter  $U(-4, 4)$  is used; and the parameters of the exogenous variables are generated using  $U(-2, 2)$ . The random variable representing noise,  $\mathbf{v}$ , is generated using a standard normal distribution. The one-sided random variable representing output technical inefficiency,  $\mathbf{u}$ , is generated considering two different distributions: (i) the half normal distribution with zero mean and standard deviation  $\sigma_u = 0.1$ ; (ii) the exponential distribution with parameter  $\lambda = 10$ . In case (i), the normal – half normal model, a very common model used in SFA, is generated. Note that the selected values for the parameters of the previous distributions are arbitrary. Nevertheless, the chosen values are considered reasonable in empirical studies.

The supports in matrix  $\mathbf{Z}$  are defined in the interval  $[0, 1]$  for the parameters of state-allocable inputs and  $[-10, 10]$  for the other parameters. This study defines five points ( $M = 5$ ) for all supports in  $\mathbf{Z}$ . The support interval for the other parameters could be narrowed, yet it is defined in this way to illustrate real cases where prior information is scarce. The supports in the matrix  $\mathbf{A}$  are defined symmetric and centered on zero, using the three-sigma rule with the empirical standard deviation of the noisy observations. The number of support points is three in each support ( $J = 3$ ).

The supports in the matrix  $\mathbf{B}$  are defined using (3.13). In the normal – half normal models, this study considers that the lowest DEA estimate<sup>8</sup> of output technical efficiency obtained in the samples is approximately 67%. Note that, using the half normal distribution defined in (i) to generate the vector  $\mathbf{u}$ , the probability of finding a value greater than 0.3 is less than 1%. Indeed, a value greater than 0.3 (the value is 0.31) is generated only once, in all the simulations

---

<sup>8</sup>The DEA method is not performed in this study, although it is recommended in real empirical analysis.

conducted. Thus, by considering the lowest technical efficiency estimate obtained by DEA as 67%, the supports in matrix  $\mathbf{B}$  are defined by  $\mathbf{b}'_n = [0, 0.01, 0.02, 0.03, 0.4]$ . In the normal – exponential models, the supports in the matrix  $\mathbf{B}$  are defined by  $\mathbf{b}'_n = [0, 0.01, 0.02, 0.03, 0.8]$ , considering that the lowest DEA estimate of technical efficiency is 45%. In all the simulations performed with the normal – exponential models, the maximum value generated for  $\mathbf{u}$  is 0.62. Note that conservative supports are used in both models (i.e., supports with an upper bound larger than necessary), illustrating that the DEA method provides lower levels of efficiency than the SFA and the state-contingent approach.

The GCE estimator is applied using the supports  $[0, 0.4]$  and  $[0, 0.8]$ , respectively in the normal – half normal models and the normal – exponential models, with five equally spaced points in both supports. Two cases are defined: the GCE1 with  $\mathbf{q}_3$  defined by (3.16) and the GCE2 with  $\mathbf{q}_3 = [0.50, 0.40, 0.05, 0.03, 0.02]'$ , where the cross-entropy objective shrinks the posterior distribution in order to have more mass near zero than with (3.16).

In this simulation study, several models are tested although only ten of them are reported here (Table 3.1 – Table 3.5). For each of these ten models the vectors  $\mathbf{u}$  and  $\mathbf{v}$  are replicated in order to create 1000 different versions of each model. To evaluate the performance of the estimators, two measures are applied: the mean squared error loss (MSEL), with  $\text{SEL}(\hat{\boldsymbol{\beta}}) = \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|^2$ , and the difference between the true and the estimated mean of technical efficiency (DMTE). The results for the different GME- $\alpha$  estimators are presented using the notation “GME $\text{name}\alpha_1\alpha_2$ ”, where  $\text{name}$  is “R” or “T” representing, respectively, the Rényi or Tsallis entropies; and  $\alpha_1, \alpha_2$  are the order of the entropy measure used for the parameters and the error components, respectively. The values  $\alpha_1, \alpha_2 = \{2, 4, 6\}$  are chosen taking into account the range of values  $[0, 6.5]$  for the order of the entropy measures used by Golan and Perloff [68] in their simulation studies.

Under those severe empirical conditions, a first conclusion is that the GCE, GME and GME- $\alpha$  estimators perform, in general, better than the ML estimator in terms of MSEL and DMTE. Even in models with only two or three states of nature, the ME estimators exhibit, in general, a better performance than the ML estimator, except in models with a small 2-norm condition number and with a reasonable number of observations in each state of nature (i.e., models with ideal empirical conditions). In those models (e.g., Model 1 in Table 3.1 and Model 8 in Table 3.4), all the estimators have similar values of MSEL.

**Table 3.1:** MSEL and DMTE for the different estimators (Model 1 and Model 2).

	Model 1		Model 2	
	[33;17] <sup>a</sup>		[4;46]	
	normal – half normal		normal – half normal	
	cond <sub>2</sub> $\mathbf{X}$ = 44		cond <sub>2</sub> $\mathbf{X}$ = 113	
	MSEL	DMTE	MSEL	DMTE
ML	2.9782	0.0714	10.4192	0.0729
GME	2.2731	0.0133	3.3616	0.0130
GCE1	2.1049	0.0272	2.9355	0.0293
GCE2	2.1214	0.0129	2.9265	0.0105
GMER22	1.0648	0.0132	1.1980	0.0130
GMER44	1.0717	0.0132	1.1568	0.0129
GMER66	1.2091	0.0131	1.2445	0.0128
GMER26	1.8072	0.0131	1.6425	0.0128
GMER62	0.7861	0.0133	0.9120	0.0130
GMET22	2.5209	0.0132	4.6930	0.0129
GMET44	3.1781	0.0109	8.2582	0.0106
GMET66	3.7674	0.0074	9.6361	0.0076
GMET26	1.1934	0.0075	1.2353	0.0075
GMET62	3.0773	0.0132	9.2867	0.0129

<sup>a</sup>The brackets define the number of observations *per* state in the corresponding model.

As the number of observations *per* state decreases and/or the number of states of nature increases, the 2-norm condition number increases substantially. In particular, the 2-norm condition number can increase exponentially when the number of observations *per* state approaches to one, illustrating the collinearity problem in the estimation of state-contingent production frontiers. For example, in all the ten models presented here, if a state of nature with only one observation is considered, the 2-norm condition number can be greater than  $10^{10}$ ! As expected, the MSEL of the ML estimator increases when the 2-norm condition number increases. For the GCE, GME and GME- $\alpha$  estimators, the increase in MSEL is smaller and, in general, the MSEL increases only slightly (except in models affected by severe collinearity, i.e., Model 6 in Table 3.3 and Model 10 in Table 3.5).

An interesting finding in the simulation study is that, in general, the ME estimators have

**Table 3.2:** MSEL and DMTE for the different estimators (Model 3 and Model 4).

	Model 3		Model 4	
	[6;94]		[3;5;242]	
	normal – half normal		normal – half normal	
	cond <sub>2</sub> $\mathbf{X}$ = 62		cond <sub>2</sub> $\mathbf{X}$ = 273	
	MSEL	DMTE	MSEL	DMTE
ML	4.6668	0.0723	13.4357	0.0738
GME	2.1684	0.0139	3.5306	0.0129
GCE1	1.5884	0.0293	2.5113	0.0296
GCE2	1.5872	0.0101	2.4053	0.0098
GMER22	1.2909	0.0138	2.2723	0.0128
GMER44	1.3665	0.0137	2.3426	0.0127
GMER66	1.5434	0.0135	2.4979	0.0124
GMER26	1.6639	0.0135	2.4382	0.0124
GMER62	1.3750	0.0138	2.3405	0.0128
GMET22	2.3208	0.0138	4.2724	0.0128
GMET44	2.7202	0.0109	5.7050	0.0099
GMET66	3.2137	0.0056	6.7481	0.0041
GMET26	1.6300	0.0058	2.3249	0.0044
GMET62	2.6685	0.0138	7.5808	0.0128

a lower DMTE than the ML estimator in most of the models presented. As expected, the ML estimator presents a larger DMTE in the normal – exponential models, since the technical efficiency estimates are computed for a normal – half normal model (illustrating the case of a wrong model specification); e.g., Kumbhakar and Lovell [96, pp. 74–78].

Among the ME estimators, the GCE2 and GME- $\alpha$  estimators provide, in general, the lowest DMTE. Note that the GME, GME- $\alpha$  and GCE (i.e., GCE1 and GCE2) estimators provide higher values of DMTE in the normal – exponential models than in the normal – half normal models. For the GME and the GME- $\alpha$  estimators, this result is due to the different supports in the matrix  $\mathbf{B}$  used in these models, with different prior means: 91.2% for the normal – half normal models and 84.2% for the normal – exponential models. Taking into account the parameters defined in the half normal and the exponential distributions in this

**Table 3.3:** MSEL and DMTE for the different estimators (Model 5 and Model 6).

	Model 5		Model 6	
	[11;3;38;7;41]		[10;40;4;12;3;39;22;13;31;26]	
	normal – half normal		normal – half normal	
	cond <sub>2</sub> $\mathbf{X}$ = 149		cond <sub>2</sub> $\mathbf{X}$ = 1198	
	MSEL	DMTE	MSEL	DMTE
ML	16.9371	0.0710	329.5597	0.0788
GME	8.7154	0.0135	17.3020	0.0124
GCE1	8.8675	0.0316	14.8480	0.0275
GCE2	8.8100	0.0086	13.9953	0.0119
GMER22	5.7226	0.0134	14.1451	0.0124
GMER44	5.9235	0.0133	13.8171	0.0123
GMER66	6.3263	0.0132	13.7514	0.0121
GMER26	6.5837	0.0132	16.6523	0.0122
GMER62	6.1228	0.0134	12.1694	0.0124
GMET22	7.1939	0.0134	18.9152	0.0123
GMET44	9.9523	0.0108	29.5623	0.0086
GMET66	12.5169	0.0158	33.9247	0.0035
GMET26	6.2566	0.0060	15.1987	0.0039
GMET62	12.6782	0.0134	47.1504	0.0123

simulation study, the positive skewness value of the exponential distribution is much higher than the positive skewness value of the half normal distribution. Thus, with a lower prior mean and a higher degree of positive skewness, the GME and the GME- $\alpha$  estimators generate higher values of DMTE in the normal – exponential models. A similar analysis can be made for the GCE estimators multiplying the support vector by the vector with prior information and taking the exponential with the symmetric of the result. The following prior means are found: 89.6% (GCE1) and 93.5% (GCE2) for the normal – half normal models, and 80.3% (GCE1) and 87.5% (GCE2) for the normal – exponential models.

Concerning only the GME- $\alpha$  class of estimators, the evaluation of the results obtained in this study is difficult. Different combinations  $(\alpha_1, \alpha_2)$  are introduced in an attempt to find the best choice for the order of the entropy measure, but the results are somewhat inconclusive.

**Table 3.4:** MSEL and DMTE for the different estimators (Model 7 and Model 8).

	Model 7		Model 8	
	[95;5]		[63;54;83]	
	normal – exponential		normal – exponential	
	cond <sub>2</sub> $\mathbf{X}$ = 57		cond <sub>2</sub> $\mathbf{X}$ = 38	
	MSEL	DMTE	MSEL	DMTE
ML	4.4704	0.0854	0.9382	0.0882
GME	2.5070	0.0674	1.2249	0.0667
GCE1	1.9763	0.0888	1.0428	0.0804
GCE2	1.9343	0.0368	0.8515	0.0283
GMER22	2.5706	0.0664	1.2603	0.0661
GMER44	2.9721	0.0657	1.2196	0.0652
GMER66	3.2642	0.0649	1.3323	0.0641
GMER26	2.6577	0.0650	1.3955	0.0642
GMER62	3.5678	0.0663	1.4279	0.0662
GMET22	2.7696	0.0665	1.1996	0.0658
GMET44	3.2794	0.0516	1.1093	0.0503
GMET66	3.7302	0.0425	1.5345	0.0388
GMET26	2.8196	0.0408	1.3572	0.0396
GMET62	3.4098	0.0666	1.1638	0.0658

However, it seems that the GME- $\alpha$  estimators with the Rényi entropy (regardless of the combination  $(\alpha_1, \alpha_2)$  considered) and the GME- $\alpha$  estimator with the Tsallis entropy (with  $\alpha_1 = 2$  and  $\alpha_2 = 6$ ) are the best choices in this study (specially in terms of MSEL). This is an important finding that deserves further investigation in future work. Different simulation studies should be conducted in order to find, for specific models, the best choice for the orders of the entropy measures, including different orders of entropy in the two-error component, rather than using the same value for both.

Among the ME estimators, the GCE estimator is probably the most adequate choice in the estimation of technical efficiency with state-contingent production frontiers. Although the GME and GME- $\alpha$  estimators are also valid options, both of them present some disadvantages. The GME implies a subjective selection of the support values defined in (3.13), particularly

**Table 3.5:** MSEL and DMTE for the different estimators (Model 9 and Model 10).

	Model 9		Model 10	
	[7;23;17;47;6]		[11;8;12;9;10;4;13;7;15;11]	
	normal – exponential		normal – exponential	
	cond <sub>2</sub> $\mathbf{X}$ = 180		cond <sub>2</sub> $\mathbf{X}$ = 689	
	MSEL	DMTE	MSEL	DMTE
ML	16.1426	0.0865	275.5349	0.0915
GME	5.5112	0.0696	13.8092	0.0668
GCE1	4.9967	0.0843	11.6348	0.0856
GCE2	4.7191	0.0335	11.4424	0.0323
GMER22	1.9657	0.0690	11.0388	0.0663
GMER44	2.0101	0.0683	10.8759	0.0657
GMER66	2.2714	0.0677	11.1634	0.0653
GMER26	3.3627	0.0677	14.0829	0.0654
GMER62	1.8406	0.0690	10.3162	0.0663
GMET22	6.8873	0.0688	13.9429	0.0661
GMET44	8.3576	0.0540	27.2592	0.0524
GMET66	11.0959	0.0415	37.7618	0.0386
GMET26	2.1965	0.0430	10.6572	0.0409
GMET62	12.0538	0.0688	44.1553	0.0661

the second, third and fourth values in the support, that determine the prior mean and the skewness. The problem of choosing the orders of the entropy measure ( $\alpha_1$  and  $\alpha_2$ ) underlies the GME- $\alpha$  estimators. In contrast, the GCE exhibits none of the previous drawbacks. Naturally, the GCE estimator involves the subjective prior weights defined in  $\mathbf{q}_3$ , although it is expected that the estimator remains stable when this prior information is incorrect; e.g., Golan et al. [69, pp. 138–144].

### 3.4 Conclusions

To the best of the authors' knowledge, this work is the first contribution to the estimation of technical efficiency with state-contingent production frontiers that combines the GME,

GME- $\alpha$  and GCE estimators. The discussion provided throughout this chapter reveals some advantages of the ME estimators. These estimators outperform the ML estimator in most of the cases analyzed: models with few observations in some states of nature and models with a large number of states of nature, that usually represent models affected by collinearity. Naturally, these results should be tempered with some caution, since more simulation studies and empirical applications are needed.

Nevertheless, some advantages of these estimators seem to be clear, namely: (1) the possibility of considering easily prior information on the parameters and errors components; (2) the traditional assumptions on the errors' distributions are not necessary; and (3) these estimators can be used in models with a large number of states of nature and even with only one observation *per* state (which represent models usually affected by severe collinearity). Note that few observations *per* state restrict the use of traditional estimators and, even with a reasonable number of observations in each state, the collinearity problem arises with an increasing number of states of nature. In these cases, researchers can reduce the number of states (increasing the number of observations in each state, but losing important information) or use the ME estimators that are robust in the presence of collinearity. Furthermore, the ME estimators can be used even if there is only one observation *per* state.

The GME- $\alpha$  estimators have never been used in the state-contingent production context and the results achieved here are, unfortunately, somewhat inconclusive. Concerning only this class of estimators, the GME- $\alpha$  estimators with the Rényi entropy (regardless of the combination  $(\alpha_1, \alpha_2)$  used) and the GME- $\alpha$  estimators with the Tsallis entropy (with  $\alpha_1 = 2$  and  $\alpha_2 = 6$ ) are probably the best choices in this study, in particular concerning MSEL values. The results achieved here support the results of Golan and Perloff [68]: some GME- $\alpha$  estimators have lower MSEL than the GME estimator. In the state-contingent production context, directions for further research may also include different orders of entropy in the two-error component rather than using the same value for both.

The results indicate that the ME estimators are powerful alternatives to the ML estimator under severe empirical conditions. Although all the estimators have drawbacks, it seems that the GCE estimator is probably more appropriate. The GME implies the subjective selection of the support values; the problem of choosing the orders of the entropy measure ( $\alpha_1$  and  $\alpha_2$ ) underlies the GME- $\alpha$  estimators. The GCE has the advantage that wrong prior weights

defined in  $\mathbf{q}_3$  are discounted if the information does not agree with the sample information.

The ME estimators are expected to contribute strongly to the increase of empirical work with state-contingent production frontiers in the near future. Besides the reasons already mentioned, the ME estimation of technical efficiency with state-contingent production frontiers, using the DEA and SFA methodologies, provides an important contribution to the efficiency analysis. Specifically, the DEA method is used to define an upper bound for the supports (the main criticism on DEA is used here as an advantage) and the composed error structure as in SFA is used here without distributional assumptions (the main criticism on SFA is avoided here). In contrast to SFA, the state-contingent approach allows producers to allocate different inputs in different states of nature and manage in this way uncertainty.



## Chapter 4

# The choice of the ridge parameter in ridge regression

“The ultimate choice of  $k$  [ridge parameter] for a specific application involving collinear explanatory variables still remains part art and part science.”

McDonald [112, p. 99].

In this chapter, a new method to estimate the ridge parameter, based on the ridge trace and the GME estimator, is presented. The performance of the new estimator is illustrated through a Monte Carlo simulation study and an empirical application to the well-known Portland cement data set.

### 4.1 Introduction

The multiple linear regression is one of the widely used models in mathematical statistics and the most common method used for estimating the regression coefficients is the ordinary least squares (OLS) method. For convenience, the usual assumptions of the general linear regression model stated in Definition 2.6 are presented next; e.g., Greene [71, p. 11–19].

**Assumption 4.1.** *The model (2.42) specifies a linear relationship between the response variable and the explanatory variables (linearity concerning the parameters).*

**Assumption 4.2.** *There are no exact linear relations between the explanatory variables in the model (2.42).*

**Assumption 4.3.** *Each error has the same finite variance,  $\sigma^2$ , and is uncorrelated with every other error in the model (2.42). This means that  $E[\mathbf{u}\mathbf{u}'|\mathbf{X}] = \sigma^2\mathbf{I}$ , where  $\mathbf{I}$  is a  $(N \times N)$  identity matrix.*

**Assumption 4.4.** *The errors are assumed to have a conditional expected value equal to zero at each observation. The expected value of each error is not a function of any explanatory variable, i.e., the explanatory variables do not have information concerning the expected value of the errors. This can be stated as  $E[\mathbf{u}|\mathbf{X}] = \mathbf{0}$ .*

**Assumption 4.5.** *The data contained in the  $\mathbf{X}$  matrix can be a mixture of constants and random variables, assuming that the mean and the variance of each error are independent of all elements of the  $\mathbf{X}$  matrix.*

**Assumption 4.6.** *The errors are normally distributed with zero mean and constant variance.*

The main interest in this chapter is the Assumption 4.2. The OLS method provides accurate results as long as the regressors are linearly independent and the errors are zero-mean normally distributed with constant variance, and uncorrelated. Under these conditions, the OLS estimator has smaller variance than any other linear unbiased estimator. This result is usually known as the Gauss-Markov theorem.<sup>1</sup>

It is well-known that the absence of collinearity is essential to the multiple regression model. In the presence of collinearity, the OLS estimator (as well as the ML estimator; see section 3.3) performs poorly since the variances of the parameter estimates can be substantially large, meaning that the estimates tend to be less precise. Therefore, the higher the collinearity, the less interpretable are the parameters; see section 2.3.

Various methods, including ridge regression, principal component regression, partial least squares regression, continuum regression, lasso, elastic net and least angle regression, are well suited to cope with collinearity problems; see, among others, Brown [19], Efron et al. [44], Hoerl and Kennard [76], Stone and Brooks [162], Tibshirani [168] and Zou and Hastie [185]

---

<sup>1</sup>The normality assumption is not needed in this theorem. Normality is mentioned here simply because it is useful for inference purposes. A detailed discussion of all the assumptions of the classical linear regression model can be found in Greene [71, p. 11–19].

for reviews. However, despite more recent approaches, ridge regression continues to play a key role in regression models affected by collinearity, and outperforms other competitors in many cases; e.g., Hastie et al. [75], McDonald [112] and Zou and Hastie [185].

The ridge regression introduced by Hoerl and Kennard [76] is a very popular parameter estimation method among practitioners to handle collinearity without removing variables from the regression model. By adding a small non-negative constant (the ridge parameter) to the diagonal of the correlation matrix of the explanatory variables, it is possible to reduce the variance of the OLS estimator through the introduction of some bias. Although the resulting estimators are biased, the biases are small enough for these estimators to be substantially more precise than the unbiased estimators. Therefore, these biased estimators are preferred to unbiased ones since they have a larger probability of being closer to the true parameter values. Concerning this issue of biased and unbiased estimators, Ryan [144, p. 466] argues that

“[...] for all practical purposes, the ordinary least squares (OLS) estimator will also generally be biased, because we can be certain that it is unbiased only when the model that is being used is the correct model. Since we cannot expect this to be true, we similarly cannot expect the OLS estimator to be unbiased. Therefore, although the choice between OLS and a ridge estimator is often portrayed as a choice between a biased estimator and an unbiased estimator, that really isn't the case.”

However, at this point, it is necessary to assume that the OLS is an unbiased estimator to better understand the importance of the ridge estimator. The mean squared error (MSE) of the ridge estimator is smaller than that of the OLS estimator provided the reduction in the variance is greater than the increase of the squared bias. In order to make this possible, it is necessary to select an adequate estimate for the ridge parameter. Arguably, the most straightforward approach is based on simply plotting the coefficients against several possible values for the ridge parameter and inspecting the resulting traces; e.g., Hoerl and Kennard [76, 77] and Zhang and McDonald [183]. As the ridge parameter increases, the variances of the coefficients estimates reduce and the coefficients become more stable. Therefore, the value of the ridge parameter is chosen at the point for which the coefficients no longer change

rapidly (this choice is quite subjective). It is important to stress, however, that stability does not imply convergence of the regression coefficients. In addition to graphical procedures, several formal methods have been proposed to estimate the ridge parameter. The literature on methods for choosing the ridge parameter includes, among many others, Alkhamisi et al. [4], Alkhamisi and Shukur [5], Gibbons [61], Golub et al. [70], Hoerl and Kennard [76, 77], Hoerl et al. [78], Khalaf and Shukur [90], Kibria [92], McDonald and Galarneau [113], Muniz and Kibria [120] and Tran [171].

The main purpose of this chapter is to introduce a new estimator for the ridge parameter (referred, hereafter, as Ridge-GME estimator), which combines the ridge trace and the GME estimator. An important disadvantage of most of the traditional analytical methods used to estimate the ridge parameter is the dependence on unknown parameters that have to be estimated from the data. In contrast, this dependence is avoided with the Ridge-GME estimator since only a GME estimate is necessary to obtain the ridge parameter estimate. A wide set of possible values for the ridge parameter (referred, hereafter, as the ridge interval), rather than a single estimate, is selected in the analysis of the ridge trace along with the Ridge-GME estimator. The rationale behind this new estimator can be summarized as follows: the ridge trace provides the prior information needed to obtain a solution for the GME estimator, as well as a wide set of possible values for the ridge parameter (i.e., the ridge interval). The estimate of the ridge parameter is obtained as being the value within the ridge interval that minimizes the infinity norm of the difference between the solution provided by the GME estimator and a solution of the ridge estimator (selected from the set of all possible solutions obtained from all possible values<sup>2</sup> for the ridge parameter defined by the ridge interval).

## 4.2 The ridge regression estimator

Consider the general linear regression model defined by (2.42) in Definition 2.6. Both the OLS and the ridge regression estimators of  $\beta$  are well-known.

**Definition 4.1.** *The OLS estimator of  $\beta$  in the model (2.42) is given by*

$$\hat{\beta}_{OLS} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}. \quad (4.1)$$

---

<sup>2</sup>A large number of possible values defined by the researcher; see an example in section 4.5.

**Definition 4.2.** *The ridge regression estimator of  $\boldsymbol{\beta}$  in the model (2.42) takes the form*

$$\widehat{\boldsymbol{\beta}}_{\text{ridge}} = (\mathbf{X}'\mathbf{X} + \eta\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}, \quad (4.2)$$

where  $\eta \geq 0$  denotes the ridge parameter and  $\mathbf{I}$  is a  $(K \times K)$  identity matrix.

Definition 4.2 represents the biased estimator proposed by Hoerl and Kennard [76].<sup>3</sup> In the ridge regression estimator, note that when  $\eta \rightarrow 0$ , the ridge estimator approaches the OLS estimator whereas the ridge estimator approaches the zero vector when  $\eta \rightarrow \infty$ . Thus, a trade-off between variance and bias is needed. Hoerl and Kennard [76, pp. 62–63] proved that the ridge estimator is superior to the OLS estimator (in a MSE sense) for a range of values of  $\eta$ , say  $0 < \eta < \sigma^2/\alpha_{\max}^2$ , where  $\alpha_{\max}^2$  is the largest squared value from a vector  $\boldsymbol{\alpha}$  that depends on  $\boldsymbol{\beta}$ ; see model (4.9) below. Is the problem of choosing  $\eta$  solved? Unfortunately the answer is no, because  $\sigma^2$  and  $\boldsymbol{\beta}$  are unknown!

**Definition 4.3.** *The MSE of the ridge estimator of  $\boldsymbol{\beta}$  in the model (2.42) is given by*

$$\text{MSE}(\widehat{\boldsymbol{\beta}}_{\text{ridge}}) = \underbrace{\sigma^2 \sum_{k=1}^K \frac{\lambda_k}{(\lambda_k + \eta)^2}}_{\text{Var}(\widehat{\boldsymbol{\beta}}_{\text{ridge}})} + \underbrace{\eta^2 \boldsymbol{\beta}'(\mathbf{X}'\mathbf{X} + \eta\mathbf{I})^{-2}\boldsymbol{\beta}}_{\left(\text{bias in } \widehat{\boldsymbol{\beta}}_{\text{ridge}}\right)^2}, \quad (4.3)$$

where the  $\lambda_k$ 's are the eigenvalues of the  $\mathbf{X}'\mathbf{X}$  matrix in correlation form.

A detailed derivation of this result can be found in Ryan [144, pp. 481–482]. Note that the variance of  $\widehat{\boldsymbol{\beta}}_{\text{ridge}}$  decreases as  $\eta$  increases, whereas the bias increases with  $\eta$ . Due to its biasedness, the ridge estimator is superior to the OLS estimator (in a MSE sense) if the reduction in the variance is greater than the increase of the squared bias. Since the range of values for which the ridge estimator is superior to the OLS estimator depends on the unknown parameters  $\boldsymbol{\beta}$  and  $\sigma^2$ , the challenge is to select an estimate of  $\eta$  such that the ridge estimator has a smaller MSE than the OLS estimator.

To better understand the drawbacks of the OLS estimator under collinearity, consider the expected squared distance between  $\widehat{\boldsymbol{\beta}}_{\text{OLS}}$  and  $\boldsymbol{\beta}$ , defined as

$$E \left[ (\widehat{\boldsymbol{\beta}}_{\text{OLS}} - \boldsymbol{\beta})'(\widehat{\boldsymbol{\beta}}_{\text{OLS}} - \boldsymbol{\beta}) \right]. \quad (4.4)$$

<sup>3</sup>There were some earlier introductions on this topic before this article; see McDonald [112, p. 93].

Considering  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_K > 0$  as the ordered eigenvalues of the  $\mathbf{X}'\mathbf{X}$  matrix in correlation form, it follows that

$$E \left[ (\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta})' (\hat{\boldsymbol{\beta}}_{OLS} - \boldsymbol{\beta}) \right] = \sigma^2 \sum_{k=1}^K \frac{1}{\lambda_k} \quad (4.5)$$

and

$$E \left[ \hat{\boldsymbol{\beta}}'_{OLS} \hat{\boldsymbol{\beta}}_{OLS} \right] = \boldsymbol{\beta}'\boldsymbol{\beta} + \sigma^2 \sum_{k=1}^K \frac{1}{\lambda_k}. \quad (4.6)$$

Thus, as  $\lambda_K$  becomes smaller (leading to an increase in collinearity), the vector  $\hat{\boldsymbol{\beta}}_{OLS}$  can be expected to be farther from the vector  $\boldsymbol{\beta}$ ; e.g., Hoerl and Kennard [77].

