



# Essays on statistical equilibrium and competition in local public goods : philosophical explorations and empirical perspectives

Juan C. Melo

## ► To cite this version:

Juan C. Melo. Essays on statistical equilibrium and competition in local public goods : philosophical explorations and empirical perspectives. Economics and Finance. Université Panthéon-Sorbonne - Paris I, 2022. English. NNT : 2022PA01E018 . tel-03714130

**HAL Id: tel-03714130**

**<https://theses.hal.science/tel-03714130>**

Submitted on 5 Jul 2022

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



**UNIVERSITÉ PARIS I PANTHÉON SORBONNE**

UFR 02

Centre d'Économie de la Sorbonne

**THÈSE**

Pour l'obtention du titre de Docteur en Sciences Économiques

Présentée et soutenue publiquement

le 22 FÉVRIER 2022 par

**Juan C. MELO**

**Essays on Statistical Equilibrium and Competition in Local  
Public Goods**

**Philosophical Explorations and Empirical Perspectives**

**Sous la direction de :**

Annie L. COT, Professeur émérite  
Université Paris I Panthéon Sorbonne

**Membres du Jury**

Duncan FOLEY, Professeur émérite  
New School for Social Research  
*Rapporteur*

Alan KIRMAN, Professeur émérite et directeur d'études  
Aix-Marseille University, École des Hautes Études en Sciences Sociales  
*Rapporteur*

Antoine MANDEL, Professeur des Universités  
Université Paris I Panthéon Sorbonne

Ellis SCHARFENAKER, Maître de conférences  
University of Utah

*L'université Paris 1 Panthéon-Sorbonne n'entend donner aucune approbation, ni improbation aux opinions émises dans cette thèse; elles doivent être considérées comme propres à leur auteur.*

# Preface

In the 19th century, physicists interested in studying the properties of matter faced the formidable challenge of thinking about how to predict the behavior of large collections of atoms moving through space and haphazardly bouncing off each other in a closed container. The collections were so large, on the order of  $10^{23}$ , that it seemed hopeless to try to track them individually by using Newton's laws of motion. James Clerk Maxwell, inspired by the growing number of statistical regularities that were being found in social data, notably by figures like Laplace, John Graunt, and Edmund Halley, proposed to tackle the challenge by taking censuses of atom populations. Starting out with the hypothesis that atom properties were gaussian, he soon realized, along with Ludwig Boltzmann, that exponential distributions were better suited for studying the bulk-scale properties of matter from a 'social-statistical' perspective. This led to the development of statistical mechanics, a field which has seen a recent upsurge of interest in the social sciences.

In many ways, the use of statistical physics ideas in the social sciences is a story about returning home, and as with any repatriation, it can be complex to navigate. The current interest in applying statistical physics concepts and models in the social sciences comes as a result of three main developments. The first is the increase in computational power and the amount of data available. The second is the realization that social institutions and collections of individuals interactively self-organize into complex structures, which do not seem to be easily graspable by control-theoretic or mechanistic perspectives. The third is the rise of interdisciplinary collaborations between economists, physicists, sociologists, biologists, and information theorists, amongst others, under the umbrella of complex systems science.

Independently of econophysicists and other social physicists, economists working in the classical political economy tradition have been exploring different ways of theorizing markets and economic institutions through the lens of statistical mechanics

frameworks for the past thirty years. Duncan Foley's 1994 paper *A statistical equilibrium theory of markets* was the first attempt in economics to theorize decentralized exchange and market transactions in terms of statistical equilibrium distributions. In a 2017 paper, Ellis Scharfenaker and Duncan Foley proposed an exponential probability distribution to study decentralized economic environments and market feedback processes. This is the quantal response statistical equilibrium (QRSE) distribution, which is used and studied in detail throughout this dissertation. The QRSE distribution is as much a theory of decentralized economic interaction as it is a descriptive probability model, similar to the well-known Gaussian and Laplace distributions, but also flexible enough to accommodate significant skew in the data.

This dissertation interweaves two distinct but interdependent projects. The first is the application of the statistical equilibrium framework for studying local public goods data. It constitutes a first step in a longer term project aimed at building a statistical mechanics picture of the complex and evolving process of local government fragmentation across the spectrum of urban locales in the US. The second is a philosophical exploration of the nature of principle-based inference in economic statistical equilibrium models. The results of this philosophical analysis, decidedly structured from the practitioner's point of view and reflective of the procedural knowledge involved in estimating statistical models, pave the way for future research at the intersection of semantic theories of information, statistical mechanics and political economy.

The dissertation is structured as follows. **Chapter 1** provides an overview of the general themes of complex systems thinking in economics, complemented by relevant historical and philosophical perspectives. The goal of this chapter is to discuss, within a broad complex systems science view, some of the major implications that come about from the use of information-theoretic methods in economics. **Chapter 2** provides a detailed and pedagogical introduction to maximum entropy methods and social statistical mechanics. It discusses historical, philosophical, and formal aspects. Presented in essay format, these two chapters serve as a general introduction to complexity and maximum entropy methods for this dissertation, and they are also the synthesis of my current understanding of these topics.