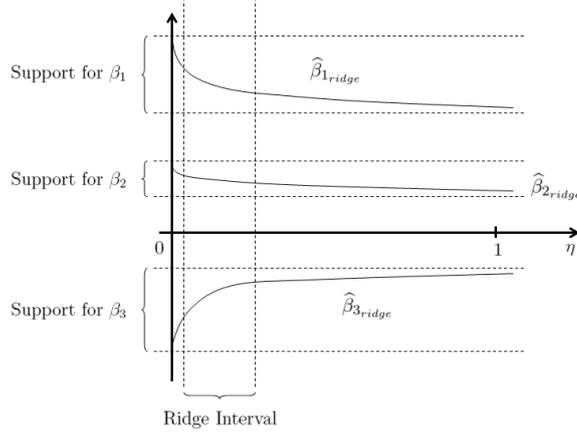
### 4.3 The Ridge-GME estimator

In this section, the Ridge-GME estimator is introduced and explained in some detail. As referred above, the idea underlying this new estimator is to combine efficiently the ridge trace and the GME estimator. The key issues are:

- (a) how can the information provided by the ridge trace be used without making a subjective selection of  $\hat{\eta}$ ?
- (b) how can the ridge trace and the GME estimator be efficiently combined?

In answering (a), note that it is commonly accepted in the literature that an estimate of the ridge parameter must be selected within the interval  $]0, 1]$ . Thus, the ridge interval selected by visual inspection of the ridge trace has to be a subset of the interval  $]0, 1]$ . Before answering (b), it is important to remember (see subsection 2.2.1, p. 34) that the practical implementation of the GME estimator is sometimes discouraged due to the fact that it needs exogenous information (which is not always available) for the parameters and error vectors (i.e., the supports in which each parameter or error is restricted to lie), and the GME coefficient estimates are sensitive to the specification of their support intervals; e.g., Caputo and Paris [22]. However, these (possible) disadvantages of the GME estimator are partially overcome by the information provided by the ridge trace. Hence, concerning (b), the ridge trace provides guidelines for the selection of the supports for the model regression parameters, which are in turn used to obtain the solution from the GME estimator. For an arbitrary ridge trace, Figure 4.1 illustrates this procedure for the Ridge-GME estimator.

**Figure 4.1:** Support intervals and ridge interval in the Ridge-GME estimator for an arbitrary ridge trace.



The solution of the GME estimator is used in the Ridge-GME estimator as a reference solution. Moreover, the ridge interval provides an admissible region for the choice of an estimate of the ridge parameter, which means that even if there exists a value outside the ridge interval such that the infinity norm of the difference between the GME and the ridge solutions is smaller than the one obtained for points inside the ridge interval, such value is never selected as an estimate of the ridge parameter. Hence, the ridge trace plays a key role in this new estimator, since it provides supports for the GME estimator and ensures that the  $\hat{\eta}$  always falls into the ridge interval (selected from visual inspection of the ridge trace). The ridge interval ensures the necessary equilibrium between the variance and the bias discussed in the previous section.

**Definition 4.4.** Taking into account Definitions 2.6, 2.7 and 4.2, the Ridge-GME estimator of  $\eta$  is given by

$$\hat{\eta}_{\text{Ridge-GME}} = \underset{\eta}{\operatorname{argmin}} \left\| \mathbf{Z}\hat{\mathbf{p}} - (\mathbf{X}'\mathbf{X} + \eta\mathbf{I})^{-1} \mathbf{X}'\mathbf{y} \right\|_{\infty} \quad (4.7)$$

subject to the ridge interval,  $\eta \in [a_1, a_2]$ , where  $a_1$  and  $a_2$  denote the lower and upper bounds of the ridge interval, respectively.<sup>4</sup>

<sup>4</sup>A MATLAB code for the Ridge-GME estimator is provided in Appendix C.

The vector  $\hat{\mathbf{p}}$  in Definition 4.4 is obtained through the GME estimator from Definition 2.7, where the matrix  $\mathbf{Z}$  is a  $(K \times KM)$  matrix of support values provided by the ridge trace as illustrated in Figure 4.1. The supports for the error component (matrix  $\mathbf{V}$ ) are usually defined by the empirical standard deviation of the noisy observations  $\mathbf{y}$ ; see subsection 2.2.1. In general, the conservative  $3\sigma$  or  $4\sigma$  rules are considered; e.g., Campbell and Hill [21] and Golan et al. [69].

#### 4.4 Simulation study

In this section, a simulation study is presented to illustrate the performance of the ridge estimator based on the Ridge-GME estimator when compared with the OLS estimator (performed through the QR decomposition)<sup>5</sup> and the ridge estimator based on the generalized cross-validation (GCV) estimator by Golub et al. [70], the HK estimator proposed by Hoerl and Kennard [76, 77], the HKB estimator by Hoerl et al. [78], the KS estimator by Khalaf and Shukur [90], and the KM4, KM5 and KM6 estimators by Muniz and Kibria [120]. Taking into account Definitions 2.6 and 4.2, all of these ridge parameter estimators are presented in the next definitions.

**Definition 4.5.** *Considering Definitions 2.6 and 4.2, the GCV estimator of the ridge parameter  $\eta$  is given by*

$$\hat{\eta}_{GCV} = \operatorname{argmin}_{\eta} \left\{ \frac{\frac{1}{N} \|(\mathbf{I} - \mathbf{A}(\eta))\mathbf{y}\|^2}{\left(\frac{1}{N} \operatorname{trace}(\mathbf{I} - \mathbf{A}(\eta))\right)^2} \right\}, \quad (4.8)$$

where  $\mathbf{A}(\eta) = \mathbf{X}(\mathbf{X}'\mathbf{X} + N\eta\mathbf{I})^{-1}\mathbf{X}'$ , and  $\hat{\boldsymbol{\beta}}_{ridge} = (\mathbf{X}'\mathbf{X} + N\eta\mathbf{I})^{-1}\mathbf{X}'\mathbf{y}$  in this case.

The remaining ridge parameter estimators are usually defined using the model (2.42) in the canonical form, i.e.,

$$\mathbf{y} = \mathbf{X}_c\boldsymbol{\alpha} + \mathbf{u}, \quad (4.9)$$

where  $\mathbf{X}_c = \mathbf{X}\mathbf{P}$ ,  $\boldsymbol{\alpha} = \mathbf{P}'\boldsymbol{\beta}$ ,  $\mathbf{P}$  is an orthogonal matrix such that  $\mathbf{P}'\mathbf{X}'\mathbf{X}\mathbf{P} = \boldsymbol{\Lambda}$  and  $\boldsymbol{\Lambda}$  is a diagonal matrix with the eigenvalues of  $\mathbf{X}'\mathbf{X}$ .

---

<sup>5</sup>This is probably the best procedure to obtain accurate OLS solutions; see Greene [71, p. 975].

**Definition 4.6.** Considering model (4.9) and Definition 4.2, the HK estimator<sup>6</sup> of the ridge parameter  $\eta$  is given by

$$\hat{\eta}_{HK} = \frac{\hat{\sigma}^2}{\hat{\alpha}_{\max}^2}, \quad (4.10)$$

where  $\hat{\alpha}_{\max}^2$  is the largest squared value of  $\hat{\alpha}_{OLS}$  (the OLS estimate of  $\alpha$ ) and  $\hat{\sigma}^2$  is an estimate of  $\sigma^2$ .

**Definition 4.7.** Considering model (4.9) and Definition 4.2, the HKB estimator of the ridge parameter  $\eta$  is given by

$$\hat{\eta}_{HKB} = \frac{K\hat{\sigma}^2}{\hat{\alpha}'_{OLS}\hat{\alpha}_{OLS}}, \quad (4.11)$$

where  $\hat{\alpha}_{OLS}$  is the OLS estimate of  $\alpha$  and  $\hat{\sigma}^2$  is an estimate of  $\sigma^2$ .

**Definition 4.8.** Considering model (4.9) and Definition 4.2, the KS estimator of the ridge parameter  $\eta$  is given by

$$\hat{\eta}_{KS} = \frac{t_{\max}\hat{\sigma}^2}{(N-K)\hat{\sigma}^2 + t_{\max}\hat{\alpha}_{\max}^2}, \quad (4.12)$$

where  $t_{\max}$  is the largest eigenvalue of the  $\mathbf{X}'\mathbf{X}$  matrix,  $\hat{\alpha}_{\max}^2$  is the largest squared value of  $\hat{\alpha}_{OLS}$  and  $\hat{\sigma}^2$  is an estimate of  $\sigma^2$ .

Considering Definition 4.2, model (4.9) and  $m_k = (\hat{\sigma}^2/\hat{\alpha}_k^2)^{\frac{1}{2}}$ , the three estimators from Muniz and Kibria [120] are defined next.

**Definition 4.9.** The KM4 estimator of the ridge parameter  $\eta$  is given by

$$\hat{\eta}_{KM4} = \left( \prod_{k=1}^K \frac{1}{m_k} \right)^{\frac{1}{K}}. \quad (4.13)$$

**Definition 4.10.** The KM5 estimator of the ridge parameter  $\eta$  is given by

$$\hat{\eta}_{KM5} = \left( \prod_{k=1}^K m_k \right)^{\frac{1}{K}}. \quad (4.14)$$

**Definition 4.11.** The KM6 estimator of the ridge parameter  $\eta$  is given by

$$\hat{\eta}_{KM6} = \text{med} \left( \frac{1}{m_k} \right), \quad (4.15)$$

where *med* represents the median.

---

<sup>6</sup>Hoerl and Kennard [76] present an estimator for  $\eta_k$ . The authors emphasize the use of the ridge trace.

Following Gibbons [61], Kibria [92] and McDonald and Galarneau [113], in order to generate different degrees of collinearity, the explanatory variables are generated through the equation

$$x_{nk} = (1 - \alpha^2)^{\frac{1}{2}} z_{nk} + \alpha z_{n6}, \quad (4.16)$$

where  $z_{nk}$ ,  $n = 1, 2, \dots, N$ ,  $k = 1, \dots, K$ , are independent standard normal pseudo-random numbers (based on the Marsaglia's ziggurat algorithm),  $K = 5$  is the number of explanatory variables, and  $\alpha$  is specified so that the correlation between any two explanatory variables is given by  $\alpha^2$ . The simulation study contemplates four different combinations of sample sizes, namely  $N = 10, 20, 50$  and  $100$ . By choosing the true coefficient vector  $\beta$  as the normalized eigenvector corresponding to the largest eigenvalue of the  $\mathbf{X}'\mathbf{X}$  matrix,<sup>7</sup> the  $N$  observations for the dependent variable are obtained by

$$y_n = \beta_1 x_{n1} + \beta_2 x_{n2} + \beta_3 x_{n3} + \beta_4 x_{n4} + \beta_5 x_{n5} + u_n, \quad (4.17)$$

where  $u_n$ ,  $n = 1, \dots, N$ , are independent normal pseudo-random numbers with mean zero and variance  $\sigma^2$ .

The variables are standardized so that the matrix  $\mathbf{X}'\mathbf{X}$  is in correlation form whereas the vector  $\mathbf{X}'\mathbf{y}$  represents the correlations of the dependent variable with each explanatory variable.<sup>8</sup> Three different values for  $\sigma$  and five different values for  $\alpha$  are tested, namely  $\sigma = 0.5, 1.0, 1.5$  and  $\alpha = 0.750, 0.900, 0.950, 0.975, 0.999$ . The estimated standardized coefficients are transformed back to the original model and, for the 1000 trials performed, the mean squared error loss (MSEL), with  $\text{SEL}(\hat{\beta}) = \|\hat{\beta} - \beta\|^2$ , is the measure used to evaluate the performance of the different estimators.

Table 4.1 – Table 4.4 show the MSEL for different ridge estimators with different ridge parameter estimators. The ridge interval for the Ridge-GME estimator is defined as  $\eta \in ]0, 1]$ . The GME estimator is performed using the support  $[-5, 5]$  for the five parameters of the model and the  $4\sigma$  rule is used to define the support for the error component, where an estimate of  $\sigma$  is obtained from the empirical standard deviation of the noisy observations  $\mathbf{y}$ .

<sup>7</sup>This choice leads to the minimization of the MSE; e.g., Gibbons [61, p. 133].

<sup>8</sup>Generally speaking, should the regressors and the dependent variable be centered and scaled so that  $\mathbf{X}'\mathbf{X}$  and  $\mathbf{X}'\mathbf{y}$  are in correlation form? This issue is quite controversial, since both options have advantages and disadvantages. Although many results in the literature are obtained from correlation forms, it seems that this choice depends on the problem under study; e.g., Belsley et al. [10], Montgomery et al. [119] and Ryan [144].

**Table 4.1:** MSEL for OLS and different ridge estimators ( $N = 10$ ).

$\alpha$	OLS	ridge HK	ridge HKB	ridge GCV	ridge KS	ridge KM4	ridge KM5	ridge KM6	ridge Ridge-GME
$\sigma = 0.5$									
0.750	0.662	0.512	0.390	0.285	0.490	0.157	0.097	0.178	0.061
0.900	2.769	1.456	1.124	0.950	1.342	0.174	0.161	0.214	0.083
0.950	3.765	1.841	1.492	1.115	1.686	0.200	0.144	0.244	0.048
0.975	5.568	2.781	2.214	1.659	2.529	0.358	0.062	0.408	0.031
0.999	67.273	29.033	25.858	20.395	26.385	0.597	0.065	0.607	0.030
$\sigma = 1.0$									
0.750	3.489	1.723	1.376	1.193	1.641	0.202	0.344	0.217	0.188
0.900	11.391	4.503	3.698	3.608	4.153	0.181	0.295	0.202	0.107
0.950	18.894	7.316	5.953	5.396	6.737	0.220	0.255	0.241	0.059
0.975	39.102	14.634	12.228	10.333	13.388	0.289	0.298	0.295	0.050
0.999	602.764	220.698	203.300	198.594	201.912	0.461	0.582	0.492	0.057
$\sigma = 1.5$									
0.750	9.477	3.931	3.132	2.960	3.693	0.218	0.442	0.226	0.213
0.900	20.431	7.981	6.682	5.669	7.341	0.191	0.512	0.201	0.132
0.950	27.808	10.657	9.453	8.364	9.881	0.194	0.767	0.193	0.192
0.975	40.101	15.880	14.088	11.818	14.578	0.224	0.650	0.226	0.099
0.999	681.248	314.637	271.974	213.268	288.182	0.508	0.693	0.562	0.061

The number of support points used for the unknown parameters and the error component are five and three, respectively.

The results in Tables 4.1 – 4.4 reveal a good performance (in terms of MSEL) of the ridge estimator based on the Ridge-GME estimator. In 46 of the 60 simulations performed, the ridge estimator based on the Ridge-GME estimator outperforms all the competitors. Moreover, in 11 of the remaining 14 simulations, the ridge estimator based on the Ridge-GME estimator is outperformed only by the ridge estimator based on the KM5 estimator. It is important to stress that the estimators KM4, KM5 and KM6 exhibit a good performance in this simulation study. For example, in several simulations where the ridge estimator based on the Ridge-GME estimator outperforms all the competitors, the MSEL of the ridge estimators

**Table 4.2:** MSEL for OLS and different ridge estimators ( $N = 20$ ).

$\alpha$	OLS	ridge HK	ridge HKB	ridge GCV	ridge KS	ridge KM4	ridge KM5	ridge KM6	ridge Ridge-GME
$\sigma = 0.5$									
0.750	0.166	0.154	0.126	0.091	0.154	0.119	0.065	0.123	0.067
0.900	0.589	0.425	0.287	0.168	0.421	0.148	0.048	0.163	0.043
0.950	0.832	0.583	0.401	0.231	0.582	0.137	0.065	0.163	0.044
0.975	2.324	1.163	0.839	0.526	1.154	0.167	0.074	0.210	0.041
0.999	37.416	13.630	10.937	5.435	13.166	0.529	0.028	0.549	0.030
$\sigma = 1.0$									
0.750	0.746	0.546	0.385	0.270	0.562	0.128	0.138	0.143	0.100
0.900	1.560	0.875	0.617	0.364	0.889	0.124	0.104	0.153	0.056
0.950	3.371	1.572	1.199	0.692	1.562	0.150	0.126	0.189	0.050
0.975	8.552	3.296	2.593	1.588	3.313	0.165	0.195	0.191	0.058
0.999	113.906	43.462	33.942	19.046	42.080	0.438	0.115	0.462	0.038
$\sigma = 1.5$									
0.750	1.220	0.766	0.543	0.378	0.838	0.144	0.202	0.156	0.140
0.900	3.413	1.702	1.240	0.805	1.809	0.130	0.244	0.151	0.108
0.950	5.835	2.416	1.942	1.140	2.545	0.130	0.258	0.151	0.088
0.975	13.787	5.426	4.348	2.674	5.561	0.151	0.386	0.163	0.103
0.999	238.179	93.334	72.369	41.096	90.501	0.387	0.264	0.422	0.050

based on the KM4, KM5 and KM6 estimators are only slightly superior than the MSEL of the ridge estimator based on the Ridge-GME estimator.

The results obtained for the OLS estimator and the different ridge estimators are, in general, consistent with the results obtained by other authors, namely (a) as the sample size increases, the MSEL decreases; (b) as  $\sigma$  increases, the MSEL increases; and (c) as  $\alpha$  increases, the MSEL increases (being this increase larger as  $\sigma$  increases). Furthermore, the results for the ridge estimator based on the Ridge-GME estimator are consistent with (a) and (b). However, as  $\alpha$  increases, in general, the MSEL decreases, regardless of the value of  $\sigma$  (note that the KM5 estimator exhibits a similar behavior). In this simulation study, the ridge estimator based on the Ridge-GME estimator outperforms all the competitors for  $\sigma = 1.0$

**Table 4.3:** MSEL for OLS and different ridge estimators ( $N = 50$ ).

$\alpha$	OLS	ridge HK	ridge HKB	ridge GCV	ridge KS	ridge KM4	ridge KM5	ridge KM6	ridge Ridge-GME
$\sigma = 0.5$									
0.750	0.064	0.062	0.056	0.039	0.062	0.117	0.036	0.106	0.049
0.900	0.121	0.114	0.095	0.047	0.114	0.135	0.022	0.131	0.037
0.950	0.235	0.207	0.154	0.063	0.207	0.162	0.020	0.169	0.033
0.975	0.431	0.340	0.231	0.090	0.341	0.198	0.019	0.217	0.032
0.999	10.385	3.737	3.040	1.429	3.719	0.510	0.012	0.553	0.029
$\sigma = 1.0$									
0.750	0.187	0.171	0.137	0.101	0.174	0.109	0.094	0.112	0.084
0.900	0.495	0.393	0.270	0.151	0.411	0.103	0.076	0.117	0.059
0.950	1.110	0.700	0.470	0.233	0.749	0.128	0.070	0.161	0.045
0.975	1.709	0.953	0.658	0.324	1.028	0.149	0.071	0.184	0.039
0.999	41.644	15.732	11.738	5.410	15.627	0.509	0.031	0.525	0.030
$\sigma = 1.5$									
0.750	0.456	0.372	0.267	0.191	0.404	0.122	0.139	0.131	0.111
0.900	0.968	0.650	0.441	0.269	0.737	0.106	0.128	0.125	0.082
0.950	2.172	1.135	0.790	0.434	1.328	0.117	0.136	0.148	0.068
0.975	3.685	1.638	1.226	0.627	1.872	0.147	0.120	0.182	0.048
0.999	92.662	33.659	25.026	11.811	33.454	0.467	0.070	0.485	0.034

and  $\sigma = 1.5$ , regardless of the values of  $\alpha$  and  $N$  (except in two cases, namely  $N = 100$ ,  $\sigma = 1.0$ , for  $\alpha = 0.75$  and  $\alpha = 0.90$ ). The simulation study reveals another interesting feature of the ridge estimator based on the Ridge-GME estimator: in Table 4.1, for  $N = 10$ , this estimator is the best in all the 15 simulations; in Table 4.2, for  $N = 20$ , this estimator is the best in 13 of the 15 simulations; in Table 4.3, for  $N = 50$ , this estimator is the best in 10 of the 15 simulations; and, finally, in Table 4.4, for  $N = 100$ , this estimator is the best in 8 of the 15 simulations. It seems that, in the case of regression models with small samples sizes affected by collinearity, the Ridge-GME estimator is probably one of the best ridge parameter estimators in the literature of ridge regression.

**Table 4.4:** MSEL for OLS and different ridge estimators ( $N = 100$ ).

$\alpha$	OLS	ridge HK	ridge HKB	ridge GCV	ridge KS	ridge KM4	ridge KM5	ridge KM6	ridge Ridge-GME
$\sigma = 0.5$									
0.750	0.027	0.027	0.026	0.022	0.027	0.126	0.039	0.104	0.059
0.900	0.056	0.054	0.049	0.029	0.054	0.137	0.019	0.121	0.037
0.950	0.103	0.098	0.083	0.037	0.098	0.152	0.015	0.149	0.033
0.975	0.233	0.208	0.157	0.057	0.208	0.219	0.013	0.225	0.031
0.999	6.257	2.535	1.973	0.964	2.563	0.509	0.010	0.567	0.028
$\sigma = 1.0$									
0.750	0.111	0.105	0.090	0.065	0.107	0.106	0.061	0.103	0.064
0.900	0.223	0.199	0.150	0.076	0.203	0.114	0.037	0.119	0.041
0.950	0.489	0.387	0.264	0.120	0.403	0.136	0.039	0.155	0.038
0.975	1.014	0.657	0.435	0.188	0.703	0.172	0.037	0.207	0.033
0.999	20.694	7.777	5.964	2.905	7.869	0.478	0.031	0.506	0.030
$\sigma = 1.5$									
0.750	0.241	0.215	0.165	0.113	0.225	0.105	0.087	0.112	0.078
0.900	0.495	0.391	0.267	0.141	0.423	0.101	0.072	0.117	0.055
0.950	1.115	0.703	0.468	0.227	0.808	0.124	0.073	0.159	0.046
0.975	1.995	1.077	0.745	0.367	1.222	0.171	0.060	0.213	0.037
0.999	47.828	17.199	13.119	5.915	17.364	0.465	0.040	0.481	0.031

## 4.5 Numerical example

In this section, the well-known Portland cement data set from Woods et al. [177] is used to illustrate the performance of the ridge estimator using the Ridge-GME estimator. This data set has received considerable attention in the literature; e.g., Hald [73], Kaçiranlar et al. [88], Liu [101], Muniz and Kibria [120] and Sakalhoğlu and Kaçiranlar [146]. The response variable is the heat evolved *per* gram of cement ( $\mathbf{y}$ ) and the four explanatory variables are the amounts of tricalcium aluminate ( $\mathbf{x}_1$ ), tricalcium silicate ( $\mathbf{x}_2$ ), tetracalcium aluminoferrite ( $\mathbf{x}_3$ ) and  $\beta$ -dicalcium silicate ( $\mathbf{x}_4$ ). The linear model without intercept presented by Woods et al. [177] does not suffer from collinearity because  $\text{cond}_2 \mathbf{X} \approx 21$ , where  $\mathbf{X}$  is the matrix

of the explanatory variables and  $\text{cond}_2$  represents the 2-norm condition number.<sup>9</sup> However, the linear model with intercept defined as

$$y_n = \beta_0 + \beta_1 x_{n1} + \beta_2 x_{n2} + \beta_3 x_{n3} + \beta_4 x_{n4} + u_n, \quad (4.18)$$

$n = 1, 2, \dots, 13$ , is affected by severe collinearity since  $\text{cond}_2 \mathbf{X} \approx 6056$ , with  $\mathbf{X}$  representing the matrix of the explanatory variables with the first column of ones.<sup>10</sup> As noted by Liu [101], this dramatic change in  $\text{cond}_2$  is explained by the fact that the sum of each row in the original  $\mathbf{X}$  matrix is approximately equal to 100 (the explanatory variables are presented in rounded percentages) and, thus, the model (4.18) is affected by severe collinearity. The error vector is assumed to be normally distributed with zero mean and constant variance. A simple inspection of a plot of the residuals against fitted values and the p-value from the Shapiro-Wilk test support these error assumptions. Based on the LTS residuals<sup>11</sup> and the Cook's squared distance, no influential observations are assumed in this regression model.

To use the Ridge-GME estimator, based on the ridge trace with non-standardized coefficients, the GME estimator is performed using the support  $[0, 100]$  for the constant and  $[0, 5]$  for the other parameters in the model, except for  $\beta_4$  whose support is  $[-1, 5]$ . The support for the error component is defined by the empirical standard deviation of the noisy observations using the  $4\sigma$  rule. The number of points in supports are ten for both the unknown parameters and the error component.

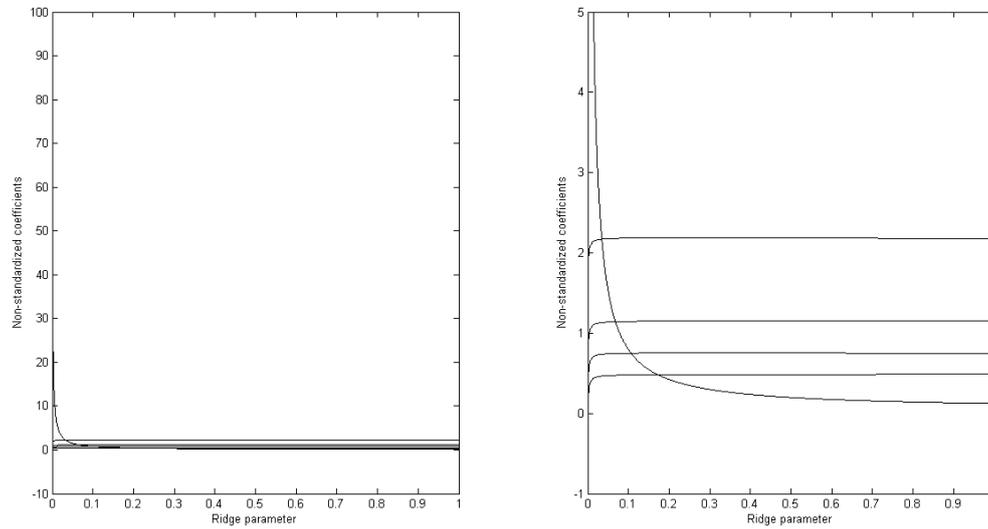
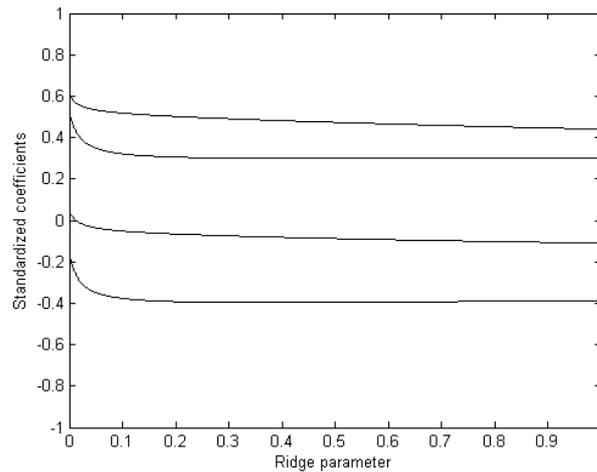
The ridge trace with non-standardized coefficients in Figure 4.2 illustrates the lack of stability of the OLS estimates and the presence of severe collinearity, i.e., large changes in the coefficients for small values of the ridge parameter. Note that the ridge trace with standardized coefficients in Figure 4.3 does not include the intercept, since the regressors and the response are centered and scaled.

From visual inspection of the ridge trace in Figure 4.2, the ridge interval is defined as  $\eta \in ]0, 0.3]$ . Figure 4.4 illustrates the selection of the ridge parameter estimate,  $\hat{\eta}$ , from the Ridge-GME estimator. The Ridge-GME estimate is  $\hat{\eta} = 0.005$  (this value is selected from

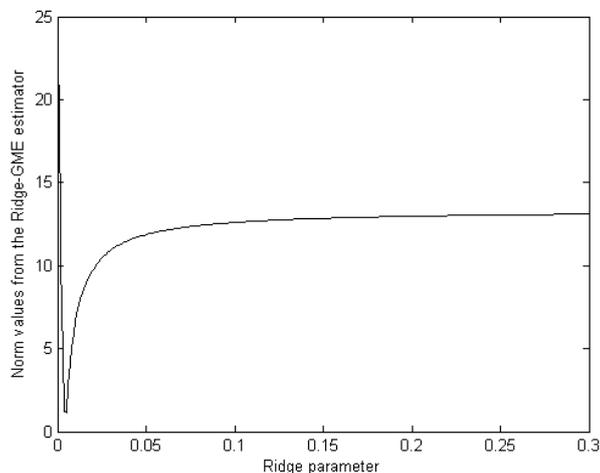
<sup>9</sup>Taking into account the footnote 8 on p. 78, note that, in this model, the 2-norm condition number increases when the  $\mathbf{X}$  matrix is centered and scaled ( $\text{cond}_2 \mathbf{X} \approx 37$ ).

<sup>10</sup>The same condition number is obtained by the definition  $k(\mathbf{X}'\mathbf{X}) = (\lambda_{\max}/\lambda_{\min})^{\frac{1}{2}}$ , where  $\lambda_{\max}$  and  $\lambda_{\min}$  are the highest and the smallest eigenvalues of  $\mathbf{X}'\mathbf{X}$ .

<sup>11</sup>Outlier diagnosis based on the OLS residuals is not always correct because OLS tries to avoid large residuals.

**Figure 4.2:** Ridge trace for the Portland cement model (non-standardized coefficients).**Figure 4.3:** Ridge trace for the Portland cement model (standardized coefficients).

a vector of 300 linear equally spaced points between 0 and 0.3) and the MSE of the ridge estimator considering the Ridge-GME estimate is approximately 2704. Table 4.5 shows the MSE for the OLS and different ridge estimators.

**Figure 4.4:** Selection of the Ridge-GME estimate for the Portland cement model.**Table 4.5:** MSE for different estimators in the Portland cement model.

	OLS	ridge HK	ridge HKB	ridge GCV	ridge KS	ridge KM4	ridge KM5	ridge KM6	ridge Ridge-GME
MSE	4912	2171	2990	3859	2180	3879	3882	3876	2704

The MSE for the ridge estimator based on the Ridge-GME estimate is greater than the MSE for the ridge estimator based on the HK and KS estimates, although it is lower than the MSE for the OLS and the others ridge estimators presented in Table 4.5. The ridge estimator based on the Ridge-GME estimate also provides a MSE less than the MSE of other ridge estimators considered by Muniz and Kibria [120, pp. 628–629] and of some estimators considered by Sakalhoğlu and Kaçiranlar [146, pp. 683–687] in the Portland cement model.

## 4.6 Conclusions

In this chapter, a new method for selecting the ridge parameter is introduced. The new estimator is based on the GME estimator and the ridge trace. The empirical application and the Monte Carlo simulation study illustrate the good performance of this new estimator.

In the simulation study, the ridge estimator based on the Ridge-GME estimator outper-

forms the OLS estimator and several other ridge estimators in most of the cases analyzed. Furthermore, the estimators KM4, KM5 and KM6 recommended by Muniz and Kibria [120] also exhibit a good performance in the simulation study, in particular the KM5 estimator.

Based on the results from the simulation study and the Portland model, it seems reasonable to state that, in the case of regression models with small samples sizes affected by collinearity, the Ridge-GME estimator is probably one of the best ridge parameter estimators in the literature of ridge regression and it may be recommended to practitioners.

Recently, the Ridge-GME estimator discussed in this chapter was analyzed and adapted for a jackknife procedure in the ridge regression by Erdugan and Akdeniz [46]. The good results found in this chapter for the Ridge-GME estimator are also achieved by these authors in their work, which supports the previous belief that the Ridge-GME estimator may be recommended to practitioners and should become part of the restricted group of ridge parameter estimators that may be considered in any ridge regression analysis with small samples sizes.

## Chapter 5

# Some developments in the maximum entropy estimation

“[...] we have found nothing wrong with the theory of quantum electrodynamics. It is, therefore, I would say, the jewel of physics – our proudest possession.”

Feynman [57, p. 8].

In this chapter, two developments in the ME estimation based on some ideas from the theory of light (quantum electrodynamics), the robust regression literature and the GME estimator are introduced. The new estimators seem to perform well in linear regression models with small samples sizes affected by collinearity and outliers.

### 5.1 Maximum entropy robust regression group estimators

#### 5.1.1 Introduction

As already mentioned in subsection 2.1.2, the ME formalism was first established by Jaynes [81, 82] based on physics (the Shannon entropy and statistical mechanics) and statistical inference. In a linear pure inverse problem, the ME principle provides the probability distribution for which the current state of knowledge is sufficient to determine the probability assignment.

The GME estimator presented in subsection 2.2.1 contributed to the development of the

ME econometrics literature in recent years. In view of the fact that ill-posed real-world problems seem to be the rule rather than the exception in applied mathematics and statistics, the GME estimator has acquired special importance in the toolkit of statistical techniques, by allowing statistical formulations free of restrictive and unnecessary assumptions. As mentioned in subsection 2.2.1, this estimator is widely used in linear regression models affected by collinearity, in models where the number of unknown parameters exceeds the number of observations, and in small samples sizes.

However, the main weakness of the GME class of estimators is that support intervals (exogenous information not always available) for the parameters and error vectors are needed. Those supports are defined as being closed and bounded intervals in which each parameter and error are restricted to lie. Giving heed to the problem of the definition of support intervals, Paris [127] developed the maximum entropy Leuven (MEL) estimator<sup>1</sup> based on some ideas from the theory of light of Feynman [57]. The MEL estimator is generated using the Shannon entropy measure and the OLS estimator. Based on the MEL estimator, as well as information theory and robust regression techniques, a general class of estimators, denoted by maximum entropy robust regression group (in short MERG) estimators, is introduced in this section.

Considering the same framework as in the MEL estimator, a general expression is defined in which the Rényi and Tsallis entropies can be also applied. Furthermore, different estimators based on robust regression, namely the least trimmed squares (LTS) estimator, the least absolute deviations (LAD) estimator, and the least median of squares (LMS) estimator, are considered; e.g., Ellis and Morgenthaler [45], Rousseeuw [141] and Rousseeuw and Leroy [142]. It is worth to mention that several other MERG estimators can be defined by merging other entropy expressions and/or other robust estimators.<sup>2</sup>

At this point one question arises: why MERG estimators? Different reasons justify this approach. First, Paris [127, 128] showed that the MEL estimator is useful in dealing with linear regression models affected by collinearity. Nonetheless, the results in subsection 5.1.5 show that the MERG estimators (which include the MEL as a particular case) outperform some traditional competitors when considering linear regression models with small samples

---

<sup>1</sup>In a later paper, Prof. Quirino Paris presented two versions of the MEL estimator; see Paris [128]. Here, only the first version of the MEL estimator is considered.

<sup>2</sup>This is an interesting topic for future research.

sizes affected by outliers and collinearity. Second, the MERG estimators allow to incorporate the Rényi and Tsallis entropies whereas the MEL estimator only uses the Shannon entropy measure. In doing so, the idea is to explore the advantages of the Rényi and Tsallis entropies over the Shannon entropy measure, expressed in the results obtained by Golan and Perloff [68] with the GME- $\alpha$  class of estimators; see subsection 2.2.3. Third, the MEL estimator only considers the OLS estimator in the objective function. Different methods widely used in robust regression can also be applied with the MERG estimators, namely the LAD, LMS and LTS estimators. For example, one advantage of the LAD estimator over the OLS estimator is its robustness in the presence of outliers in the response variable. In contrast, however, the LAD estimator is very sensitive to the presence of outliers in the independent variables. For a detailed analysis of advantages and weaknesses of LAD, LTS and LMS estimators when compared with the OLS estimator, see Huber and Ronchetti [80], Maronna et al. [109] and Rousseeuw and Leroy [142].<sup>3</sup>

### 5.1.2 The maximum entropy Leuven estimator

One important challenge when dealing with model (2.42) is the estimation of  $\beta$  under severe collinearity. As mentioned in section 2.3, collinearity inflates the variances of the estimated regression coefficients and may lead to numerical instability, meaning that statistical inference based on the regression model may be compromised.

Most of the traditional estimators developed to solve collinearity problems perform poorly in many practical situations, given their dependence on unknown parameters that must be estimated from the data; e.g., Sakalhoğlu and Kaçiranlar [146, p. 687]. The GME estimator does not depend crucially on specific unknown parameters, although it depends on exogenous information about the parameters and errors of the model in order to define their supports.

In an attempt to overcome the problem of subjective definition of support intervals in the GME estimator, Paris [127] introduced the MEL estimator.

**Definition 5.1.** *The MEL estimator of  $\beta$  in model (2.42) is given by*

$$\operatorname{argmin}_{\mathbf{p}_\beta, L_\beta, \mathbf{r}} \{ \mathbf{p}'_\beta \ln \mathbf{p}_\beta + L_\beta \ln L_\beta + \mathbf{r}' \mathbf{r} \}, \quad (5.1)$$

---

<sup>3</sup>Despite their importance, it seems that robust regression is not yet widely used in regression analysis; see, for example, Zaman et al. [179] concerning applied econometrics.

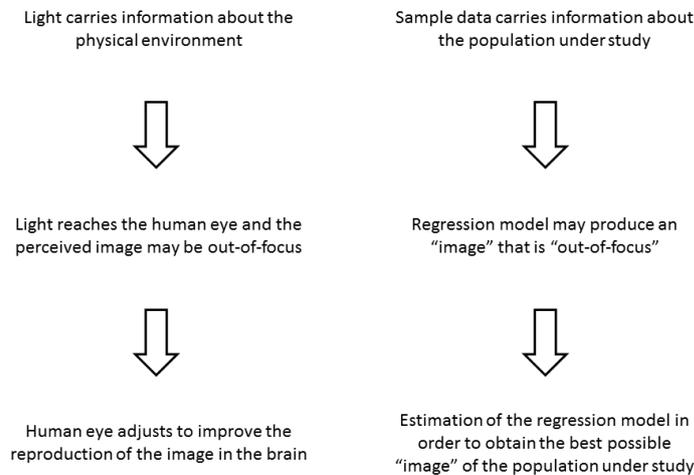
subject to

$$\begin{cases} \mathbf{y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{r} \\ L_{\beta} = \boldsymbol{\beta}'\boldsymbol{\beta} \\ \mathbf{p}_{\beta} = \frac{\boldsymbol{\beta} \ominus \boldsymbol{\beta}}{L_{\beta}} \\ \mathbf{0} < \mathbf{p}_{\beta} < \mathbf{1} \end{cases}, \quad (5.2)$$

where  $\ominus$  indicates the element-by-element Hadamard product,  $\mathbf{p}_{\beta}$  is a vector of probabilities associated to  $\boldsymbol{\beta}$  and  $\mathbf{r}'\mathbf{r}$  is the sum of squared residuals.

Note that the MEL estimator considers the Shannon entropy measure in the objective function and the OLS estimator. The MEL estimator is inspired by the theory of light and its corresponding analogy with econometric analysis can be found in Paris [127, p. 3]. Figure 5.1 presents this analogy in a more general framework (not only in economic analysis).

**Figure 5.1:** The analogy with quantum electrodynamics.



This analogy justifies Assumption 5.1 presented next. Paris [127, p. 3] states that each parameter in model (2.42)

“[...] depends on every other parameter specified in the model and its measured dimensionality is affected by the available sample information as well as by the measuring procedure. Following the theory of light, it is possible to estimate the

probability of such parameters using their revealed image. The revealed image of a parameter can be thought of as the estimable dimensionality that depends on the sample information available for the analysis.”

**Assumption 5.1.** *As in the theory of light, where the probability of a photomultiplier being hit by a photon reflected from a sheet of glass equals the square of its amplitude, the probability of  $\beta_k$  is given by the square of  $\Delta$  (i.e., the amplitude or normalized dimensionality), where*

$$\Delta = \frac{\beta_k}{\sqrt{L_\beta}}. \quad (5.3)$$

The amplitude of a photon is denoted by a vector (a “final arrow”) that summarizes the various ways in which a photon can reach the photomultiplier. Feynman [57, pp. 17–35] explains this idea with simple experiments to measure the partial reflection of light by a single or two surfaces of glass. By using arrows representing each possible way in which a photon can reach a given photomultiplier, the author illustrates how to define that “final arrow” (a sum vector) whose square represents the probability of reflection.<sup>4</sup>

The MEL estimator performs very well under collinearity and avoids the main criticism usually made to the GME estimator, i.e., the requirement of exogenous information to define the support intervals for  $\beta$  and  $\mathbf{u}$ , which is, in general, not available. Another formulation for the MEL estimator, which extends the probability specification (inspired by the theory of light) for the parameters to the error component, can be found in Paris [128].

By introducing Hölder norms, it follows from (5.2) and (5.3) that  $\beta_k$  is normalized using the Euclidean norm. More recently, Mishra [117] introduced the modular maximum entropy Leuven (MMEL) estimator in which  $\beta_k$  is normalized by taking the absolute norm. Both Paris [127] and Mishra [117] illustrated the good performance of their estimators under collinearity, using Monte Carlo simulation studies. Moreover, both estimators outperform the OLS estimator under collinearity and it seems that MMEL estimator performs better than the MEL estimator in some cases. Mishra [118] introduced a new model where, depending on the choice of different parameters, it is possible to minimize the sum of absolute residuals rather than minimizing the sum of squared residuals, as in the MEL estimator.

---

<sup>4</sup>This discussion in quantum electrodynamics is naturally beyond the scope of this thesis. The interest on quantum electrodynamics in this work lies on the analogy expressed in Assumption 5.1.

### 5.1.3 The MERG estimators

In this subsection, the class of maximum entropy robust regression group (MERG) estimators is introduced and discussed.

**Definition 5.2.** *The MERG estimators of  $\beta$  in model (2.42) are given by*

$$\operatorname{argmin}_{\mathbf{p}_\beta, L_\beta, \mathbf{r}} \left\{ \sum_{k=1}^K H_1(p_{\beta_k}) + H_2(L_\beta) + \sum_{n=1}^N \phi(r_n) \right\}, \quad (5.4)$$

*subject to the model constraint and the two constraints inspired on the theory of light,*

$$\begin{cases} y_n = \sum_{k=1}^K x_{nk} \beta_k + r_n \\ L_\beta = \sum_{k=1}^K \beta_k^2 \\ p_{\beta_k} = \frac{\beta_k^2}{L_\beta} \end{cases}, \quad (5.5)$$

where  $0 < p_{\beta_k} < 1$  is the probability of the parameter  $\beta_k$ ,  $\phi(r_n)$  is a function of the regression residuals, and the functions  $\sum_k H_1(p_{\beta_k})$  and  $H_2(L_\beta)$  are both entropy measures (e.g., Shannon, Rényi or Tsallis entropies).

Since  $H_2(L_\beta)$  is included in the objective function only to prevent the overflow of the  $L_\beta$  parameter, for simplicity, the MERG estimators are defined considering  $H_2(L_\beta) = L_\beta \ln L_\beta$  (using the Shannon entropy measure).

In models where the intercept is much larger (in absolute value) than the other parameters, Paris [128, pp. 16–17] suggests a definition of  $L_\beta$  and  $p_{\beta_k}$  only for slope parameters. The minimization problem (5.4)–(5.5) is easily adapted in such cases by a shift in  $k$ . Naturally, it is not easy to say when an intercept is “much larger” than the slope parameters.

By considering different specifications for the components in (5.4), several MERG estimators are obtained; see Table 5.1 below where the superscripts “R” and “T” denote the Rényi and Tsallis entropies, respectively.<sup>5</sup> Table 5.1 does not contain  $H_2(L_\beta)$  because, as mentioned previously, this function is defined by the Shannon entropy measure for all MERG estimators. Note that the MERG1 estimator is the MEL estimator or the MMEL estimator

<sup>5</sup>In Appendix C, a MATLAB code is provided for the MERG estimators.

**Table 5.1:** MERG estimators.

Group	$\sum_k H_1(p_{\beta_k})$	$\phi(r)$
MERG1	$\sum_k p_{\beta_k} \ln p_{\beta_k}$	$r_n^2$
MERG2	$\sum_k p_{\beta_k} \ln p_{\beta_k}$	$r_{(n:N)}^2$
MERG3	$\sum_k p_{\beta_k} \ln p_{\beta_k}$	$ r_n $
MERG4	$\sum_k p_{\beta_k} \ln p_{\beta_k}$	$med_n r_n^2$
MERG1 $^R_\alpha$	$\frac{1}{\alpha-1} \ln \sum_k (p_{\beta_k})^\alpha$	$r_n^2$
MERG1 $^T_\alpha$	$\frac{1}{1-\alpha} (1 - \sum_k (p_{\beta_k})^\alpha)$	
MERG2 $^R_\alpha$	$\frac{1}{\alpha-1} \ln \sum_k (p_{\beta_k})^\alpha$	$r_{(n:N)}^2$
MERG2 $^T_\alpha$	$\frac{1}{1-\alpha} (1 - \sum_k (p_{\beta_k})^\alpha)$	
MERG3 $^R_\alpha$	$\frac{1}{\alpha-1} \ln \sum_k (p_{\beta_k})^\alpha$	$ r_n $
MERG3 $^T_\alpha$	$\frac{1}{1-\alpha} (1 - \sum_k (p_{\beta_k})^\alpha)$	
MERG4 $^R_\alpha$	$\frac{1}{\alpha-1} \ln \sum_k (p_{\beta_k})^\alpha$	$med_n r_n^2$
MERG4 $^T_\alpha$	$\frac{1}{1-\alpha} (1 - \sum_k (p_{\beta_k})^\alpha)$	

proposed by Mishra [117] if  $\beta_k$  is normalized using the absolute norm. For this case

$$L_\beta = \sum_{k=1}^K |\beta_k| \quad \text{and} \quad p_{\beta_k} = \frac{|\beta_k|}{L_\beta}. \quad (5.6)$$

Obviously, the OLS estimator is obtained by removing the entropy components. In MERG2, the sum in  $\phi(r_n)$  has to be adjusted since  $r_{(n:N)}^2$  represents the ordered squared residuals after removing a proportion, say  $\rho$ , of the largest squared residuals. Note that the LTS estimator falls into this group if the entropy components are removed. The MERG3 estimator contains, as a special case, the LAD estimator by removing the entropy components and it is basically the same as the case  $k_1 = 1$  in the model proposed by Mishra [118].<sup>6</sup> Finally, keeping in mind that the goal of MERG4 is to minimize the median of squared residuals (instead of the sum), the sum in  $\phi(r_n)$  has to be removed. Note that the LMS estimator is obtained by removing the entropy components.

Regarding the other groups of estimators, the difference among them lies on the entropy measures considered in (5.4). Moreover, as in the MERG2 estimator, the sum in  $\phi(r_n)$  has to be adjusted in the MERG2 $^R_\alpha$  and MERG2 $^T_\alpha$  groups. Furthermore, as in the MERG4 estimator,

<sup>6</sup>Depending on the choice of other parameters such as  $k_2$  and  $k_3$ ; see Mishra [118].

the sum in  $\phi(r_n)$  has to be removed in the  $\text{MERG4}_\alpha^R$  and  $\text{MERG4}_\alpha^T$  groups. The choice of  $\alpha$  in these MERG estimators depends on the characteristics of the entropy measures; see subsection 2.1.1.

Other MERG estimators can be defined by merging other entropy expressions (e.g., Taneja [166]) and/or other robust estimators (e.g., Maronna et al. [109], Rousseeuw and Leroy [142] and Ryan [144]). This new general class of estimators attempts to create a group of estimators with high performance when dealing with regression analysis with small samples sizes exhibiting collinearity and outliers.

In order to study the structure of these estimators, the Lagrangian functions with the corresponding first-order conditions can be used. For illustrative purposes, the Lagrangian function and the first-order conditions for the MERG1 estimator are presented. In matricial form, the Lagrangian function is given by

$$L(\cdot) = \mathbf{p}'_\beta \ln \mathbf{p}_\beta + L_\beta \ln L_\beta + \mathbf{r}'\mathbf{r} + \boldsymbol{\lambda}'(\mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{r}) + \mu(L_\beta - \boldsymbol{\beta}'\boldsymbol{\beta}) + \boldsymbol{\nu}'(\mathbf{p}_\beta - \boldsymbol{\beta} \ominus \boldsymbol{\beta}/L_\beta), \quad (5.7)$$

where  $\boldsymbol{\lambda}$ ,  $\mu$  and  $\boldsymbol{\nu}$  are the Lagrange multipliers on the corresponding constraints (5.5). The first-order optimality conditions are given by

$$\frac{\partial L(\cdot)}{\partial \mathbf{p}_\beta} = \ln \mathbf{p}_\beta + \mathbf{1}_K + \boldsymbol{\nu} = \mathbf{0}, \quad (5.8)$$

$$\frac{\partial L(\cdot)}{\partial L_\beta} = \ln L_\beta + 1 + \mu + \boldsymbol{\nu}'\boldsymbol{\beta} \ominus \boldsymbol{\beta}/L_\beta^2 = 0, \quad (5.9)$$

$$\frac{\partial L(\cdot)}{\partial \mathbf{r}} = 2\mathbf{r} - \boldsymbol{\lambda} = \mathbf{0}, \quad (5.10)$$

$$\frac{\partial L(\cdot)}{\partial \boldsymbol{\lambda}} = \mathbf{y} - \mathbf{X}\boldsymbol{\beta} - \mathbf{r} = \mathbf{0}, \quad (5.11)$$

$$\frac{\partial L(\cdot)}{\partial \mu} = L_\beta - \boldsymbol{\beta}'\boldsymbol{\beta} = 0, \quad (5.12)$$

$$\frac{\partial L(\cdot)}{\partial \boldsymbol{\nu}} = \mathbf{p}_\beta - \boldsymbol{\beta} \ominus \boldsymbol{\beta}/L_\beta = \mathbf{0}. \quad (5.13)$$

The MERG estimators do not possess a closed-form representation and the solutions must be found numerically by means of nonlinear optimization techniques; see, for example, Brinkhuis and Tikhomirov [18] for a review in optimization.

The MERG estimators represent a non-standard approach to the collinearity problem in the linear regression model, but they may be regarded as belonging to the class of regulari-

zation methods.<sup>7</sup> Thus, it is interesting to compare the MERG estimators with some other traditional regularization methods that are related to (or make use of) maximum entropy; see, for example, Donoho et al. [41], Gamboa and Gassiat [60], Golan [63] and Hastie et al. [75] for further details. The discussion provided in Golan [64, 65] is also important to understand the connections among the different information-based theoretical methods.

The estimator in Donoho et al. [41], presented in the next definition, seems to have some similarities with the MERG estimators, specifically the MERG1 (i.e., the MEL) estimator.

**Definition 5.3.** *The ME estimator of  $\boldsymbol{\beta}$  in model (2.42) is given by*

$$\operatorname{argmin}_{\boldsymbol{\beta}} \left\{ \|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + 2\lambda \sum_{k=1}^K \beta_k \log \beta_k \right\}, \quad (5.14)$$

where  $\lambda$  is a regularization parameter and  $\beta_k$  must be non-negative for all  $k$ ,  $k = 1, 2, \dots, K$ .

Besides the robust regression methods and the different entropy measures used in some MERG estimators, the main difference between the MERG estimators and the ME estimator in Donoho et al. [41] lies on the probability specification for the parameters inspired by the theory of light. Given this strategy developed by Paris [127], the MERG estimators are not restricted to problems with  $\boldsymbol{\beta} \in \mathbb{R}_+^K$ . Although based on Assumption 5.1, the MERG estimators can be considered as a generalization of the ME estimator in Donoho et al. [41]. Future research may establish important connections between these information-based estimators.

#### 5.1.4 Some properties of the MERG estimators

Since the MERG estimators are based on ME and robust regression techniques, its properties can be derived from the ones already established in ME (GME and GME- $\alpha$ ) and in the regression (OLS, LTS, LAD and LMS) literature. Keeping that in mind, scale invariance, consistency and asymptotic normality are discussed for some MERG estimators.

In regression analysis, equivariance is a property which allows the evaluation of estimate changes due to data transformations and it is usually divided in regression, scale and affine

---

<sup>7</sup>Generally speaking, a regularization method is a technique to stabilize the solution of an ill-posed problem. In general, from model (2.42), a regularization method is given by the minimization of  $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\| + \lambda\mathcal{P}(\boldsymbol{\beta})$ , where  $\mathcal{P}(\boldsymbol{\beta})$  is a penalty function that represents the constraints on  $\boldsymbol{\beta}$  and  $\lambda$  is a regularization parameter that denotes a trade-off between the two components of the objective function. For example, ridge regression discussed in Chapter 4 is a regularization method.

equivariance. It is well-known that OLS, LAD, LMS, and LTS estimators are regression, scale and affine equivariant. However, it is important to note that, in some cases, this property is relaxed.

Following Paris [128, pp. 11–14], it is possible to show that the MERG estimators with the Shannon entropy measure are scale invariant, in the sense that when the model (2.42) is scaled as

$$c\mathbf{y} = (c\mathbf{X}\mathbf{A}^{-1})\boldsymbol{\beta}_s + \mathbf{u}_s, \quad (5.15)$$

where  $c$  is a known arbitrary constant and  $\mathbf{A}$  is a known non-singular square matrix, the original solution  $\boldsymbol{\beta}$  of model (2.42) can be obtained from the solution  $\boldsymbol{\beta}_s$  of model (5.15).

**Proposition 5.1.** *The MERG estimators with Shannon entropy are scale invariant.*

*Proof.* The complete proof for MERG1 (i.e., MEL) is available in Paris [128, pp. 11–14]. The proof for the others MERG estimators with Shannon entropy is analogous and, thus, it is omitted here. However, some particular details are highlighted next. From the objective function in (5.4), it follows that the MERG2 estimator takes the form

$$\operatorname{argmin}_{\mathbf{p}_{\beta}, L_{\beta}, \mathbf{r}} \left\{ \sum_{k=1}^K p_{\beta_k} \ln p_{\beta_k} + L_{\beta} \ln L_{\beta} + \sum_{n=1}^{[(1-\rho)N]+1} r_{(n:N)}^2 \right\}. \quad (5.16)$$

Since the MERG1 estimator is scale invariant, it follows that the MERG2 estimator is also scale invariant, because  $r_{(n:N)}^2$  just represents the ordered squared residuals after removing a proportion of the largest squared residuals. On the other hand, from the objective function in (5.4), the MERG3 estimator assumes the form

$$\operatorname{argmin}_{\mathbf{p}_{\beta}, L_{\beta}, \mathbf{r}} \left\{ \sum_{k=1}^K p_{\beta_k} \ln p_{\beta_k} + L_{\beta} \ln L_{\beta} + \sum_{n=1}^N |r_n| \right\}, \quad (5.17)$$

and, following the same line of reasoning as in Paris [128, pp. 11–14], it is necessary to show that the solution of the unscaled model can be found using the solution of the scaled model. The Lagrangian function and the corresponding first-order conditions are similar to those for the MERG1 estimator, i.e., the conditions defined by (5.8)–(5.13). Formally, the main difference between both problems lies on the partial derivatives of  $L(\cdot)$  with respect to the residual variable,

$$\frac{\partial L(\cdot)}{\partial \mathbf{r}} = 2\mathbf{r} - \boldsymbol{\lambda} = \mathbf{0} \text{ [MERG1]} \quad \text{and} \quad \frac{\partial L(\cdot)}{\partial \mathbf{r}} = \mathbf{1} - \boldsymbol{\lambda} = \mathbf{0} \text{ [MERG3]}, \quad (5.18)$$

where  $\boldsymbol{\lambda}$  is the vector of Lagrange multipliers for the first constraint in (5.5). This difference simplifies the proof conducted by Paris [128], meaning that the MERG3 estimator is scale invariant, such as the MERG1 estimator. Analogous arguments can be used for the MERG4 estimator given by the objective function

$$\operatorname{argmin}_{\mathbf{p}_\beta, L_\beta, \mathbf{r}} \left\{ \sum_{k=1}^K p_{\beta_k} \ln p_{\beta_k} + L_\beta \ln L_\beta + \operatorname{med}_n r_n^2 \right\}. \quad (5.19)$$

A similar discussion on the partial derivatives of  $L(\cdot)$  with respect to the residual variable can be followed. However, the similarity between the MERG1 and MERG4 estimators can be achieved from another point of view. Taking into account that the minimization of the sum of squared residuals in the OLS estimator (used in the MERG1 estimator) is equivalent to the minimization of the mean of the squared residuals, the LMS estimator (used in the MERG4 estimator) just replaces the mean by the median of the squared residuals.  $\square$

In proving consistency and asymptotic normality, it is possible to show that the asymptotic properties of the OLS, LTS, LAD or LMS estimators carry over to the respective MERG estimators in Definition 5.2.

**Proposition 5.2.** *The probability limit of the entropy components with the Shannon entropy measure in the objective function of the MERG estimators tends to zero as  $N \rightarrow \infty$ .*

*Proof.* The proof conducted by Paris [128, pp. 20–22] for the MEL estimator is sufficient to prove this proposition. Assuming finite bounds on every component of the estimators and applying the probability limit, it follows that

$$\operatorname{plim}_{N \rightarrow \infty} \frac{1}{N} \sum_{k=1}^K p_{\beta_k}^N \ln p_{\beta_k}^N = 0 \quad \text{and} \quad \operatorname{plim}_{N \rightarrow \infty} \frac{1}{N} L_\beta^N \ln L_\beta^N = 0, \quad (5.20)$$

where the superscript “ $N$ ” indicates the dependence on the sample size  $N$ .