**Chapter 3** studies the distribution of school district expenditures in the US public education market using a statistical equilibrium framework. I implement the QRSE model in order to provide a novel empirical treatment of the Tiebout

hypothesis, and to examine the role of inter-jurisdictional competition and household choice in shaping observed expenditure patterns. I find that educational returns and expenditures in school districts across the US exhibit distinctively peaked, positively skewed distributions. I fit the QRSE model to the pooled sample for the 2000-2016 period and study the role that competition and choice scale parameters play in shaping the equilibrium outcome. Bayesian inference and Markov Chain Monte Carlo (MCMC) sampling are used to recover posteriors for the unknown parameters in the model. The sample used covers all public school districts in the US for the 2000-2016 period.

In **Chapter 4**, I examine the distinct role played by Tiebout and sorting incentives in shaping the observed distributions of educational expenditures. The distinctively peaked and positively skewed distributions of school district expenditures are explained in terms of the complex interaction between jurisdictional competition and households' drive towards stratification on the basis of income and property values.

**Chapter 5** studies the structure of modal and principle-based inference in Max-Ent probability models. It builds on Marc Lange's recent account of 'non-causal' explanations in the philosophy of physics. I propose a distinct set of features that can be used to identify principle-based inference and discuss its applicability to existing statistical research in both physics and political economy.

# Contents

<b>List of Figures</b>	<b>x</b>
------------------------	----------

<b>List of Tables</b>	<b>xv</b>
-----------------------	-----------

<b>1 The complexity view in economics</b>	<b>1</b>
1.1 General complexity themes . . . . .	1
1.2 Economic complexity . . . . .	6
1.2.1 Themes in economic complexity . . . . .	6
1.2.2 Econophysics . . . . .	6
1.2.3 Maximum entropy methods and the distribution of profit rates	7
1.2.4 Quantal response statistical equilibrium . . . . .	9
1.3 A note on econometrics . . . . .	9
1.4 Looking ahead: aggregate dynamics or information ? . . . . .	12
1.5 The cognitive-historical question . . . . .	13
1.6 Semantic issues in computational complexity . . . . .	17
1.6.1 Methods of approximation and partial information . . . . .	17
1.6.2 Discrete and computational representations . . . . .	18
1.7 Pooh, the tortoise and the hare . . . . .	21
<b>2 Entropy, inference, and Boltzmann distributions</b>	<b>25</b>
2.1 Probability distributions . . . . .	25

2.2	Complex systems, statistical physics and maxent . . . . .	27
2.2.1	Bird flocks, criticality and statistics . . . . .	29
2.2.2	Why birds ? . . . . .	34
2.2.3	Animal models and analogical transfer . . . . .	36
2.2.4	Disciplinary tensions . . . . .	37
2.3	Maximum entropy : a variational principle . . . . .	39
2.3.1	What is information entropy ? . . . . .	39
2.3.2	Dies and uncertainty . . . . .	40
2.3.3	Kangaroos and logical consistency . . . . .	41
2.3.4	Monkeys and multiplicity . . . . .	43
2.3.5	Probability boards . . . . .	45
2.3.6	Critical points . . . . .	47
2.4	Boltzmann densities and microeconomic energy functionals . . . . .	56
2.4.1	An integrated perspective for economic statistical mechanics . . . . .	56
2.4.2	Approaching the optimization procedure . . . . .	57
2.4.3	Principle-based search . . . . .	60
2.4.4	Mining for energy . . . . .	62
<b>3</b>	<b>A Statistical Equilibrium Model of Public School Expenditures</b>	<b>65</b>
3.1	Introduction . . . . .	65
3.1.1	Tiebout Competition . . . . .	65
3.1.2	Unintended Outcomes, Market Efficiency and Tiebout Sorting . . . . .	69
3.1.3	Sample and Paper Structure . . . . .	70
3.2	School District Variables . . . . .	70
3.2.1	Data . . . . .	70
3.2.2	Empirical Distributions of $x$ . . . . .	71
3.3	Free Competition and The Tiebout Hypothesis . . . . .	73



3.3.1	Smith's Theory of Competition . . . . .	73
3.3.2	The Tiebout Hypothesis . . . . .	74
3.4	Local Public School Expenditures and Household Choice . . . . .	75
3.4.1	Statistical Equilibrium Modeling and Maximum Entropy Inference . . . . .	75
3.4.2	A Logit Quantal Response Function for Household Choice .	77
3.4.3	The Competitive Feedback Constraint . . . . .	80
3.4.4	QRSE Maximum Entropy Program and Density . . . . .	82
3.5	Bayesian Estimation of QRSE Model . . . . .	84
3.5.1	MAPs and Distance Measures . . . . .	84
3.5.2	Model Specification and Markov-Chain Monte Carlo Sampling	85
3.5.3	Results . . . . .	89
3.5.4	Fits . . . . .	91
3.6	Discussion . . . . .	93
3.7	Conclusion . . . . .	96
<b>4</b>	<b>Tiebout statistical equilibrium : sorting and inter-jurisdictional compe- tition across urban locales and income groups</b>	<b>97</b>
4.1	Introduction . . . . .	97
4.2	Tiebout statistical equilibrium . . . . .	99
4.2.1	The educational returns random variable . . . . .	99
4.2.2	Maximum Entropy Inference . . . . .	100
4.2.3	T-QRSE . . . . .	101
4.3	Quantal Response & Competitive Feedback . . . . .	105
4.4	School District Data . . . . .	109
4.4.1	Census and NCES Data . . . . .	109
4.4.2	Subsamples . . . . .	110