Note: this result is also valid for the MERG estimators with the Rényi and Tsallis entropies.<sup>8</sup>

$\square$

Proposition 5.2 is useful to derive the asymptotic properties of the MERG estimators, by taking into account the asymptotic properties of the OLS, LTS, LAD and LMS estimators. Note that MERG4, for example, is not asymptotically normal since the LMS estimator is not

<sup>8</sup>In the context of the thesis, i.e., for  $\alpha > 1$ ; see footnote 7 on p. 20.

asymptotically normal either. In contrast, the MERG2 estimator is consistent and asymptotically normal since, by Proposition 5.2, this estimator has the same asymptotic properties as the LTS estimator (which is consistent and asymptotically normal).

In real-world empirical applications and for inference purposes, researchers can use the asymptotic properties of the MERG estimators, established by Proposition 5.2. However, since the MERG estimators are suitable for regression models with small samples sizes, the bootstrap method may be recommended for statistical inference; see, for example, Greene [71, pp. 407–408] that also suggests bootstrap<sup>9</sup> for the LAD estimator, or Maronna [108, p. 52] that suggests bootstrap for the robust ridge regression estimator based on repeated M-estimation (the RR-MM estimator) with small samples sizes.

For example, the bootstrap estimator for the asymptotic covariance matrix can be computed as

$$\text{Var}(\hat{\boldsymbol{\beta}}_{\text{MERG}}) = \frac{1}{T} \sum_{t=1}^T (\hat{\boldsymbol{\beta}}_{\text{MERG}}(t) - \hat{\boldsymbol{\beta}}_{\text{MERG}})(\hat{\boldsymbol{\beta}}_{\text{MERG}}(t) - \hat{\boldsymbol{\beta}}_{\text{MERG}})', \quad (5.21)$$

where  $\hat{\boldsymbol{\beta}}_{\text{MERG}}$  is the MERG estimator and  $\hat{\boldsymbol{\beta}}_{\text{MERG}}(t)$  is the  $t$ th MERG estimate of  $\boldsymbol{\beta}$  based on a sample of  $N$  observations drawn with replacement from the original sample.<sup>10</sup>

### 5.1.5 Simulation study: collinearity and outliers

In this simulation study, the performance of the MERG estimators is compared with other possible competitors. The set of possible competitors includes some estimators discussed previously (e.g., GME, ridge) as well as the iteratively reweighted least squares (IRLS), the Liu-type and the RR-MM estimators. For the sake of completeness, the latter estimators are defined next.

**Definition 5.4.** *The IRLS estimator of  $\boldsymbol{\beta}$  in model (2.42) is given by*

$$\hat{\boldsymbol{\beta}}_{\text{IRLS}}^{(i+1)} = (\mathbf{X}'\mathbf{W}^{(i)}\mathbf{X})^{-1}\mathbf{X}'\mathbf{W}^{(i)}\mathbf{y}, \quad (5.22)$$

where  $\mathbf{W}^{(i)}$  is a  $(N \times N)$  diagonal matrix of weights (of the residuals) in the  $i$ th iteration.

<sup>9</sup>Greene [71, p. 407] states: “Since the efficacy of the bootstrap has been established for this purpose, the search for a formula for standard errors of the LAD estimator is not really necessary.”

<sup>10</sup>A MATLAB code to estimate standard errors using bootstrap is presented in Appendix C.

The weights assigned to the residuals at each iteration are calculated by applying robust criterion functions (Tukey's biweight, Andrews' wave, Huber, among others) to the residuals from the previous iteration; e.g., Maronna et al. [109].

**Definition 5.5.** *The Liu-type estimator of  $\boldsymbol{\beta}$  in model (2.42) is given by*

$$\widehat{\boldsymbol{\beta}}_{\eta,d} = (\mathbf{X}'\mathbf{X} + \eta\mathbf{I})^{-1}(\mathbf{X}'\mathbf{y} - d\widehat{\boldsymbol{\beta}}), \quad (5.23)$$

where  $\eta > 0$  and  $d \in \mathbb{R}$  are tuning parameters,  $\mathbf{I}$  is a  $(K \times K)$  identity matrix, and  $\widehat{\boldsymbol{\beta}}$  is any estimator of  $\boldsymbol{\beta}$ ; e.g., Liu [101].

Different choices of  $\eta$  and  $d$  are discussed in Liu [101, pp. 1013–1014]. When  $\text{cond}_2 \mathbf{X} \geq 10$ ,  $\eta$  and  $d$  can be estimated by

$$\widehat{\eta} = \frac{\lambda_1 - 100\lambda_K}{99}, \quad (5.24)$$

and

$$\widehat{d} = \frac{\sum_{k=1}^K \frac{\widehat{\sigma}^2 - \widehat{\eta}\widehat{\alpha}_k^2}{(\lambda_k + \widehat{\eta})^2}}{\sum_{k=1}^K \frac{\widehat{\sigma}^2 + \lambda_k\widehat{\alpha}_k^2}{\lambda_k(\lambda_k + \widehat{\eta})^2}}, \quad (5.25)$$

where  $\widehat{\boldsymbol{\alpha}}$  represents the ridge estimator of model (2.42) in canonical form (see model (4.9)),  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_K > 0$  are the ordered eigenvalues of  $\mathbf{X}'\mathbf{X}$ , and  $\widehat{\sigma}^2$  is an estimate of the variance using the residuals from the ridge estimator.

**Definition 5.6.** *The RR-MM estimator of  $\boldsymbol{\beta}$  in model (2.42), that combines ridge regression and MM estimation, is given by*

$$\widehat{\boldsymbol{\beta}}_{RR-MM} = \underset{\boldsymbol{\beta}}{\text{argmin}} \left\{ \widehat{\sigma}_{ini}^2 \sum_{n=1}^N \rho \left( \frac{r_n(\boldsymbol{\beta})}{\widehat{\sigma}_{ini}} \right) + \eta \|\boldsymbol{\beta}_1\|^2 \right\}, \quad (5.26)$$

where  $\boldsymbol{\beta} = (\beta_0, \boldsymbol{\beta}_1)'$ ,  $\widehat{\sigma}_{ini}$  is an  $M$ -scale estimate,  $\rho$  is the Tukey's biweight  $\rho$ -function and  $\eta$  is a penalty parameter (ridge parameter); e.g., Maronna [108].<sup>11</sup>

It is important to note that, in Definition 5.6, if

$$\widehat{\sigma}_{ini}^2 \sum_{n=1}^N \rho \left( \frac{r_n(\boldsymbol{\beta})}{\widehat{\sigma}_{ini}} \right) \quad (5.27)$$

<sup>11</sup>Prof. Ricardo Maronna provides a MATLAB code for the RR-MM estimator.

is replaced by  $\|\mathbf{y} - \mathbf{X}\boldsymbol{\beta}\|^2$ , the RR-MM estimator reduces to the ridge regression estimator. The RR-MM estimator in Maronna [108] is one of the most recent, complete and powerful estimators in the literature, concerning the estimation of regression models affected by collinearity and outliers.<sup>12</sup> Maronna [108, pp. 48–49] discusses the good performance of the RR-MM estimator as well as the drawbacks of other competitors, namely the estimators in Silvapulle [155] and Simpson and Montgomery [156].

In this simulation study, a pseudo-random number generator is used to define a  $(40 \times 3)$  matrix  $\mathbf{X}$  from a normal distribution with zero mean and unit standard deviation. Using the singular value decomposition (see Definition 2.12), the singular values of  $\mathbf{X}$  in the diagonal matrix, obtained from the decomposition, are changed to define a matrix  $\mathbf{X}_1$  with any desired condition number specified *a priori*. In this experiment,  $\text{cond}_2 \mathbf{X}_1 = 500$ . Finally, a column of ones is added to  $\mathbf{X}_1$  to define a  $(40 \times 4)$  matrix  $\mathbf{X}_2$ , whose  $\text{cond}_2 \mathbf{X}_2 \approx 1600$ .

The model is given by  $\mathbf{y} = \mathbf{X}_2\boldsymbol{\beta} + \mathbf{u}$ , where  $\boldsymbol{\beta} = (0.7, 0.1, -0.8, 0.5)'$  and  $\mathbf{u}$  is a vector of  $N(0, 1)$  errors added to form the vectors of noisy observations  $\mathbf{y}$  in each Monte Carlo trial. To create a small proportion of regression outliers, the first two elements in the second column of  $\mathbf{X}_2$  are replaced with pseudo-random values drawn from a uniform distribution  $U(10, 15)$ . After incorporating the outliers,  $\text{cond}_2 \mathbf{X}_2 \approx 98$ . In this case, outliers reduce the collinearity problem, i.e., the magnitude of the relation between the independent variables is reduced.

In each Monte Carlo trial, the first two elements of  $\mathbf{y}$  are replaced with pseudo-random values drawn from a uniform distribution  $U(10, 15)$ . For the 1000 trials performed, the mean squared error loss (MSEL), with  $\text{SEL} = \|\boldsymbol{\beta} - \widehat{\boldsymbol{\beta}}\|^2$ , is the measure used to evaluate the performance of the LTS (with  $\rho = 0.1$ ), LAD, LMS, IRLS,<sup>13</sup> OLS, ridge, Liu-type, RR-MM, GME and some MERG estimators (the MERG2 class with  $\rho = 0.1$ ); see Table 5.2.

The MERG estimators with the Tsallis and Rényi entropies are performed with  $\alpha = 4$ .<sup>14</sup> The GME estimators are performed with two different supports for the parameters,  $[-10, 10]$  and  $[-5, 5]$ , with  $M = 5$ . The supports for the errors are defined by the  $3\sigma$  rule with  $J = 3$ . The ridge estimators are performed with the HKB, KM5 and Ridge-GME estimators. The

---

<sup>12</sup>Maronna [108] emphasizes also the good performance of the ridge estimator over many other traditional competitors in models only affected by collinearity.

<sup>13</sup>The weights at each iteration are calculated by applying Tukey's biweight function to the residuals from the previous iteration.

<sup>14</sup>In the estimators with the Tsallis and Rényi entropies,  $\alpha = 4$  is always used. A detailed study on the impact of this choice in the estimation is left for future research.

**Table 5.2:** MSEL in the simulation study with outliers and collinearity.

	MSEL		MSEL		MSEL
MERG1	2.7461	MERG1 <sub>4</sub> <sup>R</sup>	2.2679	OLS	226.6157
MERG2	1.4641	MERG2 <sub>4</sub> <sup>R</sup>	1.4407	ridge (HKB)	44.3252
MERG3	2.5598	MERG3 <sub>4</sub> <sup>R</sup>	2.0023	ridge (KM5)	3.0146
MERG4	2.3033	MERG4 <sub>4</sub> <sup>R</sup>	1.3161	ridge (Ridge-GME)	3.6274
MERG1 <sub>4</sub> <sup>T</sup>	2.2679	LTS	116.7541	GME [-10, 10]	12.2597
MERG2 <sub>4</sub> <sup>T</sup>	1.4407	IRLS	153.8387	GME [-5, 5]	4.1983
MERG3 <sub>4</sub> <sup>T</sup>	1.9822	LAD	198.3922	Liu-type	2.1369
MERG4 <sub>4</sub> <sup>T</sup>	1.2866	LMS	105.3665	RR-MM	96.6137

ridge interval for the Ridge-GME estimator is defined as  $\eta \in ]0, 1]$ . The corresponding GME estimator is performed using the support  $[-5, 5]$  for the parameters (with  $M = 5$ ) and the  $3\sigma$  rule to define the support for the error component (with  $J = 3$ ).

Three important conclusions emerge from this simulation study. First, the MERG estimators perform very well and rival with two ridge estimators (using the KM5 and Ridge-GME parameter estimates), the Liu-type estimator<sup>15</sup> and the GME estimator (with the support of smaller amplitude).

Second, outliers can mask the presence of collinearity (in this case, they reduce collinearity, but the problem still remains), which means that the use of traditional robust estimators, without a careful diagnostic of collinearity problems, can be misleading. Note that the LTS, IRLS, LAD and LMS estimators perform poorly in this experiment.

Third, the MERG estimators outperform the RR-MM estimator, which performs poorly in this simulation study.<sup>16</sup> This is an important result because it highlights the performance of the MERG estimators for the combined collinearity-outliers problem in regression analysis with small samples sizes.

Another simulation study, similar to the previous one, is performed with a sample size of  $N = 30$ ,  $\text{cond}_2 \mathbf{X}_2 \approx 200$  and an outlier contamination of 20% only in  $\mathbf{y}$ . In the LTS estimator

<sup>15</sup>Other ridge and Liu-type estimators are also considered, but the results (not reported here) are very poor. These estimators depend on some parameters that must be estimated from the sample and the results are sensitive to the quality of these parameter estimates.

<sup>16</sup>All the MSEL values presented for the RR-MM estimator (in Table 5.2 and others where it is used) are calculated using a 10% upper trimmed average; see Maronna [108, p. 49]. The real values of MSEL are higher.

and the MERG2 class estimators,  $\rho = 0.25$  (the usual default value in statistical software using LTS) is used. The results from this experiment are qualitatively the same as the ones in Table 5.2. The best results (with MSEL less than 200) are achieved by the MERG estimators, with MSEL values ranging between 0.8359 (MERG3<sub>4</sub><sup>T</sup>) and 6.8959 (MERG1), whereas the MSEL for the RR-MM estimator is 29.7082. For comparison purposes, the MSEL for the OLS estimator is approximately 14236 in this experiment!

Based on these two experiments, it becomes clear the enormous difficulty in the estimation of regression models affected by outliers and collinearity. The interaction among different proportions of outliers contamination, different kind of outliers and different magnitudes in the relations among regressors makes the estimation of regression models a very difficult task. Even the best estimators, such as the RR-MM estimator, suffer from this interaction. Surprisingly, the MERG estimators reveal a high stability in different scenarios. In addition, the MERG estimators are very easy to compute and no relevant prior information is needed in order to implement them. Despite these results, more simulation studies are needed to evaluate the performance of the MERG estimators.

### 5.1.6 Examples and additional simulation studies

The aim of this subsection is to illustrate the performance of the MERG estimators when only outliers or collinearity are present in the linear regression model.

#### 5.1.6.1 HDI and Portland cement models

The first example is a data set containing information on 30 countries with the gross domestic product (GDP) *per capita* greater than 10000 PPP US\$ (purchasing power parity US dollar) and the respective human development index (HDI), both in 2005. The data is collected from the Human Development Report 2007/2008, published by the United Nations Development Programme. The HDI model<sup>17</sup> is defined as

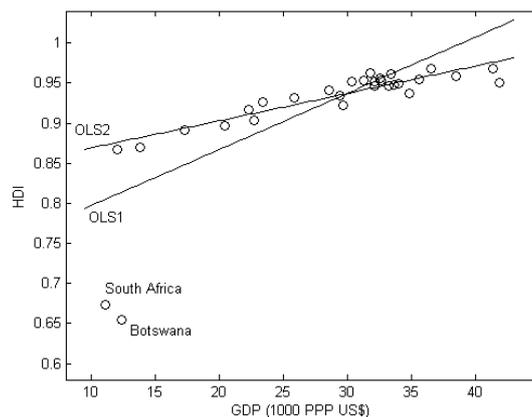
$$\text{HDI}_n = \beta_0 + \beta_1 \text{GDP}_n + u_n, \quad n = 1, \dots, 30. \quad (5.28)$$

---

<sup>17</sup>The HDI is a composite index measuring the average achievements by a country in some basic dimensions of human development, namely a long and healthy life, knowledge and a decent standard of living. The model defined here is only for illustrative purposes.

Figure 5.2 shows the OLS regression lines for the sample with 30 countries (OLS1) and the sample where South Africa and Botswana are removed (OLS2).

**Figure 5.2:** OLS regression lines in the HDI model.



**Table 5.3:** Estimates for  $\beta_0$  and  $\beta_1$  in the HDI model.

	$\hat{\beta}_0$	std. error	$\hat{\beta}_1$	std. error
OLS1	0.7269	0.0307	0.0067	0.0010
OLS2	0.8354	0.0086	0.0034	0.0003
IRLS	0.8331	0.0073	0.0035	0.0002
LTS	0.8292	0.0076	0.0036	0.0003
LAD	0.8201	0.0102	0.0039	0.0003
LMS	0.8654	0.1324	0.0026	0.0042
MERG1	0.7261	0.0294	0.0067	0.0009
MERG2	0.8284	0.0074	0.0037	0.0002
MERG3	0.8201	0.0101	0.0039	0.0003
MERG4	0.8126	0.0589	0.0040	0.0019

Table 5.3 summarizes the estimates and their corresponding standard errors for OLS1, OLS2, IRLS, LTS (with  $\rho = 0.1$ ), LAD, LMS, MERG1, MERG2 (with  $\rho = 0.1$ ), MERG3 and MERG4 estimators. The estimates of standard errors are obtained by resampling residuals with 1000 bootstrap data samples.<sup>18</sup>

<sup>18</sup>For comparison purposes, the same procedure is used for all estimators. The standard errors applying the

A closer look to Table 5.3 reveals that the results for the MERG1 (MEL) estimator are similar to those for the OLS1 estimator. Using robust methods in the non-entropy component of the objective function, the other MERG estimators outperform the OLS1 and rival with the robust regression estimators. Note that the results for the OLS2 estimator (i.e., the case without outliers) are used as the reference.

The second example consists in the estimation of the Portland cement model, already discussed in section 4.5, p. 82. It is assumed that there is no influential observations in this regression model. Even when there are no outliers, the MERG class of estimators with robust regression methods can be used. However, it is not expected significant differences between the results for this class of estimators and the results for the MERG1 class of estimators.

Considering the original Portland cement model, Table 5.4 presents the results for the OLS and some MERG estimators.<sup>19</sup> Results in Table 5.4 indicate that MERG estimators perform well in the original Portland cement model, i.e., a model with weak collinearity.

**Table 5.4:** Estimates in the original Portland cement model.

	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
OLS	2.1930	1.1533	0.7585	0.4863
MERG1	2.1531	1.1625	0.7302	0.4923
MERG3	2.0070	1.1844	0.6873	0.5202
MERG1 <sub>4</sub> <sup>T</sup>	2.1537	1.1624	0.7303	0.4922
MERG3 <sub>4</sub> <sup>T</sup>	2.0930	1.2055	0.7901	0.4634

Next, the Portland cement model with an intercept, in (4.18), is estimated using the OLS, Liu-type and ridge estimators as well as some MERG and GME estimators. Results for the Portland cement model with an intercept concerning different ridge and Liu-type estimators can be found in Liu [101], Muniz and Kibria [120] and Sakalhoğlu and Kaçiranlar [146].

The GME estimators are performed with the supports defined in Table 5.5 for the parameters, considering different levels of prior information available. Five points are used in the supports for the parameters ( $M = 5$ ) and three points are defined in the supports for the  $t$ -distribution are also calculated for the traditional estimators, but the results are similar to those from the bootstrap (with differences less than 0.001).

<sup>19</sup>Results for the MERG estimators with the Rényi entropy are not reported since they are similar to the results for the estimators with the Tsallis entropy.

error component ( $J = 3$ ), using the  $3\sigma$  rule and the empirical standard deviation of the noisy observations.

**Table 5.5:** Parameter supports for GME estimators in the Portland cement model.

	$\beta_0$	$\beta_1$	$\beta_2$	$\beta_3$	$\beta_4$
S1	$[-100, 100]$	$[-100, 100]$	$[-100, 100]$	$[-100, 100]$	$[-100, 100]$
S2	$[-5, 5]$	$[-5, 5]$	$[-5, 5]$	$[-5, 5]$	$[-5, 5]$
S3	$[-1, 1]$	$[0, 5]$	$[0, 3]$	$[0, 2]$	$[0, 2]$

The HK, HKB, KM5 and Ridge-GME are the ridge parameter estimators considered in this example for the ridge regression. The ridge interval for the Ridge-GME estimator is defined by  $\eta \in ]0, 0.3]$ , as in section 4.5. The GME estimator associated with the Ridge-GME estimator is performed using the support  $[-100, 100]$  for the constant and  $[-10, 10]$  for the other parameters of the model (with  $M = 5$ ). The supports for the error component are defined by the empirical standard deviation of the noisy observations using the  $3\sigma$  rule (with  $J = 3$ ). More conservative supports than the ones used in section 4.5 are used in order to illustrate different interpretations of the ridge trace.

Considering that the model with an intercept is not correct in the original context defined by Woods et al. [177], it can be viewed as an illustration of a real problem with a wrong model specification. Thus,  $\beta_0$  should be zero and the performance of the estimators considered in this example are compared with the results for the OLS estimator in the original model (without intercept); see the same strategy in Liu [101].

Results in Table 5.6 reveal the good performance of the MERG estimators in the model with an intercept. For the MERG estimators, the estimate of  $\beta_0$  is close to zero and the sign of  $\hat{\beta}_4$  is positive (it is negative in the OLS estimator). Considering the infinity norm of the difference between the vector of OLS estimates of the original model in Table 5.4 and the vector of estimates of the model with an intercept in Table 5.6, the three best results are obtained for  $\text{MERG1}_4^T$ , MERG1 and the ridge estimator with the KM5 parameter estimate.

It seems that the MERG estimators can be a good choice in linear regression models with small samples sizes under severe collinearity. Although the ridge and Liu-type estimators are also proper choices to deal with collinearity, their practical implementation is often dis-

**Table 5.6:** Estimates in the Portland cement model with intercept.

	$\hat{\beta}_0$	$\hat{\beta}_1$	$\hat{\beta}_2$	$\hat{\beta}_3$	$\hat{\beta}_4$
OLS	62.4054	1.5511	0.5102	0.1019	-0.1441
MERG1	0.1003	2.1521	1.1615	0.7291	0.4913
MERG3	0.0433	2.1176	1.1582	0.8711	0.4564
MERG1 $_4^T$	0.0728	2.1530	1.1616	0.7296	0.4915
MERG3 $_4^T$	0.0484	2.0519	1.1724	0.8040	0.4797
ridge (HK)	27.6068	1.9090	0.8688	0.4680	0.2075
ridge (HKB)	8.5870	2.1046	1.0648	0.6681	0.3996
ridge (KM5)	0.1107	2.1756	1.1560	0.7456	0.4877
ridge (Ridge-GME)	3.6270	2.1555	1.1160	0.7202	0.4497
Liu-type	0.0308	0.8574	1.4055	0.0789	0.6439
GME-S1	0.3649	2.1860	1.1504	0.7518	0.4832
GME-S2	0.0382	1.4116	1.3224	0.2563	0.5947
GME-S3	-0.0080	2.2726	1.0804	0.8869	0.5885

couraged due to the fact that, as stressed by Sakalhoğlu and Kaçıranlar [146, p. 687], these estimators depend upon the unknown parameters  $\beta_k$  and  $\sigma^2$ , as well as the choice of  $\eta$  and  $d$ .<sup>20</sup> When these parameters are replaced by their corresponding estimates, the solutions of these estimators may be substantially affected by using different parameter estimators. On other hand, the problem of choosing arbitrarily the supports for the parameters is a possible disadvantage of the GME estimator.

### 5.1.6.2 Additional simulation studies

Based on the first simulation study presented in subsection 5.1.5 (with  $N = 40$ ), another simulation study is conducted with  $\text{cond}_2 \mathbf{X}_2 \approx 5$  and only two outliers in  $\mathbf{y}$ . Table 5.7 presents the results where, as expected, the LTS, IRLS and LAD estimators perform well. Unexpectedly, the LMS estimator performs poorly. However, this result is not entirely surprising since this estimator is usually not recommended as a stand-alone regression procedure. Finally, the MERG estimators reveal the best performance in this experiment!

<sup>20</sup>The ridge regression estimator with the Ridge-GME estimator discussed in Chapter 4 avoids such criticism.

**Table 5.7:** MSEL in the simulation study with outliers.

	MSEL		MSEL		MSEL
MERG1	1.6756	MERG3 <sub>4</sub> <sup>T</sup>	0.8525	LTS	3.5814
MERG2	0.9978	MERG4 <sub>4</sub> <sup>T</sup>	1.2930	IRLS	3.1911
MERG3	0.9012	MERG1 <sub>4</sub> <sup>R</sup>	1.7069	LAD	5.1867
MERG4	1.1957	MERG2 <sub>4</sub> <sup>R</sup>	0.9774	LMS	13.7382
MERG1 <sub>4</sub> <sup>T</sup>	1.7069	MERG3 <sub>4</sub> <sup>R</sup>	0.9026	OLS	24.4410
MERG2 <sub>4</sub> <sup>T</sup>	0.9774	MERG4 <sub>4</sub> <sup>R</sup>	1.2723		

Another experiment is performed with  $N = 40$ ,  $\text{cond}_2 \mathbf{X}_2 \approx 250$  and no outliers. As expected, the OLS estimator performs poorly in this ill-conditioned model. The good results for the MERG estimators and some competitors are presented in Table 5.8. Although there are no outliers, the MERG class of estimators with robust methods can be applied. In this experiment,  $\rho = 0.05$  is used in the MERG2 class of estimators.

**Table 5.8:** MSEL in the simulation study with collinearity.

	MSEL		MSEL		MSEL
MERG1	0.7801	MERG3 <sub>4</sub> <sup>T</sup>	0.7114	OLS	1548.9017
MERG2	0.8324	MERG4 <sub>4</sub> <sup>T</sup>	1.0739	ridge (KM5)	0.9870
MERG3	0.7057	MERG1 <sub>4</sub> <sup>R</sup>	0.6833	ridge (Ridge-GME)	0.9937
MERG4	1.0816	MERG2 <sub>4</sub> <sup>R</sup>	0.7107	GME $[-10, 10]$	1.5372
MERG1 <sub>4</sub> <sup>T</sup>	0.6833	MERG3 <sub>4</sub> <sup>R</sup>	0.7359	GME $[-5, 5]$	1.1533
MERG2 <sub>4</sub> <sup>T</sup>	0.7107	MERG4 <sub>4</sub> <sup>R</sup>	1.1054	Liu-type	1.3625

In addition, these simulation studies with collinearity and/or outliers were replicated under different conditions, namely for  $N = 20$  and  $30$ ,  $K = 3, 4$  and  $5$ , and different combinations of  $\beta_k$  between  $-3$  and  $3$ . For these additional experiments, the MERG estimators reveal, in general, a good performance. Some convergence problems with the GME estimator are found in some models as well as some variability in the estimates with different supports. A large instability is found for the LTS estimator and different ridge regression estimators.

The MERG estimators appear to be a good choice in models with small samples sizes affected not only by outliers and collinearity simultaneously, but also only by collinearity

or outliers separately. Naturally, this statement should be tempered with caution, since more simulation studies and empirical applications are needed. Moreover, it is important to mention that the ideas from the quantum electrodynamics used in the MERG estimators may not be always valid in different regression models. The violation of Assumption 5.1 motivated an extension of the MERG estimators; see section 5.2.

### 5.1.7 Conclusions

An extension of the MEL estimator in Paris [127, 128] is introduced in this section. Two important features of the MERG estimators emerge in this study: they are easy to compute and no relevant prior information is needed. It seems that this new general class of estimators rivals (and in some cases outperforms) with some traditional competitors in linear regression models with small samples sizes affected by collinearity and/or outliers. However, more simulation studies and empirical applications are needed.

Following a similar procedure, more estimators can be derived simply by merging other entropy measures with another robust estimators. For instance, the use of an S-estimator within the MERG estimators may be an interesting option. Note that it is possible to show that the OLS, LTS, LAD and LMS estimators are particular cases of the S-estimator defined by the minimization of the dispersion of the residuals,  $s(r_1(\boldsymbol{\beta}), r_2(\boldsymbol{\beta}), \dots, r_N(\boldsymbol{\beta}))$ , with final scale estimate  $\hat{\sigma} = s(r_1(\hat{\boldsymbol{\beta}}), r_2(\hat{\boldsymbol{\beta}}), \dots, r_N(\hat{\boldsymbol{\beta}}))$ . It follows that the dispersion is given as the solution of

$$\sum_{n=1}^N \rho\left(\frac{r_n}{s}\right) = N\tau, \quad (5.29)$$

where  $\rho(\cdot)$  is a function to be selected and  $\tau$  is a consistency constant; see, for example, Maronna et al. [109] and Rousseeuw and Yohai [143] for further details.

## 5.2 An extension of the maximum entropy robust regression group estimators

### 5.2.1 Introduction

The MERG estimators, presented in subsection 5.1.3, are based on the MEL estimator in Paris [127, 128], as well as the maximum entropy and robust regression literatures. It seems that

MERG estimators are useful in linear regression models with small samples sizes affected by outliers and/or collinearity. The MEL estimator incorporates only the Shannon entropy and the OLS estimator, while the MERG estimators allow to incorporate the Rényi and Tsallis entropies, and different methods used in robust regression literature. The MERG estimators (where the MEL is a particular case) avoid the possible subjective exogenous information, needed in the GME estimator, to define the support intervals for the parameters and errors in linear regression models. Moreover, the MERG estimators do not require the estimation of regularization parameters that are needed in the majority of the regularization methods.

In this section, the MERG estimators are extended to incorporate supports for the parameters as in the GME estimator. For notational simplicity, this extension is denoted as MERGE estimators. This acronym is the initials of the words *maximum entropy robust regression group extended*, and reflects the ambitious objective to *merge* several estimators in one group with high performance in linear regression models with small samples sizes affected by collinearity and outliers.

Why this extension of MERG estimators? First, it is not yet fully understood whether the theory of light (quantum electrodynamics) in Feynman [57], used in the MERG estimators, is always valid in different regression models. More research is needed to assess the reasonability of the Assumption 5.1. Second, there are regression models where the supports of the parameters can be defined by the researcher's experience and/or provided by the theory (e.g., in economics, estimating the marginal propensity to consume in a Keynesian consumption function). Third, the use of the cross-entropy formalism and the possibility to impose parameter inequality restrictions through the parameter support matrix (as in the GME estimator) are easily handled with this extension. Fourth, the supports for the errors used in the GME estimator are not needed with the MERGE estimators.

### 5.2.2 The MERGE estimators

The MERGE estimators are presented next.

**Definition 5.7.** *The MERGE estimators of  $\beta$  in model (2.42) are given by*

$$\operatorname{argmin}_{\mathbf{p}, \mathbf{r}} \left\{ (1 - \theta) \sum_{k=1}^K \sum_{m=1}^M H(p_{km}) + \theta \sum_{n=1}^N \phi(r_n) \right\}, \quad (5.30)$$

subject to the model constraint and the additivity constraint

$$\left\{ \begin{array}{l} y_n = \sum_{k=1}^K \sum_{m=1}^M x_{nk} z_{km} p_{km} + r_n \\ \sum_{m=1}^M p_{km} = 1, k = 1, 2, \dots, K \end{array} \right., \quad (5.31)$$

where  $p_{km} > 0$ ,  $k = 1, 2, \dots, K$ ,  $m = 1, 2, \dots, M$ , are probabilities,  $z_{km}$  are the supports for the parameters, the function  $\sum_k \sum_m H(p_{km})$  is an entropy measure (e.g., Shannon, Rényi or Tsallis entropies) and  $\phi(r_n)$  is a function of the residuals.

In the objective function (5.30), a parameter  $\theta \in (0, 1)$  is introduced to assign different weights on the components of the objective function. The MERG estimators, presented in Definition 5.2, assume implicitly equal weights in the components of the objective function. Note that the Assumption 5.1 is not used in the MERGE estimators, contrary to MERG, and no supports are needed for the error component, in contrast to the GME estimator. Assuming  $H(p_{km}) = p_{km} \ln p_{km}$ , Table 5.9 presents different MERGE estimators using the OLS, LTS, LAD and LMS estimators.

**Table 5.9:** MERGE estimators.

	MERGE1	MERGE2	MERGE3	MERGE4
$\sum_{n=1}^N \phi(r_n)$	$\sum_{n=1}^N r_n^2$	$\sum_{n=1}^{[(1-\rho)N]+1} r_{(n:N)}^2$	$\sum_{n=1}^N  r_n $	$med_n r_n^2$

When the Tsallis and Rényi entropies are used in the objective function (5.30), the notation is similar to the one used in the MERG estimators, i.e.,  $MERGEi_\alpha^T$  or  $MERGEi_\alpha^R$ , respectively, considering  $i = 1, 2, 3, 4$  and  $\alpha$  the order of the entropy measure.

Following the idea expressed in the conclusions of section 5.1, p. 108, the function  $\phi(r_n)$  in the objective function (5.30) can be generalized using an S-estimator and, in this case, the MERGE estimators can be defined in a more general framework; see Definition 5.8. The S-estimators are regression, scale and affine equivariant and, by a convenient choice of the constants involved, their breakdown point can attain 50%. Note also that, as already mentioned previously, by allowing different types of dispersion measures, the OLS, LTS, LAD and LMS estimators are S-estimators; e.g., Maronna et al. [109], Rousseeuw and Leroy [142] and Rousseeuw and Yohai [143].

**Definition 5.8.** The MERGE estimators of  $\beta$  in model (2.42) are given by

$$\operatorname{argmin}_{\mathbf{p}, s} \left\{ (1 - \theta) \sum_{k=1}^K \sum_{m=1}^M H(p_{km}) + \theta s(r_1(\beta), r_2(\beta), \dots, r_N(\beta)) \right\}, \quad (5.32)$$

subject to

$$\begin{cases} y_n = \sum_{k=1}^K \sum_{m=1}^M x_{nk} z_{km} p_{km} + r_n \\ \sum_{m=1}^M p_{km} = 1, k = 1, 2, \dots, K \\ \sum_{n=1}^N \rho\left(\frac{r_n}{s}\right) = N\tau \end{cases}, \quad (5.33)$$

where  $p_{km} > 0$ ,  $k = 1, 2, \dots, K$ ,  $m = 1, 2, \dots, M$ , are probabilities,  $z_{km}$  are the supports for the parameters, the function  $\sum_k \sum_m H(p_{km})$  is an entropy measure (e.g., Shannon, Rényi or Tsallis entropies),  $r_n$  are the residuals,  $\rho(\cdot)$  is a function to be selected (e.g., the Tukey's biweight  $\rho$ -function) and  $\tau$  is a consistency constant.

The MERGE estimators have no closed-form solution and the solution must be found numerically as in the GME or MERG estimators. The structure of the estimators can be discussed with the Lagrangian functions. The Lagrangian function and the first-order optimality conditions for the MERGE1 estimator are presented next. The same procedure can be followed for the other MERGE estimators. In matricial form, the Lagrangian function is given by

$$L(\mathbf{p}, \mathbf{r}, \boldsymbol{\lambda}, \boldsymbol{\mu}) = (1 - \theta) \mathbf{p}' \ln \mathbf{p} + \theta \mathbf{r}' \mathbf{r} + \boldsymbol{\lambda}' (\mathbf{y} - \mathbf{XZp} - \mathbf{r}) + \boldsymbol{\mu}' (\mathbf{1}_K - (\mathbf{I}_K \otimes \mathbf{1}'_M) \mathbf{p}), \quad (5.34)$$

where  $\boldsymbol{\lambda}$  and  $\boldsymbol{\mu}$  are, respectively, a  $(N \times 1)$  and a  $(K \times 1)$  vectors of Lagrange multipliers on the corresponding constraints. The first-order optimality conditions are

$$\frac{\partial L(\cdot)}{\partial \mathbf{p}} = (1 - \theta)(\ln \mathbf{p} + \mathbf{1}_{KM}) - \mathbf{Z}' \mathbf{X}' \boldsymbol{\lambda} - (\mathbf{I}_K \otimes \mathbf{1}_M) \boldsymbol{\mu} = \mathbf{0}, \quad (5.35)$$

$$\frac{\partial L(\cdot)}{\partial \mathbf{r}} = 2\theta \mathbf{r} - \boldsymbol{\lambda} = \mathbf{0}, \quad (5.36)$$

$$\frac{\partial L(\cdot)}{\partial \boldsymbol{\lambda}} = \mathbf{y} - \mathbf{XZp} - \mathbf{r} = \mathbf{0}, \quad (5.37)$$

$$\frac{\partial L(\cdot)}{\partial \boldsymbol{\mu}} = \mathbf{1}_K - (\mathbf{I}_K \otimes \mathbf{1}'_M) \mathbf{p} = \mathbf{0}. \quad (5.38)$$

For example, solving for  $\mathbf{p}$  yields the solution

$$\hat{\mathbf{p}} = \exp \left( (1 - \theta)^{-1} \mathbf{Z}' \mathbf{X}' \hat{\boldsymbol{\lambda}} \right) \exp \left( (1 - \theta)^{-1} (\mathbf{I}_K \otimes \mathbf{1}_M) \hat{\boldsymbol{\mu}} - \mathbf{1}_{KM} \right). \quad (5.39)$$

In real-world empirical applications and for inference purposes, the bootstrap method is recommended to inferring statistical properties over the MERGE estimators, since these estimators, like the GME and MERG estimators, are suitable for regression models with small samples sizes. The bootstrap estimator for the asymptotic covariance matrix can be computed as

$$\text{Var}(\hat{\boldsymbol{\beta}}_{\text{MERGE}}) = \frac{1}{T} \sum_{t=1}^T (\hat{\boldsymbol{\beta}}_{\text{MERGE}}(t) - \hat{\boldsymbol{\beta}}_{\text{MERGE}}) (\hat{\boldsymbol{\beta}}_{\text{MERGE}}(t) - \hat{\boldsymbol{\beta}}_{\text{MERGE}})', \quad (5.40)$$

where  $\hat{\boldsymbol{\beta}}_{\text{MERGE}}$  is the MERGE estimator and  $\hat{\boldsymbol{\beta}}_{\text{MERGE}}(t)$  is the  $t$ th MERGE estimate of  $\boldsymbol{\beta}$  based on a sample of  $N$  observations drawn with replacement from the original sample.

### 5.2.3 Improvements for MERGE estimators

In this subsection, two important improvements for the MERGE estimators are presented: the use of the cross-entropy formalism and the possibility to impose parameter inequality restrictions in model (2.42) through the parameter support matrix defined in (5.31).

#### 5.2.3.1 Cross-entropy formalism

Considering only the Shannon entropy measure and making use of the cross-entropy formalism presented in Definition 2.8, the objective function (5.30) is rewritten as

$$\underset{\mathbf{p}, \mathbf{r}}{\text{argmin}} \left\{ (1 - \theta) \sum_{k=1}^K \sum_{m=1}^M p_{km} \ln(p_{km}/q_{km}) + \theta \sum_{n=1}^N \phi(r_n) \right\}, \quad (5.41)$$

subject to restrictions (5.31). The vector  $\mathbf{q}$  represents prior knowledge about  $\mathbf{p}$ , in the form of a prior distribution of probabilities. The objective function (5.41) can be redefined as

$$\underset{\mathbf{p}, \mathbf{r}}{\text{argmin}} \left\{ (1 - \theta) \left[ \sum_{k=1}^K \sum_{m=1}^M p_{km} \ln p_{km} - \sum_{k=1}^K \sum_{m=1}^M p_{km} \ln q_{km} \right] + \theta \sum_{n=1}^N \phi(r_n) \right\}. \quad (5.42)$$

When there is prior information on  $\mathbf{p}$  in the form of a prior distribution of probabilities, the cross-entropy formalism can be easily used in the MERGE estimators.

### 5.2.3.2 Parameter inequality restrictions

The approach used to impose parameter inequality restrictions in model (2.42) using the MERGE estimators is similar to the one presented in Campbell and Hill [21] for the GME estimator. An important feature of this approach is the implementation of the restrictions through the parameter support matrix rather than adding more constraints in (5.31). Naturally, the parameter support matrix is no longer block diagonal as in the GME estimator. However, this approach is simpler, from the computational point of view, than the one imposing more constraints in (5.31).

To illustrate this approach, consider the model (2.42) without constant and three explanatory variables. Assuming  $M = 3$  points in the supports, the vector  $\beta$  defined in the model constraint (5.31) is given by

$$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} & z_{13} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & z_{21} & z_{22} & z_{23} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & z_{31} & z_{32} & z_{33} \end{bmatrix} \begin{bmatrix} p_{11} \\ p_{12} \\ p_{13} \\ p_{21} \\ \vdots \\ p_{33} \end{bmatrix}. \quad (5.43)$$

Suppose that the inequality restriction  $\beta_2 \geq \beta_3$  is further imposed. Thus, the vector  $\beta$  can be defined by

$$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} & z_{13} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & z_{21} & z_{22} & z_{23} & z_{31} & z_{32} & z_{33} \\ 0 & 0 & 0 & 0 & 0 & 0 & z_{31} & z_{32} & z_{33} \end{bmatrix} \begin{bmatrix} p_{11} \\ p_{12} \\ p_{13} \\ p_{21} \\ \vdots \\ p_{33} \end{bmatrix}, \quad (5.44)$$

where  $z_{21}$ ,  $z_{22}$  and  $z_{23}$  are non-negative. Given this parameter support matrix (that is not block diagonal as in the GME estimator), it follows that

$$\begin{aligned} \hat{\beta}_2 &= z_{21}\hat{p}_{21} + z_{22}\hat{p}_{22} + z_{23}\hat{p}_{23} + z_{31}\hat{p}_{31} + z_{32}\hat{p}_{32} + z_{33}\hat{p}_{33} \\ &= z_{21}\hat{p}_{21} + z_{22}\hat{p}_{22} + z_{23}\hat{p}_{23} + \hat{\beta}_3, \end{aligned} \quad (5.45)$$

meaning that  $\beta_2 \geq \beta_3$ . Suppose now that the inequality restriction  $\beta_1 + \beta_2 \leq \beta_3$  is imposed. The vector  $\boldsymbol{\beta}$  is now defined as

$$\begin{bmatrix} \beta_1 \\ \beta_2 \\ \beta_3 \end{bmatrix} = \begin{bmatrix} z_{11} & z_{12} & z_{13} & 0 & 0 & 0 & 0 & 0 & 0 \\ -z_{11} & -z_{12} & -z_{13} & -z_{21} & -z_{22} & -z_{23} & z_{31} & z_{32} & z_{33} \\ 0 & 0 & 0 & 0 & 0 & 0 & z_{31} & z_{32} & z_{33} \end{bmatrix} \begin{bmatrix} p_{11} \\ p_{12} \\ p_{13} \\ p_{21} \\ \vdots \\ p_{33} \end{bmatrix}, \quad (5.46)$$

where  $z_{21}$ ,  $z_{22}$  and  $z_{23}$  are non-negative. With this parameter support matrix, it follows that

$$\hat{\beta}_2 = -\hat{\beta}_1 - z_{21}\hat{p}_{21} - z_{22}\hat{p}_{22} - z_{23}\hat{p}_{23} + \hat{\beta}_3, \quad (5.47)$$

and, as requested,

$$\hat{\beta}_1 + \hat{\beta}_2 = \hat{\beta}_3 - (z_{21}\hat{p}_{21} + z_{22}\hat{p}_{22} + z_{23}\hat{p}_{23}) \leq \hat{\beta}_3. \quad (5.48)$$

The approach proposed by Campbell and Hill [21] for the GME estimator is easily extended to MERGE estimators. See Campbell and Hill [21] for further details on this approach with the GME estimator.

#### 5.2.4 Simulation study

The following simulation study illustrates the performance of the MERGE estimators in the estimation of linear regression models with small samples sizes affected by outliers and collinearity. The main purpose is to illustrate that the MERGE estimators may outperform the MERG estimators rather than to provide a full comparison between these estimation techniques and other methods. However, results are also presented for the RR-MM estimator, the main competitor of MERG and MERGE estimators (MERG(E) in short) in regression models with collinearity and outliers, and the OLS estimator.

As in the experiment in subsection 5.1.5, a pseudo-random number generator is used as well as the singular value decomposition to define matrices with a desired condition number. Different  $(N \times K)$  matrices  $\mathbf{X}$  are generated from a normal distribution with zero mean and unit standard deviation. Changing the singular values obtained from the decomposition, different matrices  $\mathbf{X}_1$  with  $\text{cond}_2 \mathbf{X}_1 = 150$  are defined. To create a proportion  $\delta$  of outliers,

a similar strategy in Ferretti et al. [56] and Golan and Perloff [68] is followed, i.e., different models (without intercept) given by  $\mathbf{y} = \mathbf{X}_1\boldsymbol{\beta} + \mathbf{u}$  are defined, where  $\mathbf{y} = \mathbf{u}$  is randomly generated from a normal distribution with zero mean and unit standard deviation. For  $N\delta$  observations, the elements in  $\mathbf{y}$  are replaced by the value 6 and the corresponding elements in the first column of  $\mathbf{X}_1$  by the value 10 (being this the  $\mathbf{X}_2$  matrix). This simulation study considers the following possibilities:  $N = 10$  and  $N = 30$ ;  $K = 3$  and  $K = 5$ ;  $\delta = 0.1$  and  $\delta = 0.3$ . The MERGE estimators are performed using Definition 5.7, with  $\theta = 0.5$ , and two different supports for the parameters, namely  $[-5, 5]$  and  $[-1, 1]$ , both with  $M = 5$ . For the 1000 trials performed, the MSEL is the measure used to evaluate the performance of different estimators. Tables 5.10 and 5.11 present the results.

**Table 5.10:** MSEL for the estimators in the simulation study ( $N = 10$ ).

		$K = 3$		$K = 5$	
		$\text{cond}_2 \mathbf{X}_2$		$\text{cond}_2 \mathbf{X}_2$	
		$(\approx) 203$	$(\approx) 95$	$(\approx) 108$	$(\approx) 96$
		$\delta = 0.1$	$\delta = 0.3$	$\delta = 0.1$	$\delta = 0.3$
OLS		559.1269	45.2705	149.8799	22.5019
RR-MM		154.0145	8.3637	109.0394	14.1674
MERG	1st best	0.8673	0.7271	1.0039	0.7147
	2nd best	0.9130	0.7455	1.0265	0.8303
	3rd best	0.9285	0.8669	1.0463	0.8722
	worst	2.8191	1.1978	1.7521	2.7399
MERGE $[-5, 5]$	1st best	0.1179	0.1232	0.2117	0.2286
	2nd best	0.1250	0.1412	0.2230	0.3057
	3rd best	1.3939	1.4748	2.9431	3.2880
	worst	11.4988	10.7308	16.5803	10.1718
MERGE $[-1, 1]$	1st best	0.1034	0.1472	0.0884	0.2615
	2nd best	0.1417	0.2923	0.2531	0.3354
	3rd best	0.2426	0.4533	0.5441	0.4734
	worst	1.8883	1.8337	3.0394	2.8983

This simulation study reveals two main results. First, and concerning the main goal of this experiment, the MERGE estimators may outperform the MERG estimators, although this

**Table 5.11:** MSEL for the estimators in the simulation study ( $N = 30$ ).

		$K = 3$		$K = 5$	
		cond <sub>2</sub> $\mathbf{X}_2$		cond <sub>2</sub> $\mathbf{X}_2$	
		( $\approx$ ) 68	( $\approx$ ) 146	( $\approx$ ) 255	( $\approx$ ) 101
		$\delta = 0.1$	$\delta = 0.3$	$\delta = 0.1$	$\delta = 0.3$
OLS		24.6791	29.1554	273.1555	14.1823
RR-MM		1.8419	0.4840	0.6286	0.4145
MERG	1st best	0.7618	0.6332	0.8927	0.8257
	2nd best	0.9432	0.6973	1.0067	0.8649
	3rd best	0.9669	0.8127	1.0410	1.0292
	worst	1.4752	1.8351	2.1807	1.6949
MERGE $[-5, 5]$	1st best	0.1363	0.1316	0.2176	0.2101
	2nd best	0.1522	0.1614	0.2295	0.2541
	3rd best	1.2663	1.4268	2.4975	2.6356
	worst	12.7448	6.9831	19.7499	6.3764
MERGE $[-1, 1]$	1st best	0.1012	0.1533	0.2691	0.2537
	2nd best	0.2090	0.3199	0.2936	0.3271
	3rd best	0.3919	0.3785	0.6495	0.4090
	worst	1.9006	1.5951	2.8381	3.0892

performance seems to depend on the amplitude of the supports. Thus, as a precaution, the MERGE estimators should be used only in cases of fully correct prior information about the parameters (supports of smaller amplitude).<sup>21</sup> The worst results for the MERGE estimators, particularly the ones with the supports  $[-5, 5]$ , are almost exclusively for those estimators using the OLS and LMS estimators, i.e., the  $\text{MERGE}_i$ ,  $\text{MERGE}_4^R$  and  $\text{MERGE}_4^T$ , for  $i = 1, 4$ . In contrast, the lower values of MSEL are almost exclusively achieved by MERGE estimators using the LTS and LAD estimators.

Second, the comparison between the MERGE estimators and the RR-MM estimator<sup>22</sup> depends on the sample size  $N$ . For very small samples, such as  $N = 10$ , almost all the MERGE estimators outperform the RR-MM estimator. However, for  $N = 30$ , it is only

<sup>21</sup>The exogenous parameter weighting between signal and noise is  $\theta = 0.5$  in this study. Naturally, this parameter can be changed in order to reflect different weights in the components of the objective function.

<sup>22</sup>The MSEL values for the RR-MM estimator are calculated using a 10% upper trimmed average.

possible to say that, in general, the MERGE estimators rival with the RR-MM estimator. This result is not fully unexpected, taking into account that the ME theory is usually used to extract information from limited data.

### 5.2.5 Conclusions

In this section, an extension of the MERG estimators is presented. This extension allows the introduction of supports for the parameters as in the traditional GME and GME- $\alpha$  estimators. The cross-entropy formalism and the possibility to impose parameter inequality restrictions through the parameter support matrix are easily incorporated in the MERGE estimators.

The simulation study reveals that the MERGE estimators may outperform the MERG estimators in linear regression models with small samples sizes affected by collinearity and outliers, yet this performance depends on the definition of the supports. Moreover, the MERGE estimators rival (and may outperform in very small samples) with the RR-MM estimator, probably the most powerful estimator in the literature concerning the estimation of regression models affected by collinearity and outliers. Undoubtedly, more future research is needed on the MERGE estimators.



## Chapter 6

# Concluding remarks

“Is it possible to make the theory easier to apply for the practitioners?”

Golan [66, p. 386] concerning next directions on IEE.

Some final conclusions and topics for future research are presented in this final chapter.

### 6.1 Final conclusions

In this thesis, two new applications of ME estimation are proposed: the estimation of technical efficiency with state-contingent production frontiers and the estimation of the ridge parameter in the ridge regression. Two developments with ME estimation, namely the MERG and the MERGE estimators, are also discussed.

Concerning the analysis of linear regression models with small samples sizes, possibly affected by collinearity and/or outliers, this study provides significant contributions to

- the production economics literature, in general, and the efficiency literature, in particular, by estimating technical efficiency with state-contingent production frontiers using ME estimators;
- the ridge regression literature by the definition and evaluation of a new ridge parameter estimator based on the ridge trace and the GME estimator;
- the ME literature by the introduction and discussion of two new estimation procedures.

In the estimation of technical efficiency with state-contingent production frontiers, this study is the first one that combines the GME, GME- $\alpha$  and GCE estimators. The discussion, presented in Chapter 3, reveals that these ME estimators outperform the ML estimator in most of the cases analyzed: models with few observations in some states of nature and models with a large number of states of nature, that usually represent models affected by (severe) collinearity. Some advantages of these estimators are the following: (1) the possibility of considering easily prior information on the parameters and errors components; (2) the traditional assumptions on the errors' distributions are not necessary; and (3) these estimators can be used in models with a large number of states of nature and even with only one observation *per* state (which represent models usually affected by severe collinearity). The ME estimators are expected to provide a strong contribution to the increase of empirical work with state-contingent production frontiers, one of the most complete approaches to investigate economic production under uncertainty.

A new method to select the ridge parameter in the ridge regression model, called the Ridge-GME estimator, is presented in Chapter 4. This new estimator is based on the GME estimator and the ridge trace. The empirical application and the simulation study illustrate the good performance of the Ridge-GME estimator of the ridge parameter. The simulation study reveals that the Ridge-GME estimator is probably one of the best ridge parameter estimators in the literature on the ridge regression, in the case of regression models with small samples sizes affected by collinearity. The good performance of the Ridge-GME estimator is also identified in a recent paper by Erdugan and Akdeniz [46], where the Ridge-GME estimator is analyzed and adapted for a jackknife procedure in the ridge regression.

The main weakness of the GME class of estimators is the possible instability of the solution given different support intervals. Because of this problem, Paris [127] developed the MEL estimator based on some ideas from the theory of light (quantum electrodynamics) in Feynman [57], the Shannon entropy measure and the OLS estimator. In Chapter 5, section 5.1, considering the same framework as in the MEL estimator, the MERG estimators are defined through a general expression in which the Rényi and Tsallis entropies can be applied, as well as different robust regression estimators. The MERG estimators (which include the MEL estimator as a particular case) have two important features: they are easy to compute and, the most important, no relevant prior information is needed to implement them. Several

simulation studies illustrate an excellent performance of the MERG estimators. Besides many other estimators, these experiments include the RR-MM estimator, which is one of the most powerful estimators in the literature concerning the estimation of regression models affected by collinearity and outliers.

Finally, the MERGE estimators are introduced in section 5.2. These estimators are an extension of the MERG estimators, in the sense that they allow the introduction of supports for the parameters, as in the traditional GME and GME- $\alpha$  estimators. Those supports can be useful when there is prior information about the parameters to be estimated. The analogy with the theory of light no longer holds with the MERGE estimators, i.e., Assumption 5.1 is not used. The simulation study reveals that the MERGE estimators may outperform the MERG estimators and the RR-MM estimator in linear regression models with small samples sizes affected by collinearity and outliers.

Based on the experiments conducted in Chapter 5 (and others not reported here), the MERG2 and MERG3 class of estimators are probably the most adequate choices in the case of no prior information on the parameters. Similarly, the MERGE2 and MERGE3 class of estimators are likely the most proper choices if there is a correct prior information on the parameters. However, some questions on the MERG(E) estimators remain open, such as: which entropy measure should be used? and what should be the order of the entropy measure? It seems that, in some cases, the MERG(E) estimators with the Tsallis and Rényi entropies provide a lower MSEL than the estimators with the Shannon entropy. Further research on this issue is necessary in order to identify the most proper estimator in each case. Moreover,  $\alpha = 4$  is always used in this study, but other values of  $\alpha$  should be tested in real-world problems.

In summary, *simplicity* and *freedom* are probably two words that best characterize the MERG(E) estimators. Indeed, these estimators are easy to compute and no relevant prior information is needed to implement them. For comparison, see the complexity of the RR-MM estimator in Maronna [108], probably one of the most powerful estimators in the literature concerning the estimation of regression models affected by collinearity and outliers, and the main rival of the MERG(E) estimators.

## 6.2 Future work

Guidelines for future research are presented next, some of which are already in progress.

### 1. ME estimators.

- (a) The ME estimators are widely used in econometrics, but they still have a relative small contribution in general statistics. More work on ME estimators, including comparisons with usual competitors, needs to be done in some traditional areas of statistics to illustrate the potential of the ME estimation;
- (b) The GME- $\alpha$  estimators deserve much more investigation; see the final comments in Golan and Perloff [68, p. 209]. In addition to further analytical research and simulation studies, there is an issue that needs to be explored: the use of other entropy measures. Taneja [166, pp. 333–336] provides a list of 25 different entropy measures. For example, the Aczél-Daróczy entropy measure,

$$H_{\alpha}^{AD}(p_1, p_2, \dots, p_K) = -\frac{\sum_{k=1}^K (p_k)^{\alpha} \ln p_k}{\sum_{k=1}^K (p_k)^{\alpha}}, \quad (6.1)$$

with  $\alpha > 0$ , is an interesting option to be explored in the context of the GME- $\alpha$  estimators; see Golan [64] concerning the connections between different generalized entropy measures.

### 2. Technical efficiency with state-contingent production frontiers.

- (a) Including different orders of entropy in the two-error component rather than using the same value for both. Thus, the objective function (3.10) is defined by

$$H(\mathbf{p}, \mathbf{w}, \boldsymbol{\rho}) = \sum_{r=1}^R H_{\alpha_1}^e(\beta_r) + \sum_{n=1}^N H_{\alpha_2}^e(v_n) + \sum_{n=1}^N H_{\alpha_3}^e(u_n). \quad (6.2)$$

Since the two-error component has different characteristics (e.g., skewness), the use of different orders of entropy may improve the performance of the GME- $\alpha$  estimators. Also, the use of different orders of entropy may allow to identify the optimal order ( $\alpha_3$ ) of the entropy measure for the inefficiency component;

- (b) Developing a more complete simulation study using a model based on the production technology (3.6); testing other distributions for the error inefficiency component (e.g., truncated normal and gamma distributions) with different parameters; and using different numbers of state-allocable inputs and exogenous variables;
- (c) Developing new models of state-contingent production based on flexible functional forms, such as the generalized quadratic or translog, for the production technology.

### 3. Ridge-GME estimator.

- (a) It would be interesting to conduct a larger simulation study that includes all (or almost all) the estimators in the literature over the past forty years;
- (b) Since different users looking at the same ridge trace can provide different supports for the GME estimator, the sensitivity of the Ridge-GME estimator should be investigated, based on different prior information provided by the ridge trace;
- (c) Developing a computational procedure to choose the different supports from the ridge trace, leaving to the user only the need of choosing the ridge interval.

### 4. MERG(E) estimators.

- (a) Other MERG(E) estimators can be defined by merging other entropy measures and/or other robust estimators. This is an interesting topic for future research;
- (b) Which entropy measure should be used in the MERG(E) estimators? This is one of the most important issues that deserve further research in the future;
- (c) How to choose the best  $\alpha$  when the Tsallis or Rényi entropies are used in the MERG(E) estimators? This is the same problem faced by Golan and Perloff [68] with the GME- $\alpha$  estimators. The choice of  $\alpha$  is likely to be more dependent on the presence (and type) of influential observations than the magnitude of collinearity;
- (d) How the performance of MERG(E) estimators is affected by the type of influential observations? Detailed studies with each type of outlier are necessary;
- (e) A complete analysis on the properties of the MERG(E) estimators is necessary, as well as an axiomatic derivation using other information-based theoretical methods;
- (f) Developing the MERGE estimators in Definition 5.8 using an S-estimator;

- (g) More simulation studies and real-world empirical applications with a larger set of possible competitor estimators must be conducted in order to investigate the potential of the MERG(E) estimators.

These are some of the topics that deserve further investigation. The ME estimation will likely play an important role in econometrics, as well as in new developments in other fields of statistics.

# References

- [1] S. Abe. Axioms and uniqueness theorem for Tsallis entropy. *Physics Letters A*, 271(1-2):74–79, 2000.
- [2] S. N. Afriat. Efficiency estimation of production functions. *International Economic Review*, 13(3):568–598, 1972.
- [3] D. Aigner, C. A. K. Lovell, and P. Schmidt. Formulation and estimation of stochastic frontier production function models. *Journal of Econometrics*, 6(1):21–37, 1977.
- [4] M. Alkhamisi, G. Khalaf, and G. Shukur. Some modifications for choosing ridge parameters. *Communications in Statistics - Theory and Methods*, 35(11):2005–2020, 2006.
- [5] M. A. Alkhamisi and G. Shukur. A Monte Carlo study of recent ridge parameters. *Communications in Statistics - Simulation and Computation*, 36(3):535–547, 2007.
- [6] C. Amado and A. M. Pires. Robust bootstrap with non-random weights based on the influence function. *Communications in Statistics - Simulation and Computation*, 33(2):377–396, 2004.
- [7] K. J. Arrow and G. Debreu. Existence of an equilibrium for a competitive economy. *Econometrica*, 22(3):265–290, 1954.
- [8] R. Ash. *Information Theory*. Interscience Publishers, New York, 1965.
- [9] G. E. Battese and G. S. Corra. Estimation of a production frontier model: with application to the pastoral zone of Eastern Australia. *Australian Journal of Agricultural and Resource Economics*, 21(3):169–179, 1977.

- [10] D. A. Belsley, E. Kuh, and R. E. Welsch. *Regression Diagnostics - Identifying Influential Data and Sources of Collinearity*. John Wiley & Sons, Hoboken, New Jersey, 2004.
- [11] A. Ben-Naim. *Entropy Demystified - The Second Law Reduced to Plain Common Sense with Seven Simulated Games*. World Scientific Publishing, Singapore, 2008.
- [12] A. Ben-Naim. *Discover Entropy and the Second Law of Thermodynamics - A Playful Way of Discovering a Law of Nature*. World Scientific Publishing, Singapore, 2010.
- [13] A. K. Bera and Y. Biliyas. The MM, ME, ML, EL, EF and GMM approaches to estimation: a synthesis. *Journal of Econometrics*, 107(1-2):51–86, 2002.
- [14] J. N. Boles. Efficiency squared - efficient computation of efficiency indexes. In *Proceedings of the Thirty Ninth Annual Meeting of the Western Farm Economics Association*, pages 137–142, 1966.
- [15] R. G. Bressler. The measurement of productive efficiency. In *Proceedings of the Thirty Ninth Annual Meeting of the Western Farm Economics Association*, pages 129–136, 1966.
- [16] W. Briec. A graph-type extension of Farrell technical efficiency measure. *Journal of Productivity Analysis*, 8(1):95–110, 1997.
- [17] L. Brillouin. *Science and Information Theory*. Academic Press Inc. Publishers, New York, 1956.
- [18] J. Brinkhuis and V. Tikhomirov. *Optimization: Insights and Applications*. Princeton University Press, Princeton, New Jersey, 2005.
- [19] P. J. Brown. *Measurement, Regression, and Calibration*. Clarendon Press, Oxford, 1994.
- [20] R. Campbell, K. Rogers, and J. Rezek. Efficient frontier estimation: a maximum entropy approach. *Journal of Productivity Analysis*, 30(3):213–221, 2008.
- [21] R. C. Campbell and R. C. Hill. Imposing parameter inequality restrictions using the principle of maximum entropy. *Journal of Statistical Computation and Simulation*, 76(11):985–1000, 2006.

- [22] M. R. Caputo and Q. Paris. Comparative statics of the generalized maximum entropy estimator of the general linear model. *European Journal of Operational Research*, 185(1):195–203, 2008.
- [23] R. G. Chambers and J. Quiggin. *Uncertainty, Production, Choice, and Agency - The State-Contingent Approach*. Cambridge University Press, Cambridge, 2000.
- [24] R. G. Chambers and J. Quiggin. The state-contingent properties of stochastic production functions. *American Journal of Agricultural Economics*, 84(2):513–526, 2002.
- [25] R. G. Chambers and J. Quiggin. Dual approaches to the analysis of risk aversion. *Economica*, 74(294):189–213, 2007.
- [26] R. G. Chambers, Y. Chung, and R. Färe. Benefit and distance functions. *Journal of Economic Theory*, 70(2):407–419, 1996.
- [27] R. G. Chambers, Y. Chung, and R. Färe. Profit, directional distance functions, and Nerlovian efficiency. *Journal of Optimization Theory and Applications*, 98(2):351–364, 1998.
- [28] A. Charnes, W. W. Cooper, and E. Rhodes. Measuring the efficiency of decision making units. *European Journal of Operational Research*, 2(6):429–444, 1978.
- [29] J. P. Chavas. A cost approach to economic analysis under state-contingent production uncertainty. *American Journal of Agricultural Economics*, 90(2):435–446, 2008.
- [30] R. Clausius. *The Mechanical Theory of Heat, with its Applications to the Steam-Engine and to the Physical Properties of Bodies*. John Van Voorst, London, 1867.
- [31] T. Coelli, D. S. Prasada Rao, and G. E. Battese. *An Introduction to Efficiency and Productivity Analysis*. Kluwer Academic Publishers, Boston, 1999.
- [32] W. W. Cooper, L. M. Seiford, and K. Tone. *Data Envelopment Analysis - A Comprehensive Text with Models, Applications, References and DEA-Solver Software*. Springer, New York, 2nd edition, 2007.
- [33] I. Csiszár. Why least squares and maximum entropy? An axiomatic approach to inference for linear inverse problems. *The Annals of Statistics*, 19(4):2032–2066, 1991.

- [34] E. M. F. Curado and C. Tsallis. Generalized statistical mechanics: connection with thermodynamics. *Journal of Physics A*, 24(2):L69–L72, 1991.
- [35] A. C. Davison and D. V. Hinkley. *Bootstrap methods and their application*. Cambridge University Press, Cambridge, 1997.
- [36] G. Debreu. The coefficient of resource utilization. *Econometrica*, 19(3):273–292, 1951.
- [37] R. P. Di Sisto, S. Martínez, R. B. Orellana, A. R. Plastino, and A. Plastino. General thermostistical formalisms, invariance under uniform spectrum translations, and Tsallis q-additivity. *Physica A*, 265(3):590–613, 1999.
- [38] A. Dionísio, R. Menezes, and D. A. Mendes. Entropy-based independence test. *Non-linear Dynamics*, 44(1-4):351–357, 2006.
- [39] A. Dionísio, R. Menezes, and D. A. Mendes. An econophysics approach to analyse uncertainty in financial markets: an application to the Portuguese stock market. *The European Physical Journal B*, 50(1-2):161–164, 2006.
- [40] A. Dionísio, A. H. Reis, and L. Coelho. Utility function estimation: the entropy approach. *Physica A*, 387(15):3862–3867, 2008.
- [41] D. L. Donoho, I. M. Johnstone, J. C. Hoch, and A. S. Stern. Maximum entropy and the nearly black object. *Journal of the Royal Statistical Society, Series B*, 54(1):41–81, 1992.
- [42] J. S. Dugdale. *Entropy and its Physical Meaning*. Taylor & Francis Group, London, 1996.
- [43] B. Efron. Bootstrap methods: another look an the jackknife. *The Annals of Statistics*, 7(1):1–26, 1979.
- [44] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of Statistics*, 32(2):407–499, 2004.
- [45] S. P. Ellis and S. Morgenthaler. Leverage and breakdown in  $L_1$  regression. *Journal of the American Statistical Association*, 87(417):143–148, 1992.

- [46] F. Erdugan and F. Akdeniz. Computational method for jackknifed generalized ridge tuning parameter based on generalized maximum entropy. *Communications in Statistics - Simulation and Computation*, 41(8):1411–1429, 2012.
- [47] R. Färe and S. Grosskopf. Theory and application of directional distance functions. *Journal of Productivity Analysis*, 13(2):93–103, 2000.
- [48] R. Färe and S. Grosskopf. *New Directions: Efficiency and Productivity*. Kluwer Academic Publishers, Boston, 2004.
- [49] R. Färe and C. A. K. Lovell. Measuring the technical efficiency of production. *Journal of Economic Theory*, 19(1):150–162, 1978.
- [50] R. Färe and D. Primont. *Multi-Output Production and Duality: Theory and Applications*. Kluwer Academic Publishers, Boston, 1995.
- [51] R. Färe, S. Grosskopf, and C. A. K. Lovell. *The Measurement of Efficiency of Production*. Kluwer Academic Publishers, Boston, 1985.
- [52] R. Färe, S. Grosskopf, and C. A. K. Lovell. *Production Frontiers*. Cambridge University Press, Cambridge, 1994.
- [53] R. Färe, S. Grosskopf, and W. L. Weber. The effect of risk-based capital requirements on profit efficiency in banking. *Applied Economics*, 36(15):1731–1743, 2004.
- [54] M. J. Farrell. The measurement of productive efficiency. *Journal of the Royal Statistical Society, Series A*, 120(3):253–290, 1957.
- [55] P. Ferreira, A. Dionísio, and C. Pires. Adopt the euro? The GME approach. *Journal of Economic Interaction and Coordination*, 5(2):231–247, 2010.
- [56] N. Ferretti, D. Kelmansky, V. J. Yohai, and R. H. Zamar. A class of locally and globally robust regression estimates. *Journal of the American Statistical Association*, 94(445):174–188, 1999.
- [57] R. P. Feynman. *QED - The Strange Theory of Light and Matter*. Penguin Group, London, 1985.

- [58] I. Fraser. An application of maximum entropy estimation: the demand for meat in the United Kingdom. *Applied Economics*, 32(1):45–59, 2000.
- [59] L. Galleani and R. Garello. The minimum entropy mapping spectrum of a DNA sequence. *IEEE Transactions on Information Theory*, 56(2):771–783, 2010.
- [60] F. Gamboa and E. Gassiat. Bayesian methods and maximum entropy for ill-posed inverse problems. *The Annals of Statistics*, 25(1):328–350, 1997.
- [61] D. G. Gibbons. A simulation study of some ridge estimators. *Journal of the American Statistical Association*, 76(373):131–139, 1981.
- [62] J. W. Gibbs. *Elementary principles in statistical mechanics developed with especial reference to the rational foundation of thermodynamics*. Charles Scribner’s Sons, New York, 1902.
- [63] A. Golan. A simultaneous estimation and variable selection rule. *Journal of Econometrics*, 101(1):165–193, 2001.
- [64] A. Golan. Information and entropy econometrics - editor’s view. *Journal of Econometrics*, 107(1-2):1–15, 2002.
- [65] A. Golan. Information and entropy econometrics - a review and synthesis. *Foundations and Trends<sup>®</sup> in Econometrics*, 2(1-2):1–145, 2006.
- [66] A. Golan. Information and entropy econometrics - volume overview and synthesis. *Journal of Econometrics*, 138(2):379–387, 2007.
- [67] A. Golan and V. Dose. A generalized information theoretical approach to tomographic reconstruction. *Journal of Physics A*, 34(7):1271–1283, 2001.
- [68] A. Golan and J. M. Perloff. Comparison of maximum entropy and higher-order entropy estimators. *Journal of Econometrics*, 107(1-2):195–211, 2002.
- [69] A. Golan, G. Judge, and D. Miller. *Maximum Entropy Econometrics: Robust Estimation with Limited Data*. John Wiley & Sons, Chichester, 1996.
- [70] G. H. Golub, M. Heath, and G. Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

- [71] W. H. Greene. *Econometric Analysis*. Pearson Prentice Hall, Upper Saddle River, New Jersey, 6th edition, 2008.
- [72] M. Grendár Jr. and M. Grendár. Maximum entropy: clearing up mysteries. *Entropy*, 3(2):58–63, 2001.
- [73] A. Hald. *Statistical Theory with Engineering Applications*. John Wiley & Sons, New York, 1952.
- [74] R. V. L. Hartley. Transmission of information. *The Bell System Technical Journal*, 7(3):535–563, 1928.
- [75] T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning - Data Mining, Inference, and Prediction*. Springer, New York, 2nd edition, 2009.
- [76] A. E. Hoerl and R. W. Kennard. Ridge regression: biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970.
- [77] A. E. Hoerl and R. W. Kennard. Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1):69–82, 1970.
- [78] A. E. Hoerl, R. W. Kennard, and K. F. Baldwin. Ridge regression: some simulations. *Communications in Statistics - Simulation and Computation*, 4(2):105–123, 1975.
- [79] D. Holste, I. Grosse, and H. Herzl. Bayes' estimators of generalized entropies. *Journal of Physics A*, 31(11):2551–2566, 1998.
- [80] P. J. Huber and E. M. Ronchetti. *Robust Statistics*. John Wiley & Sons, Hoboken, New Jersey, 2nd edition, 2009.
- [81] E. T. Jaynes. Information theory and statistical mechanics. *Physical Review*, 106(4):620–630, 1957.
- [82] E. T. Jaynes. Information theory and statistical mechanics. II. *Physical Review*, 108(2):171–190, 1957.
- [83] E. T. Jaynes. Information theory and statistical mechanics. In K. W. Ford, editor, *Statistical Physics*, Brandeis University Summer Institute Lectures in Theoretical Physics, pages 182–218. W. A. Benjamin, Inc., New York, 1963.

- [84] E. T. Jaynes. Where do we go from here? In C. Ray Smith and W. T. Grandy, Jr., editors, *Maximum-Entropy and Bayesian Methods in Inverse Problems*, pages 21–58. D. Reidel Publishing Company, 1985.
- [85] E. T. Jaynes. The relation of Bayesian and maximum entropy methods. In G. J. Erickson and C. R. Smith, editors, *Maximum-Entropy and Bayesian Methods in Science and Engineering (Vol. I)*, pages 25–29. Kluwer Academic Publishers, 1988.
- [86] E. T. Jaynes. *Probability Theory - The Logic of Science*. Cambridge University Press, Cambridge, 2003. Note: this book was available for free on the internet during many years. It is no longer available due to copyright reasons.
- [87] G. G. Judge and R. C. Mittelhammer. *An Information Theoretic Approach to Econometrics*. Cambridge University Press, Cambridge, 2012.
- [88] S. Kaçiranlar, S. Sakalhoğlu, F. Akdeniz, G. P. H. Styan, and H. J. Werner. A new biased estimator in linear regression and a detailed analysis of the widely-analysed dataset on Portland cement. *The Indian Journal of Statistics, Series B*, 61(3):443–459, 1999.
- [89] K. P. Kalirajan and R. T. Shand. Frontier production functions and technical efficiency measures. *Journal of Economic Surveys*, 13(2):149–172, 1999.
- [90] G. Khalaf and G. Shukur. Choosing ridge parameter for regression problems. *Communications in Statistics - Theory and Methods*, 34(5):1177–1182, 2005.
- [91] A. I. Khinchin. *Mathematical Foundations of Information Theory*. Dover Publications, New York, 1957.
- [92] B. M. G. Kibria. Performance of some new ridge regression estimators. *Communications in Statistics - Simulation and Computation*, 32(2):419–435, 2003.
- [93] S. Kullback. Certain inequalities in information theory and the Cramer-Rao inequality. *Annals of Mathematical Statistics*, 25(4):745–751, 1954.
- [94] S. Kullback. *Information Theory and Statistics*. John Wiley & Sons, New York, 1959.

- [95] S. Kullback and R. A. Leibler. On information and sufficiency. *Annals of Mathematical Statistics*, 22(1):79–86, 1951.
- [96] S. C. Kumbhakar and C. A. K. Lovell. *Stochastic Frontier Analysis*. Cambridge University Press, Cambridge, 2000.
- [97] A. O. Lansink, E. Silva, and S. Stefanou. Inter-firm and intra-firm efficiency measures. *Journal of Productivity Analysis*, 15(3):185–199, 2001.
- [98] S. H. Lence and D. J. Miller. Estimation of multi-output production functions with incomplete data: a generalised maximum entropy approach. *European Review of Agricultural Economics*, 25(2):188–209, 1998.
- [99] R. D. Levine and M. Tribus. *The Maximum Entropy Formalism*. MIT Press, Cambridge, Massachusetts, 1979.
- [100] D. V. Lindley. On a measure of the information provided by an experiment. *Annals of Mathematical Statistics*, 27(4):986–1005, 1956.
- [101] K. Liu. Using Liu-type estimator to combat collinearity. *Communications in Statistics - Theory and Methods*, 32(5):1009–1020, 2003.
- [102] E. Maasoumi. A compendium to information theory in economics and econometrics. *Econometric Reviews*, 12(2):137–181, 1993.
- [103] P. Macedo, M. Scotto, and E. Silva. A general class of estimators for the linear regression model affected by collinearity and outliers. *Communications in Statistics - Simulation and Computation*, 39(5):981–993, 2010.
- [104] P. Macedo, M. Scotto, and E. Silva. On the choice of the ridge parameter: a maximum entropy approach. *Communications in Statistics - Simulation and Computation*, 39(8):1628–1638, 2010.
- [105] P. Macedo, E. Silva, and M. Scotto. Maximum entropy estimators to assess technical efficiency with state-contingent production frontiers. In *Proceedings of the 58th World Statistics Congress of the International Statistical Institute*, Dublin, Ireland, 2011.

- [106] P. Macedo, E. Silva, and M. Scotto. Technical efficiency with state-contingent production frontiers using maximum entropy estimators. *Journal of Productivity Analysis*, (accepted for publication), 2012.
- [107] P. G. L. Mana. Consistency of the Shannon entropy in quantum experiments. *Physical Review A*, 69(6):1–12, 2004.
- [108] R. A. Maronna. Robust ridge regression for high-dimensional data. *Technometrics*, 53(1):44–53, 2011.
- [109] R. A. Maronna, R. D. Martin, and V. J. Yohai. *Robust Statistics - Theory and Methods*. John Wiley & Sons, Chichester, 2006.
- [110] J. Maxwell. Illustrations of the dynamical theory of gases. On the motions and collisions of perfectly elastic spheres. *Philosophical Magazine*, 19:19–32, 1860.
- [111] J. Maxwell. Illustrations of the dynamical theory of gases. On the process of diffusion of two or more kinds of moving particles among one another. *Philosophical Magazine*, 20:21–37, 1860.
- [112] G. C. McDonald. Ridge regression. *Wiley Interdisciplinary Reviews: Computational Statistics*, 1(1):93–100, 2009.
- [113] G. C. McDonald and D. I. Galarneau. A Monte Carlo evaluation of some ridge-type estimators. *Journal of the American Statistical Association*, 70(350):407–416, 1975.
- [114] W. Meeusen and J. van den Broeck. Efficiency estimation from Cobb-Douglas production functions with composed error. *International Economic Review*, 18(2):435–444, 1977.
- [115] N. Merhav. Physics of the Shannon limits. *IEEE Transactions on Information Theory*, 56(9):4274–4285, 2010.
- [116] G. Miller and D. Horn. Probability density estimation using entropy maximization. *Neural Computation*, 10(7):1925–1938, 1998.
- [117] S. K. Mishra. Multicollinearity and maximum entropy Leuven estimator. *Economics Bulletin*, 3(25):1–11, 2004.

- [118] S. K. Mishra. Estimation under multicollinearity: application of restricted Liu and maximum entropy estimators to the Portland cement dataset, 2004. Accessed in June, 2009 at <http://ssrn.com/abstract=559861>.
- [119] D. C. Montgomery, E. A. Peck, and G. G. Vining. *Introduction to Linear Regression Analysis*. John Wiley & Sons, Hoboken, New Jersey, 4th edition, 2006.
- [120] G. Muniz and B. M. G. Kibria. On some ridge regression estimators: an empirical comparisons. *Communications in Statistics - Simulation and Computation*, 38(3):621–630, 2009.
- [121] C. Nauges, C. O’Donnell, and J. Quiggin. Uncertainty and technical efficiency in Finnish agriculture. In *Proceedings of the 53rd Annual Conference of the Australian Agricultural and Resource Economics Society*, Cairns, Australia, 2009.
- [122] H. Nyquist. Certain factors affecting telegraph speed. *The Bell System Technical Journal*, 3(2):324–346, 1924.
- [123] H. Nyquist. Certain topics in telegraph transmission theory. *Transactions of the American Institute of Electrical Engineers*, 47(2):617–644, 1928.
- [124] C. J. O’Donnell and W. E. Griffiths. Estimating state-contingent production frontiers. *American Journal of Agricultural Economics*, 88(1):249–266, 2006.
- [125] C. J. O’Donnell, R. G. Chambers, and J. Quiggin. Efficiency analysis in the presence of uncertainty. *Journal of Productivity Analysis*, 33(1):1–17, 2010.
- [126] F. O’Sullivan. A statistical perspective on ill-posed inverse problems. *Statistical Science*, 1(4):502–518, 1986.
- [127] Q. Paris. Multicollinearity and maximum entropy estimators. *Economics Bulletin*, 3(11):1–9, 2001.
- [128] Q. Paris. Maximum entropy Leuven estimators and multicollinearity. *Statistica*, 64(1):3–22, 2004.
- [129] Q. Paris and R. E. Howitt. An analysis of ill-posed production problems using maximum entropy. *American Journal of Agricultural Economics*, 80(1):124–138, 1998.

- [130] S. Y. Park and A. K. Bera. Maximum entropy autoregressive conditional heteroskedasticity model. *Journal of Econometrics*, 150(2):219–230, 2009.
- [131] A. Plastino and A. R. Plastino. Tsallis entropy and Jaynes’ information theory formalism. *Brazilian Journal of Physics*, 29(1):50–60, 1999.
- [132] S. Polettini. Maximum entropy simulation for microdata protection. *Statistics and Computing*, 13(4):307–320, 2003.
- [133] P. V. Preckel. Least squares and entropy: a penalty function perspective. *American Journal of Agricultural Economics*, 83(2):366–377, 2001.
- [134] J. Quiggin and R. G. Chambers. The state-contingent approach to production under uncertainty. *The Australian Journal of Agricultural and Resource Economics*, 50(2):153–169, 2006.
- [135] M. G. Raizen. Demons, entropy and the quest for absolute zero. *Scientific American*, 304:54–59, 2011.
- [136] S. Rasmussen. Criteria for optimal production under uncertainty. The state-contingent approach. *The Australian Journal of Agricultural and Resource Economics*, 47(4):447–476, 2003.
- [137] S. Rasmussen and K. Karantininis. Estimating state-contingent production functions. In *Proceedings of the XIth Conference of the European Association of Agricultural Economists*, Copenhagen, Denmark, 2005.
- [138] A. E. Rastegin. Some general properties of unified entropies. *Journal of Statistical Physics*, 143(6):1120–1135, 2011.
- [139] A. Rényi. On measures of entropy and information. In *Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, pages 547–561. University of California Press, 1961.
- [140] A. Rényi. *Probability Theory*. North-Holland Publishing Company, Amsterdam, 1970.
- [141] P. J. Rousseeuw. Least median of squares regression. *Journal of the American Statistical Association*, 79(388):871–880, 1984.

- [142] P. J. Rousseeuw and A. M. Leroy. *Robust Regression and Outlier Detection*. John Wiley & Sons, New York, 1987.
- [143] P. J. Rousseeuw and V. Yohai. Robust regression by means of S-estimators. In J. Franke, W. Härdle, and D. Martin, editors, *Robust and Nonlinear Time Series Analysis*, Lecture Notes in Statistics, No. 26, pages 256–272. Springer, New York, 1984.
- [144] T. P. Ryan. *Modern Regression Methods*. John Wiley & Sons, Hoboken, New Jersey, 2nd edition, 2009.
- [145] A. Saboia, F. Toscano, and S. P. Walborn. Family of continuous-variable entanglement criteria using general entropy functions. *Physical Review A*, 83(3):1–8, 2011.
- [146] S. Sakalhoğlu and S. Kaçiranlar. A new biased estimator based on ridge estimation. *Statistical Papers*, 49(4):669–689, 2008.
- [147] R. J. V. Santos. Generalization of Shannon’s theorem for Tsallis entropy. *Journal of Mathematical Physics*, 38(8):4104–4107, 1997.
- [148] J. P. Sethna. *Statistical Mechanics: Entropy, Order Parameters, and Complexity*. Oxford University Press, Oxford, 2006.
- [149] C. E. Shannon. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3):379–423, 1948.
- [150] C. E. Shannon. Prediction and entropy of printed English. *The Bell System Technical Journal*, 30(1):50–64, 1951.
- [151] E. Z. Shen and J. M. Perloff. Maximum entropy and Bayesian approaches to the ratio problem. *Journal of Econometrics*, 104(2):289–313, 2001.
- [152] R. W. Shephard. *Cost and Production Functions*. Princeton University Press, Princeton, 1953.
- [153] R. W. Shephard. *Theory of Cost and Production Functions*. Princeton University Press, Princeton, 1970.

- [154] J. E. Shore and R. W. Johnson. Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on Information Theory*, IT-26(1):26–37, 1980.
- [155] M. J. Silvapulle. Robust ridge regression based on an M-estimator. *Australian Journal of Statistics*, 33(3):319–333, 1991.
- [156] J. R. Simpson and D. C. Montgomery. A biased-robust regression technique for the combined outlier-multicollinearity problem. *Journal of Statistical Computation and Simulation*, 56(1):1–22, 1996.
- [157] K. Singh. Breakdown theory for bootstrap quantiles. *The Annals of Statistics*, 26(5):1719–1732, 1998.
- [158] J. Skilling. Data analysis: the maximum entropy method. *Nature*, 309(5971):748–749, 1984.
- [159] J. Skilling. The axioms of maximum entropy. In G. J. Erickson and C. R. Smith, editors, *Maximum-Entropy and Bayesian Methods in Science and Engineering (Vol. I)*, pages 173–187. Kluwer Academic Publishers, 1988.
- [160] E. S. Soofi. Capturing the intangible concept of information. *Journal of the American Statistical Association*, 89(428):1243–1254, 1994.
- [161] E. S. Soofi and J. J. Retzer. Information indices: unification and applications. *Journal of Econometrics*, 107(1-2):17–40, 2002.
- [162] M. Stone and R. J. Brooks. Continuum regression: cross-validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal components regression. *Journal of the Royal Statistical Society, Series B*, 52(2):237–269, 1990.
- [163] D. F. Styer. Insight into entropy. *American Journal of Physics*, 68(12):1090–1096, 2000.
- [164] D. F. Styer. Entropy and evolution. *American Journal of Physics*, 76(11):1031–1033, 2008.

- [165] H. Suyari. Generalization of Shannon-Khinchin axioms to nonextensive systems and the uniqueness theorem for the nonextensive entropy. *IEEE Transactions on Information Theory*, 50(8):1783–1787, 2004.
- [166] I. J. Taneja. On generalized information measures and their applications. *Advances in Electronics and Electron Physics*, 76:327–413, 1989.
- [167] A. H. Tavares. *Aspectos Matemáticos da Entropia*. Master thesis, University of Aveiro, Aveiro, 2003.
- [168] R. Tibshirani. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society, Series B*, 58(1):267–288, 1996.
- [169] A. Tonini and R. Jongeneel. Modelling dairy supply for Hungary and Poland by generalised maximum entropy using prior information. *European Review of Agricultural Economics*, 35(2):219–246, 2008.
- [170] A. Tonini and V. Pedde. A generalized maximum entropy stochastic frontier measuring productivity accounting for spatial dependency. *Entropy*, 13(11):1916–1927, 2011.
- [171] M. N. Tran. Penalized maximum likelihood principle for choosing ridge parameter. *Communications in Statistics - Simulation and Computation*, 38(8):1610–1624, 2009.
- [172] M. Tribus and E. C. McIrvine. Energy and information. *Scientific American*, 3(225):179–188, 1971.
- [173] C. Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, 52(1-2):479–487, 1988.
- [174] M. Vila, A. Bardera, M. Feixas, and M. Sbert. Tsallis mutual information for document classification. *Entropy*, 13(9):1694–1707, 2011.
- [175] H. D. Vinod and J. López-de-Lacalle. Maximum entropy bootstrap for time series: the meboot R package. *Journal of Statistical Software*, 29(5):1–19, 2009.
- [176] N. Wiener. *Cybernetics: or Control and Communication in the Animal and the Machine*. MIT Press, Cambridge, Massachusetts, 1948.

- [177] H. Woods, H. H. Steinour, and H. R. Starke. Effect of composition of Portland cement on heat evolved during hardening. *Industrial and Engineering Chemistry*, 24(11):1207–1214, 1932.
- [178] A. Yaglom and I. Yaglom. *Probability and Information*. D. Reidel Publishing Company, Dordrecht, 1983.
- [179] A. Zaman, P. J. Rousseeuw, and M. Orhan. Econometric applications of high-breakdown robust regression techniques. *Economics Letters*, 71(1):1–8, 2001.
- [180] A. Zellner. Optimal information processing and Bayes’s theorem. *The American Statistician*, 42(4):278–284, 1988.
- [181] A. Zellner. Bayesian methods and entropy in economics and econometrics. In W. T. Grandy, Jr. and L. H. Schick, editors, *Maximum Entropy and Bayesian Methods*, Fundamental Theories of Physics, pages 17–31. Kluwer Academic Publishers, 1991.
- [182] A. Zellner. Information processing and Bayesian analysis. *Journal of Econometrics*, 107(1-2):41–50, 2002.
- [183] R. Zhang and G. C. McDonald. Characterization of ridge trace behavior. *Communications in Statistics - Theory and Methods*, 34(7):1487–1501, 2005.
- [184] K. D. Zieschang. An extended Farrell technical efficiency measure. *Journal of Economic Theory*, 33(2):387–396, 1984.
- [185] H. Zou and T. Hastie. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society, Series B*, 67(2):301–320, 2005.

# Index

- Collinearity, 39
- Cross-entropy, 36
- Data envelopment analysis, 45
- Directional technology distance function, 149
- Entropy, 14
- Generalized cross-entropy estimator, 36
- Generalized maximum entropy estimator, 32
- Higher-order generalized maximum entropy estimators, 37
- Ill-posed model, 2
- Information and entropy econometrics, 2
- Least absolute deviations estimator, 42
- Least median of squares estimator, 42
- Least trimmed squares estimator, 42
- Linear regression model, 31
- Maximum entropy, 25
- Maximum entropy Leuven estimator, 89
- Maximum entropy robust regression group (MERG) estimators, 92
- Maximum entropy robust regression group extended (MERGE) estimators, 109
- Outliers, 41
- Rényi entropy, 19
- Ridge parameter, 73
- Ridge regression, 73
- Ridge trace, 83
- Ridge-GME estimator, 75
- Robust regression, 42
- Shannon entropy, 16
- Small samples sizes, 3
- State-contingent production, 50
- Stochastic frontier analysis, 46
- Technical efficiency, 43
- Theory of light (quantum electrodynamics), 91
- Tsallis entropy, 20



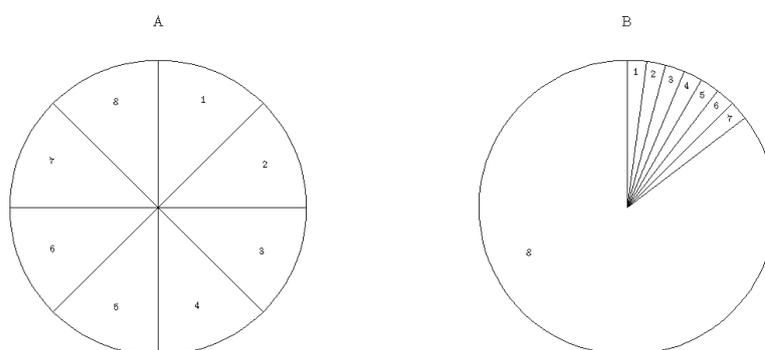
# Appendices



## Appendix A

In this appendix, the Shannon entropy is illustrated as a measure of information. Suppose that there are two roulette wheels each with eight pockets, as described in Figure A.1. The games are played by two individuals and both of them know the structure of the roulette wheels. One of them is the croupier, and the other one (who does not see where the ball falls) needs to find out the number of the pocket where the ball falls by asking binary questions, i.e., questions to which the answer is simply “yes” or “no”. If the two individuals play these games for a large number of times, what is, on average, the minimum number of questions needed to find out the number of the pocket where the ball falls? The answer to this question is provided by the Shannon entropy measure, presented in Definition 2.1.

**Figure A.1:** Shannon entropy as a measure of information.



In the game with the roulette A, the number of the pocket where the ball falls is always discovered with three questions. For example, the first question can be “The number is less than or equal to four?” and, depending on the answer (“yes” or “no”), the second question can be “The number is less than or equal to two?” or “The number is less than or equal to six?”. Again, depending on the answer, the number of the pocket where the ball falls is found with an adequate third question!

Note that if the individuals repeat this game for a large number of times, this is, on

average, the smallest number of questions needed to find out the number of the pocket where the ball falls. The individual, who is trying to find out the number of the pocket where the ball falls, can begin by asking, for example, “Is it the number seven?”, and he can hit the number! In this case, one question is enough to find out the pocket number, yet its probability is  $1/8$ . If he failed the first try, he can ask for another specific number. Now, the probability of success is  $1/7$ , and so on. Naturally, this is not the right strategy, especially when the number of pockets in the roulette wheel is very large.<sup>1</sup>

Consider the game with the roulette B in Figure A.1. The main difference between the two roulettes is the distribution of the pockets: uniform in roulette A and non-uniform in roulette B. Given the distribution of pockets in roulette B, the strategy used in the game with roulette A is feasible but it is not optimal. Note that it seems reasonable to start with the question “Is it the number eight?”, since the probability of the ball fall into this pocket is much higher than the probability of falling into one of the other pockets. Naturally, the answer can be “no”, and, in this case, more questions are needed to find out the pocket number. However, if the individuals repeat this game for a large number of times, on average, the smallest number of questions needed to find out the pocket number is one. By using information from the distribution of the pockets, it is possible to reduce the number of questions from three (if the same strategy in the game with the roulette A is applied) to only one question.

Since the game is performed with binary questions, the logarithm of base 2 is used in the Shannon entropy to measure the information in bits (binary digits). The Shannon entropy is viewed as a measure of the size of the missing information in these games. Since the distribution is uniform in roulette A, then  $H(p_1, p_2, \dots, p_K) = c \ln K$ . Thus, for  $K = 8$  and considering  $c = 1/\ln 2$ , it follows that  $H(p_1, p_2, \dots, p_8) = \log_2 8 = 3$  questions. This is the average number of binary questions needed to find out the number of the pocket where the ball falls in roulette A, as already discussed. Given that the distribution of the pockets is non-uniform in roulette B, then

$$H(p_1, p_2, \dots, p_8) = - \sum_{k=1}^8 p_k \log_2 p_k = -\frac{41}{48} \log_2 \frac{41}{48} - \frac{7}{48} \log_2 \frac{1}{48} \approx 1. \quad (\text{A.1})$$

If the individuals repeat this game with roulette B for a large number of times, on average,

---

<sup>1</sup>Considering  $N$  as the number of pockets (or the number of possibilities in a general game), the average number of questions needed with this strategy is  $(N + 1)/2$ .

the smallest number of questions needed to find out the pocket number is one.<sup>2</sup>

Several others interesting examples that illustrate the Shannon entropy as a measure of information can be found, among others, in Brillouin [17], Yaglom and Yaglom [178] and, at an introductory level, in Ben-Naim [11, 12].

---

<sup>2</sup>If the two individuals decide to play a game of guessing letters in an English text, how many binary questions *per* letter are needed, on average? Since the distribution of the 26 letters in English texts is not uniform, the number of questions needed is less than  $\log_2 26 \approx 4.7$ ; see Shannon [150].



## Appendix B

A directional distance function provides a complete representation of a production technology and is, by nature, a measure of technical inefficiency. Directional distance functions are discussed in Chambers et al. [26, 27], Färe and Grosskopf [47, 48], among others. Shephard's distance functions are special cases of the directional distance functions. The selection of the directional vector is an important challenge in directional distance functions, since different results emerge from different directional vectors. In this appendix, the most popular directional vectors used in the empirical literature are reviewed and two new measures of technical inefficiency are proposed.

The production technology  $T$ , presented in Definition 2.15, satisfies the traditional regularity conditions, namely,

- $T$  is a closed set;
- $T$  is a convex set;
- If  $(\mathbf{x}, \mathbf{y}) \in T$ ,  $\mathbf{x}^* \geq \mathbf{x}$  and  $\mathbf{y}^* \leq \mathbf{y}$ , then  $(\mathbf{x}^*, \mathbf{y}^*) \in T$ ;
- If  $(\mathbf{x}, \mathbf{y}) \in T$  and  $\mathbf{x} = \mathbf{0}$ , then  $\mathbf{y} = \mathbf{0}$ ;
- $(\mathbf{0}, \mathbf{0}) \in T$ .

The directional technology distance function is presented in the next definition; e.g., Chambers et al. [27].

**Definition B.1.** *Considering the production technology in Definition 2.15, the directional technology distance function is defined by*

$$\vec{D}_T(\mathbf{x}, \mathbf{y}; \mathbf{g}_x, \mathbf{g}_y) = \sup\{\beta : (\mathbf{x} - \beta\mathbf{g}_x, \mathbf{y} + \beta\mathbf{g}_y) \in T\}, \quad (\text{B.1})$$

where  $\mathbf{g} = (\mathbf{g}_x, \mathbf{g}_y) \in \mathbb{R}_+^N \times \mathbb{R}_+^M$  is a nonzero directional vector that defines the direction in which inputs are contracted and outputs are expanded.

Following Chambers et al. [27], the distance of  $(\mathbf{x}, \mathbf{y})$  to the production frontier, measured by (B.1), can be interpreted as an inefficiency measure, i.e., how much outputs can be

expanded and inputs can be contracted and still the input-output combination be technically feasible.

An important question arises: how to select the directional vector  $\mathbf{g}$ ? Table B.1 presents the directional vectors  $\mathbf{g} = (\mathbf{g}_x, \mathbf{g}_y)$  usually used in empirical work. A directional vector needs to be chosen by the researcher, taking into account that different choices lead to different results; e.g., Färe et al. [53].

**Table B.1:** Most popular directional vectors.

$(\mathbf{g}_x, \mathbf{g}_y)$	
$(\mathbf{x}, \mathbf{0})$	Inputs are contracted while holding outputs constant (the directional input distance function can be related to the Shephard input distance function).
$(\mathbf{0}, \mathbf{y})$	Outputs are expanded while holding inputs constant (the directional output distance function can be related to the Shephard output distance function).
$(\mathbf{x}, \mathbf{y})$	Simultaneous contraction of inputs and expansion of outputs in the direction that is determined by the input and output vector for each observation.
$(\mathbf{1}, \mathbf{1})$	Simultaneous contraction of inputs and expansion of outputs in the direction determined by $\mathbf{g} = (\mathbf{1}, \mathbf{1})$ . The directional vector is the same for all the observations.
$(\bar{\mathbf{x}}, \bar{\mathbf{y}})$	Simultaneous contraction of inputs and expansion of outputs in the direction determined by the mean of the data. The directional vector is the same for all the observations.

The first proposal for a new technical inefficiency measure is given by the distance of an input-output vector to the production frontier, measured using the directional vector  $\mathbf{g} = (\text{med } \mathbf{x}, \text{med } \mathbf{y})$ , where *med* represents the median of the data, i.e., the direction is determined by the median of the data.

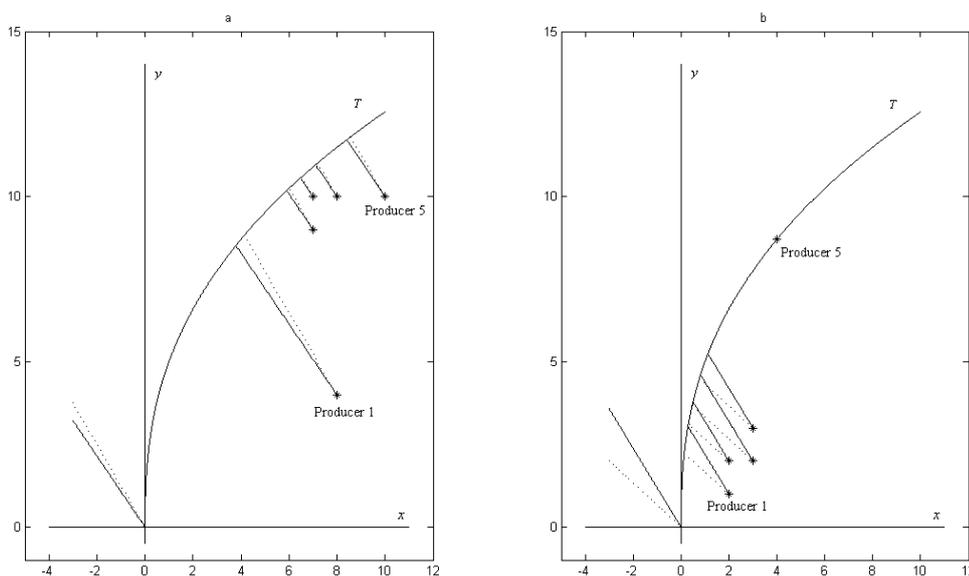
**Definition B.2.** *Considering the production technology in Definition 2.15 and the directional distance function in Definition B.1, a measure of technical inefficiency is given by*

$$\vec{D}_T(\mathbf{x}, \mathbf{y}; \text{med } \mathbf{x}, \text{med } \mathbf{y}). \quad (\text{B.2})$$

In the presence of outliers, the measure (B.2) is expected to provide more realistic values of technical inefficiency than other measures, in particular, the measure based on the directional

vector  $\mathbf{g} = (\bar{x}, \bar{y})$ . In order to illustrate the new measure of technical inefficiency in Definition B.2, consider two simple examples with a Cobb-Douglas production technology where five producers use one input,  $x$ , to produce one output,  $y$ . Figure B.1 illustrates this new measure of technical inefficiency (in dotted line) when compared with the measure  $\vec{D}_T(\mathbf{x}, \mathbf{y}; \bar{x}, \bar{y})$  (in solid line).

**Figure B.1:** Technical inefficiency measures based on the mean and median of the data.



**Table B.2:** Results from Figure B.1.

	Figure B.1a		Figure B.1b	
	$(\bar{x}, \bar{y})$	$(med\ x, med\ y)$	$(\bar{x}, \bar{y})$	$(med\ x, med\ y)$
Producer 1	6.18	6.18	2.66	2.24
Producer 2	1.60	1.59	2.34	2.03
Producer 3	0.76	0.75	3.40	3.04
Producer 4	1.31	1.29	2.92	2.68
Producer 5	2.34	2.31	0.00	0.00

Table B.2 presents the values of technical inefficiency, i.e., the distance of the input-output vector of each producer to the production frontier, measured using  $\mathbf{g} = (\text{med } \mathbf{x}, \text{med } \mathbf{y})$  and  $\mathbf{g} = (\bar{\mathbf{x}}, \bar{\mathbf{y}})$ . These examples suggest that, in the presence of outliers,  $\vec{D}_T(\mathbf{x}, \mathbf{y}; \text{med } \mathbf{x}, \text{med } \mathbf{y})$  is probably more adequate than  $\vec{D}_T(\mathbf{x}, \mathbf{y}; \bar{\mathbf{x}}, \bar{\mathbf{y}})$ . Further studies are needed on this technical inefficiency measure.

As already noted, the directional vector plays a key role in the measure of technical inefficiency with directional distance functions, since different directions lead to different distances and, consequently, different measures of technical inefficiency. Thus, how can this subjective choice be avoided? One possible answer is provided by the next technical inefficiency measure.

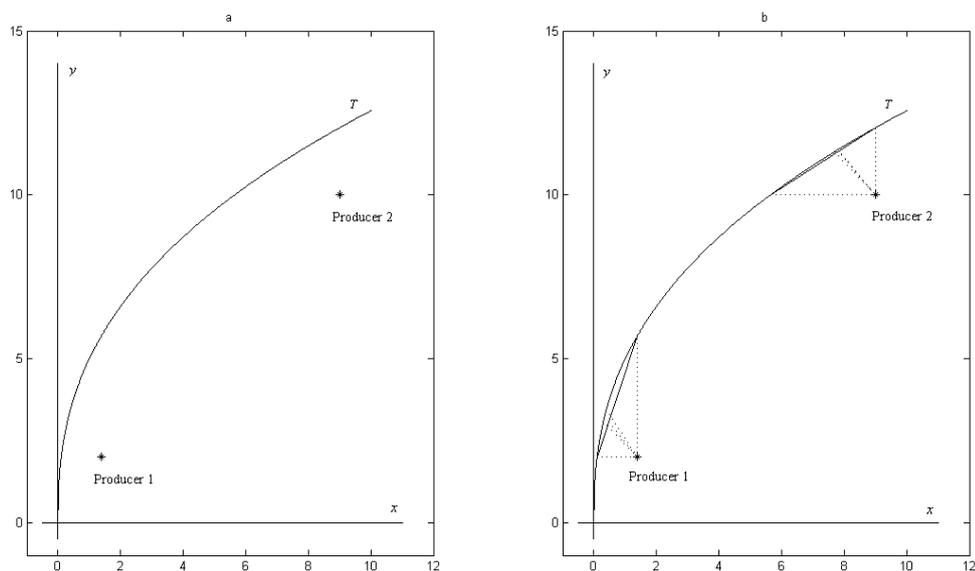
**Definition B.3.** *Considering the production technology in Definition 2.15 and the directional distance function in Definition B.1, the triangular measure of technical inefficiency, based on the directional vectors  $\mathbf{g} = (\mathbf{x}, \mathbf{0})$  and  $\mathbf{g} = (\mathbf{0}, \mathbf{y})$ , is given by*

$$\sqrt{\left[\vec{D}_T(\mathbf{x}, \mathbf{y}; \mathbf{x}, \mathbf{0})\right]^2 + \left[\vec{D}_T(\mathbf{x}, \mathbf{y}; \mathbf{0}, \mathbf{y})\right]^2}. \quad (\text{B.3})$$

The main idea underlying this new measure of technical inefficiency is to incorporate the information provided by the directional input distance function and the directional output distance function. Naturally, this measure of technical inefficiency takes the value of zero for an efficient producer. Note also that in a single input-output production technology, this measure is defined by the Pythagorean theorem.

To illustrate the technical inefficiency measure in Definition B.3, consider a simple example with a Cobb-Douglas production technology where two producers use one input,  $x$ , to produce one output,  $y$ . From figure B.2a, two questions arise: which one is the most efficient producer? and what is the value of inefficiency for each producer? The answers to these questions are not straightforward and depend on the directional vector chosen. Figure B.2b illustrates the two new measures of technical inefficiency presented in Definition B.2 and Definition B.3, and other different technical inefficiency measures based on the directional vectors presented in Table B.1. Table B.3 presents the results of technical inefficiency from Figure B.2b.

In this example, Producer 1 is the most efficient producer if  $\mathbf{g} = (\mathbf{x}, \mathbf{0})$ , but he becomes the most inefficient producer if  $\mathbf{g} = (\mathbf{0}, \mathbf{y})$ . Incorporating the information provided by the directional input distance function and the directional output distance function, the triangular measure of technical inefficiency (B.3) considers that both producers are equally inefficient.

**Figure B.2:** Different measures of technical inefficiency.**Table B.3:** Results from Figure B.2b.

	$(x, 0)$	$(0, y)$	$(x, y)$	$(1, 1)$	$(\bar{x}, \bar{y})$	(B.2)	(B.3)
Producer 1	1.29	3.70	1.75	1.54	1.61	1.61	3.92
Producer 2	3.34	2.04	1.83	1.85	1.82	1.82	3.92

In this appendix, two new measures of technical inefficiency are proposed using directional distance functions. The examples reveal that both measures can be useful: the first one in economic models affected by outliers and the second one appears to provide a more realistic evaluation of technical inefficiency. These results deserve further investigation that will be undertaken in the near future.



## Appendix C

In this appendix, some MATLAB codes are provided for some estimators presented in this thesis. Although different in most of the cases, these codes are based on the ones used in this work and are intentionally designed for users that are not familiar with the ME estimators.

In order to cope with convergence problems and feasible but non-optimal solutions arising in some examples and simulation studies, different optimization algorithms were tested using different programs, such as GAUSS, GAMS, LIMDEP and MATLAB (with different optimizers, namely SIMPS, CONSTR and FMINCON). The codes, presented next, are developed for MATLAB (using the FMINCON function), with possible efficiency losses when compared with the original codes used in this thesis. It is important to note that these codes were developed at the end of this work and have not been yet extensively tested.

Experienced users of MATLAB can easily verify that some features have been omitted for the sake of simplicity. Moreover, some codes are intentionally defined with some unnecessary detail (and probably computational efficiency losses), just to make the concepts and ideas understandable to a wider audience, such as functions, matrices and vectors definitions.

It is important to note that some errors that usually occur during execution are not due to incorrect programming codes, but simply because inputs, m-files and other necessary information are not properly defined by the user (e.g., for simplicity purposes, some functions are only suggested in the codes and the user needs to define them separately). Note also that different versions of MATLAB can cause difficulties with the use of some functions.

These and other codes will be soon available in <http://www.ua.pt/mat>, as well as new updates, examples and additional information on the ME estimators. It will also be available some of these codes for GAMS and Microsoft Excel to allow a more quick dissemination of these techniques.



```

function [p,lambda]=ME(x,y)
%
% Maximum Entropy (ME) formalism.
%
% [p,lambda]=ME(x,y)
% (version 1)
%
% p: estimate of the unknown parameters/probabilities;
% lambda: Lagrange multipliers;
% y: average value;
% x: row vector with possible outcomes.
%
% References:
% [1] Jaynes, E. T. (1957). Information theory and statistical mechanics.
%     Physical Review, 106, 620-630.
% [2] Golan, A., Judge, G. and Miller, D. (1996). Maximum Entropy
%     Econometrics: Robust Estimation With Limited Data. Wiley, Chichester.
%
% Pedro Macedo (pmacedo@ua.pt), CIDMA, DMat-UA, April 2012.

disp(' ')
disp('+++++')
disp('                Maximum Entropy (ME) formalism                ')
disp('+++++')
disp(' ')

[n k]=size(x);

if nargin ~= 2, error('ME requires two input arguments. See help ME.');
```

```

end

p=(1/k)*ones(k,1)';
lb=(1e-10)*ones(size(p));
ub=ones(size(p));
Aeq=[ones(size(p));x];
Beq=[1;y];
% See help fmincon. There are many interesting options available.
[p,fval,exitflag,output,lambda]=fmincon('FunME',p,[],[],Aeq,Beq,lb,ub,[],[]);
% FunME is the m-file with Shannon entropy that must be defined separately.
% For example:
%
% function f=FunME(p)
% p=p';
% f=-(-p'*log(p));

disp(' ')
disp('+++++')
disp(' ')
disp(sprintf('The ME objective value is %d.',abs(fval)));
disp(' ')
disp('The estimated ME distribution with MATLAB function FMINCON is')
```



```

function b=GMES(Y,X,method)
%
% Generalized Maximum Entropy estimator with the Shannon entropy (GMES).
%
% b=GMES(Y,X,method)
% (version 1)
%
% b: estimate of the unknown parameters;
% Y: vector of noisy observations;
% X: matrix with explanatory variables;
% method: optimization procedure
%       1 = FMINCON function from MATLAB;
%       2 = Newton's method with the dual objective function, designed by
%           Prof. Ximing Wu, University of California, Berkeley.
%
% References:
% [1] Golan, A., Judge, G. and Miller, D. (1996). Maximum Entropy
%     Econometrics: Robust Estimation With Limited Data. Wiley, Chichester.
% [2] Matlab programs for maximum entropy estimation available in:
%     - the Info-Metrics Institute, American University, Washington, D. C.
%       (http://www.american.edu/cas/economics/info-metrics);
%     - the web page of Prof. Ximing Wu, University of California, Berkeley
%       (http://agecon2.tamu.edu/people/faculty/wu-ximing);
%     - the web page of Prof. Neil Chriss, New York University, New York
%       (http://www.math.nyu.edu/faculty/chriss).
%
% Pedro Macedo (pmacedo@ua.pt), CIDMA, DMat-UA, April 2012.

disp(' ')
disp('+++++')
disp('      Generalized Maximum Entropy estimator with the Shannon entropy (GMES)      ')
disp('+++++')
disp(' ')

if nargin ~= 3, error('GMES requires three input arguments. See help GMES.'); end
if method ~= 1 && method ~= 2, error('Method must be 1 or 2. See help GMES.'); end

[n k]=size(X);
disp('Do you want to specify the same support interval for all the unknown')
resp=input('parameters of the model (yes/no)? [yes] ', 's');
disp(' ')
if isempty(resp)
    int1=input('Insert the support interval, [a,b], for all unknown parameters: ');
    disp(' ')
    disp('INFORMATION: usually the estimation is performed with five points in the')
    disp('parameter supports. Naturally, you can define a higher value.')
    disp(' ')
    m=input('Insert the number of points in each parameter support: ');
    incl=(int1(2)-int1(1))/(m-1);
    s1=int1(1):incl:int1(2);
    Z=zeros(k,k*m);
    for i=1:k
        pos=(i-1)*m+1;
        Z(i,pos:pos+m-1)=s1;
    end
else

```

```

disp('INFORMATION: to insert different supports you need to specify them in')
disp('this format: "[[a,b];[c,d];[e,f];...]"')
disp(sprintf('Note that, in this model, you must specify %d supports.',k))
disp(' ')
intg=input('Insert all the support intervals for the unknown parameters: ');
disp(' ')
disp('INFORMATION: usually the estimation is performed with five points in')
disp('the parameter supports. Naturally, you can define a higher value.')
disp(' ')
m=input('Insert the number of points in each parameter support: ');
s1=zeros(size(intg,1),m);
for i=1:size(intg,1)
    inc=(intg(i,2)-intg(i,1))/(m-1);
    s1(i,1:m)=intg(i,1):inc:intg(i,2);
end
Z=zeros(k,k*m);
for i=1:k
    pos=(i-1)*m+1;
    Z(i,pos:pos+m-1)=s1(i,1:m);
end
end
st1=round(-3*std(Y)); st2=round(3*std(Y));
st3=round(-4*std(Y)); st4=round(4*std(Y));
if st1==0, st1=-1; end
if st2==0, st2=1; end
if st3==0, st3=-1; end
if st4==0, st4=1; end
if issparse(X)
    R=qr(X);
else
    R=triu(qr(X));
end
b_ols=R\(R\'\'(X'*Y));
msres=((Y-X*b_ols)'*(Y-X*b_ols))/(n-k);
st1a=round(-3*sqrt(msres)); st2a=round(3*sqrt(msres));
st3a=round(-4*sqrt(msres)); st4a=round(4*sqrt(msres));
if st1a==0, st1a=-1; end
if st2a==0, st2a=1; end
if st3a==0, st3a=-1; end
if st4a==0, st4a=1; end
disp(' ')
disp('INFORMATION: the supports for the error component are usually defined by');
disp('the 3-sigma or 4-sigma rules, with sigma being the standard deviation of')
disp('the noisy observations, or an estimate of the error standard deviation');
disp('from the OLS estimation.');
```

```

disp(' ')
disp('Do you want to define the error supports using an estimate of the error');
resp1=input('standard deviation from the OLS residuals (yes/no)? [yes] ','s');
disp(' ')
if isempty(resp1)
    disp('The error supports can be defined by:')
    disp(sprintf(['%d,%d] (using the 3-sigma rule);',st1a,st2a))
    disp(sprintf(['%d,%d] (using the 4-sigma rule).',st3a,st4a))
else
    disp('Alternatively, using the standard deviation of the noisy observations,')
    disp('the error supports can be defined by:')
    disp(sprintf(['%d,%d] (using the 3-sigma rule);',st1,st2))

```

```

        disp(sprintf(['%d,%d] (using the 4-sigma rule).',st3,st4))
    end
    disp(' ')
    int2=input('Insert the support interval, [-a,a], for the error component: ');
        disp(' ')
        disp('INFORMATION: usually the estimation is performed with three points in')
        disp('the error supports. Naturally, you can define a higher value.')
        disp(' ')
    j=input('Insert the number of points in each error support: ');
    disp(' ')
    inc2=(int2(2)-int2(1))/(j-1);
    s2=int2(1):inc2:int2(2);
    V=zeros(n,n*j);
    for i=1:n
        pos=(i-1)*j+1;
        V(i,pos:pos+j-1)=s2;
    end
    disp('+++++')

    if method==1
        % FMINCON function from MATLAB
        p=(1/m)*ones(k*m,1);
        w=(1/j)*ones(n*j,1);
        pw=[p',w'];
        dp=length(p);
        lb=(1e-10)*ones(size(pw));
        ub=ones(size(pw));
        XZ=X*Z;
        matrixadditivity1=kron(eye(k,k),ones(1,m));
        matrixadditivity2=kron(eye(n,n),ones(1,j));
        Aeq=[XZ,V;matrixadditivity1,zeros(k,n*j);zeros(n,k*m),matrixadditivity2];
        Beq=[Y;ones(n+k,1)];
        % See help fmincon. There are many interesting options available.
        a=fmincon('FunGMES',pw,[],[],Aeq,Beq,lb,ub,[],[],dp);
        % FunGMES is the m-file with the objective function that must be defined
        % separately. For example:
        %
        % function f=FunGMES(pw,dp)
        % p=pw(1:dp)';
        % w=pw(dp+1:end)';
        % f=-(-p'*log(p)-w'*log(w));
        %
        p=a(1:dp);
        b=Z*p';
    else
        % Based on the Newton's method with the dual objective function, designed by
        % Prof. Ximing Wu, University of California, Berkeley.
        i=size(Z,2);
        S=repmat(s2,n,1);
        t=1;
        m=1;
        lambda=zeros(n,1);
        increm=zeros(n,1);
        iter=0;
        while (m>1e-5)
            iter=iter+1;

```

```

lambda=lambda+incred;
newz=exp(-X'*lambda*ones(1,i).*Z);
p9=newz./(newz*ones(i,i));
newv=exp(-lambda*ones(1,j).*S);
w9=newv./(newv*ones(j,j));
g9=Y-X*(Z.*p9*ones(i,1))-S.*w9*ones(j,1);
inv_z=diag((sum((p9.*(Z.^2)),2)-sum((p9.*Z),2).^2).^(-1));
inv_v=diag((sum((w9.*(S.^2)),2)-sum((w9.*S),2).^2).^(-1));
temp=inv_v*X;
inv_H=-inv_v+temp*inv(inv_z+X'*temp)*temp';
incred=inv_H*g9;
t0=t;
t=g9'*incred;
m=abs(t-t0);
end
b=sum((p9.*Z),2);
end

disp(' ')
disp('+++++')
disp(' ')
if method==1
disp('[Solution with the FMINCON function from MATLAB]')
else
disp('[Solution using the Newton's method with the dual objective function]')
end
end

```

```

function eta=RidgeGME(Y,X)
%
% Ridge-GME estimator of the ridge parameter in ridge regression.
%
% eta=RidgeGME(Y,X)
% (version 1)
%
% eta: estimate of the ridge parameter;
% Y: vector of noisy observations;
% X: matrix with explanatory variables.
%
% References:
% [1] Macedo, P., Scotto, M. and Silva, E. (2010). On the choice of the ridge
% parameter: a maximum entropy approach. Communications in Statistics -
% Simulation and Computation, 39(8), 1628-1638.
% [2] Hoerl, A. E. and Kennard, R. W. (1970). Biased estimation for
% nonorthogonal problems. Technometrics, 12(1), 55-67.
% [3] Golan, A., Judge, G. and Miller, D. (1996). Maximum Entropy
% Econometrics: Robust Estimation With Limited Data. Wiley, Chichester.
%
% Pedro Macedo (pmacedo@ua.pt), CIDMA, DMat-UA, April 2012.

disp(' ')
disp('+++++')
disp('          Ridge-GME estimator of the ridge parameter in ridge regression          ')
disp('+++++')
disp(' ')

if nargin ~= 2,
    error('RidgeGME requires two input arguments. See help RidgeGME.');
```

```

end

[n k]=size(X);
disp('INFORMATION: the ridge trace plays a key-role in the Ridge-GME estimator')
disp('and, thus, it is automatically generated at the beginning of this code.')
```

```

eta=linspace(0,1,1000);
base=zeros(k,1000);
for i=1:1000
b=inv(X'*X+eta(i)*eye(k))*X'*Y;
base(1:k,i)=b;
end
b=base;
plot(eta,b');
title('Ridge trace with non-standardized coefficients')
```

```

disp(' ')
disp('+++++')
disp(' ')
disp('INFORMATION: based on the ridge trace you need to specify the ridge interval.')
```

```

disp(' ')
ridgeinterval=input('Insert the ridge interval in this format [a,b]: ');
disp(' ')
disp('INFORMATION: based on the ridge trace you need to specify the supports for')
```

```

disp('the associated GME estimator. To insert the supports you need to specify')
disp('them in this format: "[[a,b];[c,d];[e,f];...]".')
```

```

disp(sprintf('Note that, in this model, you must specify %d supports.',k))
disp(' ')

```

```

intg=input('Insert all the support intervals for the unknown parameters: ');
disp(' ')
disp('INFORMATION: usually the estimation is performed with five points in')
disp('the parameter supports. Naturally, you can define a higher value.')
disp(' ')
m=input('Insert the number of points in each parameter support: ');
s1=zeros(size(intg,1),m);
for i=1:size(intg,1)
    inc=(intg(i,2)-intg(i,1))/(m-1);
    s1(i,1:m)=intg(i,1):inc:intg(i,2);
end
Z=zeros(k,k*m);
for i=1:k
    pos=(i-1)*m+1;
    Z(i,pos:pos+m-1)=s1(i,1:m);
end
st1=round(-3*std(Y)); st2=round(3*std(Y));
st3=round(-4*std(Y)); st4=round(4*std(Y));
if st1==0, st1=-1; end
if st2==0, st2=1; end
if st3==0, st3=-1; end
if st4==0, st4=1; end
if issparse(X)
    R=qr(X);
else
    R=triu(qr(X));
end
b_ols=R\(R\'(X'*Y));
msres=((Y-X*b_ols)'*(Y-X*b_ols))/(n-k);
st1a=round(-3*sqrt(msres)); st2a=round(3*sqrt(msres));
st3a=round(-4*sqrt(msres)); st4a=round(4*sqrt(msres));
if st1a==0, st1a=-1; end
if st2a==0, st2a=1; end
if st3a==0, st3a=-1; end
if st4a==0, st4a=1; end
disp(' ')
disp('INFORMATION: the supports for the error component are usually defined by');
disp('the 3-sigma or 4-sigma rules, with sigma being the standard deviation of')
disp('the noisy observations, or an estimate of the error standard deviation');
disp('from the OLS estimation. ');
disp(' ')
disp('Do you want to define the error supports using an estimate of the error');
resp1=input('standard deviation from the OLS residuals (yes/no)? [yes] ', 's');
disp(' ')
if isempty(resp1)
    disp('The error supports can be defined by:')
    disp(sprintf(['%d,%d] (using the 3-sigma rule);',st1a,st2a))
    disp(sprintf(['%d,%d] (using the 4-sigma rule).',st3a,st4a))
else
    disp('Alternatively, using the standard deviation of the noisy observations,')
    disp('the error supports can be defined by:')
    disp(sprintf(['%d,%d] (using the 3-sigma rule);',st1,st2))
    disp(sprintf(['%d,%d] (using the 4-sigma rule).',st3,st4))
end
disp(' ')
int2=input('Insert the support interval, [-a,a], for the error component: ');
disp(' ')

```

```

disp('INFORMATION: usually the estimation is performed with three points in')
disp('the error supports. Naturally, you can define a higher value.')
disp(' ')
j=input('Insert the number of points in each error support: ');
disp(' ')
inc2=(int2(2)-int2(1))/(j-1);
s2=int2(1):inc2:int2(2);
V=zeros(n,n*j);
for i=1:n
    pos=(i-1)*j+1;
    V(i,pos:pos+j-1)=s2;
end
disp('+++++')

p=(1/m)*ones(k*m,1);
w=(1/j)*ones(n*j,1);
pw=[p',w'];
dp=length(p);
lb=(1e-10)*ones(size(pw));
ub=ones(size(pw));
XZ=X*Z;
matrixadditivity1=kron(eye(k,k),ones(1,m));
matrixadditivity2=kron(eye(n,n),ones(1,j));
Aeq=[XZ,V;matrixadditivity1,zeros(k,n*j);zeros(n,k*m),matrixadditivity2];
Beq=[Y;ones(n+k,1)];
% See help fmincon. There are many interesting options available.
a=fmincon('FunGMES',pw,[],[],Aeq,Beq,lb,ub,[],[],dp);
% FunGMES is the m-file with the objective function that must be defined
% separately. For example:
%
% function f=FunGMES(pw,dp)
% p=pw(1:dp)';
% w=pw(dp+1:end)';
% f=-(-p'*log(p)-w'*log(w));
%
p=a(1:dp);
b_gme=Z*p';

a=ridgeinterval(1);
b=ridgeinterval(2);
d=zeros(1000,1);
etavec=linspace(a,b,1000);
for i=1:1000
    b_ridge=inv(X'*X+etavec(i)*eye(k))*X'*Y;
    difnorminf=norm(b_gme-b_ridge,inf);
    d(i,1)=difnorminf;
end
[Y1,I1]=min(d);
eta=etavec(I1);
disp('+++++')
disp(' ')
disp('The estimate of the ridge parameter is')

```



```

function b=MERG(Y,X)
%
% Maximum Entropy Robust Regression Group (MERG) estimators.
%
% b=MERG(Y,X)
% (experimental version)
%
% b: estimate of the unknown parameters;
% Y: vector of noisy observations;
% X: matrix with explanatory variables.
%
% References:
% [1] Macedo, P., Scotto, M. and Silva, E. (2010). A general class of
%     estimators for the linear regression model affected by collinearity
%     and outliers. Communications in Statistics - Simulation and
%     Computation, 39(5), 981-993.
% [2] Paris, Q. (2001). Multicollinearity and maximum entropy estimators.
%     Economics Bulletin, 3(11), 1-9.
%
% Pedro Macedo (pmacedo@ua.pt), CIDMA, DMat-UA, May 2012.

disp(' ')
disp('+++++')
disp('          Maximum Entropy Robust Regression Group (MERG) estimators          ')
disp('+++++')
disp(' ')

if nargin ~= 2, error('MERG requires two input arguments. See help MERG.');
```

```

end

[n k]=size(X);
disp('INFORMATION: this code needs an initial solution for beta (the unknown')
disp('parameters). In this experimental version, four options are available:')
disp('1. A vector of ones is used;')
disp('2. The solution from the ordinary least squares estimator is used;')
disp('3. The solution from the iteratively reweighted least squares estimator')
disp('   is used (applying the bisquare weighting function);')
disp('4. The code uses the initial solution provided by you.')
disp(' ')
answer=input('Select your option: ');
if answer == 1
    b=ones(k,1);
elseif answer == 2
    if issparse(X)
        R=qr(X);
    else
        R=triu(qr(X));
    end
    b_ols=R\(R\'\'(X'*Y));
    b=b_ols;
elseif answer == 3
    b=robustfit(X,Y,'','','off');
else
    disp(' ')
    disp('INFORMATION: the solution must be specified in a column vector.')
    disp(' ')
    b=input('Insert the initial solution for beta: ');

```

```

end
disp(' ')
disp('+++++')
res=Y-X*b;
br=[b',res'];
lb=[-inf*ones(k,1)',-inf*ones(n,1)'];
ub=[inf*ones(k,1)',inf*ones(n,1)'];
Aeq=[X,eye(n)];
Beq=Y;
% See help fmincon. There are many interesting options available.
[a,exitflag]=fmincon('FunMERG',br,[],[],Aeq,Beq,lb,ub,[],[],k);
% FunMERG is the m-file with the objective function that must be defined
% separately. For example, for MERG1:
%
% function f=FunMERG(br,k)
% b=br(1:k)';
% res=br(k+1:end)';
% lbeta=b'*b;
% p=b.^2/lbeta;
% f=p'*log(p+1e-6)+lbeta*log(lbeta+1e-6)+res'*res;
%
b=a(1:k)';

disp('+++++')
% See help fmincon for details on exitflag.
display(exitflag);

```

```

function [b,StdError]=BootstrapStdError(Y,X,nt)
%
% Estimation of standard errors for the parameters in a linear regression
% model by resampling residuals (using the bootstrp function in MATLAB).
%
% [b,StdError]=BootstrapStdError(Y,X,nt)
% (experimental version)
%
%
% b: estimate of the unknown parameters;
% StdError: estimate of the standard errors;
% Y: vector of noisy observations;
% X: matrix with explanatory variables;
% nt: number of trials (a large number, e.g., nt>1000).
%
% Pedro Macedo (pmacedo@ua.pt), CIDMA, DMat-UA, May 2012.

disp(' ')
disp('+++++')
disp('          Estimation of standard errors by resampling residuals          ')
disp('+++++')
disp(' ')

if nargin ~= 3,
    error('This function requires three input arguments. See help.');
```

```

end

[n k]=size(X);
disp('INFORMATION: in this experimental version, only three methods are available:')
disp('1. Ridge regression estimator;')
disp('2. MERG1 estimator;')
disp('3. GME estimator.')
disp(' ')
answer=input('Select your option: ');

if answer == 1

disp(' ')
disp('+++++')
disp(' ')
disp('INFORMATION: the Ridge-GME estimator is recommended to select an estimate of')
disp('the ridge parameter. See the RidgeGME code.')
disp(' ')
eta=input('Insert an estimate of the ridge parameter: ');
b=inv(X'*X+eta*eye(k))*X'*Y;
Yhat=X*b;
res=Y-X*b;
StdError=std(bootstrp(nt,@(bootr) RIDGEbootstrp(Yhat+bootr,X,eta),res))';
% RIDGEbootstrp is the m-file with the ridge regression estimator that must
% be defined separately. For example:
%
% function b=RIDGEbootstrp(Y,X,eta)
% [n k]=size(X);
% b=inv(X'*X+eta*eye(k))*X'*Y;

elseif answer == 2

```

```

disp(' ')
disp('+++++')
disp(' ')
disp('INFORMATION: this code needs an initial solution for beta (the unknown')
disp('parameters). In this experimental version, four options are available:')
disp('1. A vector of ones is used;')
disp('2. The solution from the ordinary least squares estimator is used;')
disp('3. The solution from the iteratively reweighted least squares estimator')
disp(' is used (applying the bisquare weighting function);')
disp('4. The code uses the initial solution provided by you.')
disp(' ')
answer9=input('Select your option: ');
disp(' ')
disp('+++++')
if answer9 == 1
    b9=ones(k,1);
elseif answer9 == 2
    if issparse(X)
        R=qr(X);
    else
        R=triu(qr(X));
    end
    b_ols=R\(R\'\'(X'*Y));
    b9=b_ols;
elseif answer9 == 3
    b9=robustfit(X,Y,','',','off');
else
    disp(' ')
    disp('INFORMATION: the solution must be specified in a column vector.')
    disp(' ')
    b9=input('Insert the initial solution for beta: ');
end
res=Y-X*b9;
br=[b9',res'];
lb=[-inf*ones(k,1)',-inf*ones(n,1)'];
ub=[inf*ones(k,1)',inf*ones(n,1)'];
Aeq=[X,eye(n)];
Beq=Y;
a=fmincon('FunMERC',br,[],[],Aeq,Beq,lb,ub,[],[],k);
b=a(1:k)';
Yhat=X*b;
res=Y-X*b;
StdError=std(bootstrp(nt,@(bootr) MERGbootstrap(Yhat+bootr,X,b9),res))';
% MERGbootstrap is the m-file with a simplified version of the MERG1 estimator
% that must be defined separately. For example:
%
% function b=MERCbootstrap(Y,X,b9)
% [n k]=size(X);
% res=Y-X*b9;
% br=[b9',res'];
% lb=[-inf*ones(k,1)',-inf*ones(n,1)'];
% ub=[inf*ones(k,1)',inf*ones(n,1)'];
% Aeq=[X,eye(n)];
% Beq=Y;
% a=fmincon('FunMERC',br,[],[],Aeq,Beq,lb,ub,[],[],k);
% b=a(1:k)';

```

```

else

disp(' ')
disp('+++++')
disp(' ')
disp('INFORMATION: the supports for the parameters of the model must be defined')
disp('in this format: "[[a,b];[c,d];[e,f];...]"')
disp(sprintf('Note that, in this model, you must specify %d supports.',k))
disp(' ')
intg=input('Insert all the supports for the unknown parameters: ');
disp(' ')
disp('INFORMATION: usually the estimation is performed with five points in')
disp('the parameter supports. Naturally, you can define a higher value.')
disp(' ')
m=input('Insert the number of points in each parameter support: ');
s1=zeros(size(intg,1),m);
for i=1:size(intg,1)
    inc=(intg(i,2)-intg(i,1))/(m-1);
    s1(i,1:m)=intg(i,1):inc:intg(i,2);
end
Z=zeros(k,k*m);
for i=1:k
    pos=(i-1)*m+1;
    Z(i,pos:pos+m-1)=s1(i,1:m);
end
st1=round(-3*std(Y)); st2=round(3*std(Y));
st3=round(-4*std(Y)); st4=round(4*std(Y));
if st1==0, st1=-1; end
if st2==0, st2=1; end
if st3==0, st3=-1; end
if st4==0, st4=1; end
if issparse(X)
    R=qr(X);
else
    R=triu(qr(X));
end
b_ols=R\(R\'(X\'*Y));
msres=((Y-X*b_ols)\'*(Y-X*b_ols))/(n-k);
st1a=round(-3*sqrt(msres)); st2a=round(3*sqrt(msres));
st3a=round(-4*sqrt(msres)); st4a=round(4*sqrt(msres));
if st1a==0, st1a=-1; end
if st2a==0, st2a=1; end
if st3a==0, st3a=-1; end
if st4a==0, st4a=1; end
disp(' ')
disp('INFORMATION: the supports for the error component are usually defined by');
disp('the 3-sigma or 4-sigma rules, with sigma being the standard deviation of')
disp('the noisy observations, or an estimate of the error standard deviation');
disp('from the OLS estimation.');
```

```

disp(' ')
disp('Do you want to define the error supports using an estimate of the error');
respl=input('standard deviation from the OLS residuals (yes/no)? [yes] ','s');
disp(' ')
if isempty(respl)
    disp('The error supports can be defined by:')
    disp(sprintf('[%d,%d] (using the 3-sigma rule)',st1a,st2a))
    disp(sprintf('[%d,%d] (using the 4-sigma rule).',st3a,st4a))
end

```

```

else
    disp('Alternatively, using the standard deviation of the noisy observations,')
    disp('the error supports can be defined by:')
    disp(sprintf(['%d,%d] (using the 3-sigma rule);',st1,st2))
    disp(sprintf(['%d,%d] (using the 4-sigma rule).',st3,st4))
end
disp(' ')
int2=input('Insert the support interval, [-a,a], for the error component: ');
disp(' ')
disp('INFORMATION: usually the estimation is performed with three points in')
disp('the error supports. Naturally, you can define a higher value.')
disp(' ')
j=input('Insert the number of points in each error support: ');
disp(' ')
inc2=(int2(2)-int2(1))/(j-1);
s2=int2(1):inc2:int2(2);
V=zeros(n,n*j);
for i=1:n
    pos=(i-1)*j+1;
    V(i,pos:pos+j-1)=s2;
end
disp('+++++')
p=(1/m)*ones(k*m,1);
w=(1/j)*ones(n*j,1);
pw=[p',w'];
dp=length(p);
lb=(1e-10)*ones(size(pw));
ub=ones(size(pw));
XZ=X*Z;
matrixadditivity1=kron(eye(k,k),ones(1,m));
matrixadditivity2=kron(eye(n,n),ones(1,j));
Aeq=[XZ,V;matrixadditivity1,zeros(k,n*j);zeros(n,k*m),matrixadditivity2];
Beq=[Y;ones(n+k,1)];
a=fmincon('FunGMES',pw,[],[],Aeq,Beq,lb,ub,[],[],dp);
p=a(1:dp);
b=Z*p';
Yhat=X*b;
res=Y-X*b;
StdError=std(bootstrp(nt,@(bootr) GMESbootstrap(Yhat+bootr,X,intg,m,int2,j),res));
% GMESbootstrap is the m-file with a simplified version of the GMES estimator
% that must be defined separately. For example:
%
% function b=GMESbootstrap(Y,X,intg,m,int2,j)
% [n k]=size(X);
% s1=zeros(size(intg,1),m);
% for i=1:size(intg,1)
%     inc=(intg(i,2)-intg(i,1))/(m-1);
%     s1(i,1:m)=intg(i,1):inc:intg(i,2);
% end
% Z=zeros(k,k*m);
% for i=1:k
%     pos=(i-1)*m+1;
%     Z(i,pos:pos+m-1)=s1(i,1:m);
% end
% inc2=(int2(2)-int2(1))/(j-1);
% s2=int2(1):inc2:int2(2);
% V=zeros(n,n*j);

```





```

function [b,CovMatrix,StdError]=Covariance(Y,X,R)
%
% Bootstrap estimator for the asymptotic covariance matrix.
%
% [b,CovMatrix,StdError]=Covariance(Y,X,R)
% (experimental version)
%
% b: estimate of the unknown parameters;
% CovMatrix: estimate of the asymptotic covariance matrix;
% StdError: estimate of the standard errors;
% Y: vector of noisy observations;
% X: matrix with explanatory variables;
% R: number of replications.
%
% References:
% [1] Greene, W. H. (2008). Econometric Analysis. 6th ed., Pearson Prentice
%     Hall, Upper Saddle River, New Jersey.
%
% Pedro Macedo (pmacedo@ua.pt), CIDMA, DMat-UA, May 2012.

disp(' ')
disp('+++++')
disp('          Bootstrap estimator for the asymptotic covariance matrix          ')
disp('+++++')
disp(' ')

if nargin ~= 3,
    error('This function requires three input arguments. See help.');
```

```

end

[n k]=size(X);
disp('INFORMATION: in this experimental version, only three methods are available:');
disp('1. Ridge regression estimator;');
disp('2. MERG1 estimator;');
disp('3. GME estimator.');
```

```

disp(' ')
answer=input('Select your option: ');

if answer == 1

disp(' ')
disp('+++++')
disp(' ')
disp('INFORMATION: the Ridge-GME estimator is recommended to select an estimate of')
disp('the ridge parameter. See the RidgeGME code.')
```

```

disp(' ')
eta=input('Insert an estimate of the ridge parameter: ');
b0=inv(X'*X+eta*eye(k))*X'*Y;
M=zeros(k,k);
for ii=1:R
[bootstat,bootsam]=bootstrp(1,[],X);
X1=zeros(n,k);
for i=1:n
    X1(i,1:k)=X(bootsam(i),1:k);
end
Y1=zeros(n,1);
```

```

for i=1:n
    Y1(i,1)=Y(bootsam(i),1);
end
b1=inv(X1'*X1+eta*eye(k))*X1'*Y1;
dp1=(b1-b0)*(b1-b0)';
M=M+dp1;
end
b=b0;
CovMatrix=(1/R)*M;
StdError=sqrt(diag(CovMatrix));

elseif answer == 2

disp(' ')
disp('+++++')
disp(' ')
disp('INFORMATION: this code needs an initial solution for beta (the unknown')
disp('parameters). In this experimental version, four options are available:')
disp('1. A vector of ones is used;')
disp('2. The solution from the ordinary least squares estimator is used;')
disp('3. The solution from the iteratively reweighted least squares estimator')
disp(' is used (applying the bisquare weighting function);')
disp('4. The code uses the initial solution provided by you.')
disp(' ')
answer9=input('Select your option: ');
disp(' ')
disp('+++++')
if answer9 == 1
    b=ones(k,1);
elseif answer9 == 2
    if issparse(X)
        Rm=qr(X);
    else
        Rm=triu(qr(X));
    end
    b_ols=Rm\(Rm\'(X'*Y));
    b=b_ols;
elseif answer9 == 3
    b=robustfit(X,Y,','',','off');
else
    disp(' ')
    disp('INFORMATION: the solution must be specified in a column vector.')
    disp(' ')
    b=input('Insert the initial solution for beta: ');
end
res=Y-X*b;
br=[b',res'];
lb=[-inf*ones(k,1)',-inf*ones(n,1)'];
ub=[inf*ones(k,1)',inf*ones(n,1)'];
Aeq=[X,eye(n)];
Beq=Y;
a=fmincon('FunMERG',br,[],[],Aeq,Beq,lb,ub,[],[],k);
b0=a(1:k)';
M=zeros(k,k);
for ii=1:R
    [bootstat,bootsam]=bootstrp(1,[],X);
    X1=zeros(n,k);

```

```

for i=1:n
    X1(i,1:k)=X(bootsam(i),1:k);
end
Y1=zeros(n,1);
for i=1:n
    Y1(i,1)=Y(bootsam(i),1);
end
res=Y1-X1*b;
br=[b',res'];
lb=[-inf*ones(k,1)',-inf*ones(n,1)'];
ub=[inf*ones(k,1)',inf*ones(n,1)'];
Aeq=[X1,eye(n)];
Beq=Y1;
a=fmincon('FunMERG',br,[],[],Aeq,Beq,lb,ub,[],[],k);
b1=a(1:k)';
dp1=(b1-b0)*(b1-b0)';
M=M+dp1;
end
b=b0;
CovMatrix=(1/R)*M;
StdError=sqrt(diag(CovMatrix));

else

disp(' ')
disp('+++++')
disp(' ')
disp('INFORMATION: the supports for the parameters of the model must be defined')
disp('in this format: "[[a,b];[c,d];[e,f];...]"')
disp(sprintf('Note that, in this model, you must specify %d supports.',k))
disp(' ')
intg=input('Insert all the supports for the unknown parameters: ');
disp(' ')
disp('INFORMATION: usually the estimation is performed with five points in')
disp('the parameter supports. Naturally, you can define a higher value.')
disp(' ')
m=input('Insert the number of points in each parameter support: ');
s1=zeros(size(intg,1),m);
for i=1:size(intg,1)
    inc=(intg(i,2)-intg(i,1))/(m-1);
    s1(i,1:m)=intg(i,1):inc:intg(i,2);
end
Z=zeros(k,k*m);
for i=1:k
    pos=(i-1)*m+1;
    Z(i,pos:pos+m-1)=s1(i,1:m);
end
st1=round(-3*std(Y)); st2=round(3*std(Y));
st3=round(-4*std(Y)); st4=round(4*std(Y));
if st1==0, st1=-1; end
if st2==0, st2=1; end
if st3==0, st3=-1; end
if st4==0, st4=1; end
if issparse(X)
    Rm=qr(X);
else
    Rm=triu(qr(X));

```

```

end
b_ols=Rm\Rm\'(X'*Y);
msres=((Y-X*b_ols)'*(Y-X*b_ols))/(n-k);
st1a=round(-3*sqrt(msres)); st2a=round(3*sqrt(msres));
st3a=round(-4*sqrt(msres)); st4a=round(4*sqrt(msres));
if st1a==0, st1a=-1; end
if st2a==0, st2a=1; end
if st3a==0, st3a=-1; end
if st4a==0, st4a=1; end
disp(' ')
disp('INFORMATION: the supports for the error component are usually defined by');
disp('the 3-sigma or 4-sigma rules, with sigma being the standard deviation of');
disp('the noisy observations, or an estimate of the error standard deviation');
disp('from the OLS estimation.');
```

disp(' ')

```
disp('Do you want to define the error supports using an estimate of the error');
resp1=input('standard deviation from the OLS residuals (yes/no)? [yes] ','s');
disp(' ')
if isempty(resp1)
    disp('The error supports can be defined by:');
    disp(sprintf(['%d,%d] (using the 3-sigma rule);',st1a,st2a))
    disp(sprintf(['%d,%d] (using the 4-sigma rule).',st3a,st4a))
else
    disp('Alternatively, using the standard deviation of the noisy observations,')
    disp('the error supports can be defined by:');
    disp(sprintf(['%d,%d] (using the 3-sigma rule);',st1,st2))
    disp(sprintf(['%d,%d] (using the 4-sigma rule).',st3,st4))
end
disp(' ')
int2=input('Insert the support interval, [-a,a], for the error component: ');
disp(' ')
disp('INFORMATION: usually the estimation is performed with three points in')
disp('the error supports. Naturally, you can define a higher value.')
```

disp(' ')

```
j=input('Insert the number of points in each error support: ');
disp(' ')
inc2=(int2(2)-int2(1))/(j-1);
s2=int2(1):inc2:int2(2);
V=zeros(n,n*j);
for i=1:n
    pos=(i-1)*j+1;
    V(i,pos:pos+j-1)=s2;
end
disp('+++++')
p=(1/m)*ones(k*m,1);
w=(1/j)*ones(n*j,1);
pw=[p',w'];
dp=length(p);
lb=(1e-10)*ones(size(pw));
ub=ones(size(pw));
XZ=X*Z;
matrixadditivity1=kron(eye(k,k),ones(1,m));
matrixadditivity2=kron(eye(n,n),ones(1,j));
Aeq=[XZ,V;matrixadditivity1,zeros(k,n*j);zeros(n,k*m),matrixadditivity2];
Beq=[Y;ones(n+k,1)];
a=fmincon('FunGMES',pw,[],[],Aeq,Beq,lb,ub,[],[],dp);
p=a(1:dp);
```