4.4.3	Urban Locales . . . . .	110
4.4.4	Income Groups for the period 2013-2016 . . . . .	112
4.5	Bayesian Estimation . . . . .	114
4.5.1	Model Specification and Markov-Chain Monte Carlo Sampling	114
4.6	School District Ensembles . . . . .	116
4.7	Results . . . . .	119
4.7.1	Urban Locales . . . . .	119
4.7.2	Income Groups . . . . .	122
4.8	Discussion . . . . .	124
4.8.1	Tiebout sorting and competition across urban locales . . . .	124
4.8.2	Tiebout sorting and competition across income groups . . .	128
4.8.3	Median property values and $\alpha$ . . . . .	130
4.9	Conclusion . . . . .	131
<b>5</b>	<b>Constructive statistical explanations: testing causal and non-causal principles in economic data with MaxEnt</b>	<b>132</b>
5.1	Introduction . . . . .	132
5.2	Statistical Explanation . . . . .	133
5.2.1	The Inductive-Statistical Schema . . . . .	133
5.2.2	Amending The IS Schema . . . . .	134
5.2.3	Statistical Explanation as Subsumption Under Generality . .	136
5.2.4	Probability Models . . . . .	137
5.2.5	The Probabilistic Basis of Statistical Explanation . . . . .	138
5.3	Constructive Statistical Explanations . . . . .	141
5.3.1	Key Properties . . . . .	141
5.4	Prediction, Causality and Modality . . . . .	143
5.4.1	Prediction . . . . .	143

---

5.4.2	Causal and Non-Causal Explanations . . . . .	146
5.5	Distributional Supposition and Statistical Autonomy . . . . .	154
5.5.1	Distributional Supposition . . . . .	154
5.5.2	Statistical Autonomy . . . . .	156
5.6	Constructive Statistical Explanations in Political Economy . . . . .	158
5.6.1	Explaining Profit Rates . . . . .	159
5.7	Conclusion . . . . .	167
	<b>References</b>	<b>169</b>
	<b>Appendix A Résumé de thèse</b>	<b>182</b>

# List of Figures

1.1	Emergence of communities and structure in weighted networks. See the model originally proposed by Kumpula et al. (2007), which studies the relationship between network topology and interaction strengths. Image taken from Dirk Brockmann's 'Complexity Explorables' (CC BY-NC-ND 4.0). For Brockmann's interactive version of this list of examples see: <a href="https://complexityexplained.github.io/">https://complexityexplained.github.io/</a> .	2
1.2	Bird flocking is a common example of emergent collective behavior. Individual birds interact via local and inter-connected neighborhoods, which are used to propagate directional cues and achieve global coordination. . . . .	3
1.3	The double pendulum is a classic example of chaotic motion in dynamical systems. Unlike single pendular systems (B), double pendulums (B) are highly sensitive to initial conditions. Two double pendulums suspended from the same initial conditions will quickly diverge in their trajectories. . . . .	4
1.4	The English botanist Robert Brown (1827) observed how pollen seeds suspended in water swarmed following irregular and motions. The complex many-body system which produces brownian motion cannot be accounted for by looking at the individual equations of motion. When faced with complex many-body problems, physicists take a census of the population's energy and reason in terms of statistics. Einstein inferred the existence of atoms through his statistical molecular theory of liquids (1905). . . . .	5
1.5	Winnie-the-Pooh: The Wrong Sort of Bees, by A.A. Milne . . . . .	22

1.6	Zeno's paradox. The tortoise always remains slightly ahead of the hare, no matter how many times it tries to catch up. . . . .	23
1.7	Discrete approximation. . . . .	24
2.1	On the left, the Laplace distribution: $f_x = \frac{k}{2}e^{-k x }$ . On the right, the Normal distribution: $f_x = \frac{1}{\sqrt{2\pi}\sigma}e^{-x^2/2\sigma^2}$ . Reproduced from Kotz and Kozubowski (2001) . . . . .	26
2.2	(A) Bird flocking event. (B) Velocity vectors from snapshot. Graph from Bialek et al. (2012) . . . . .	30
2.3	Maxent assignment for a six-faced die with an observed average of 4.5. Reproduced from Sivia and Skilling (2006) . . . . .	41
2.4	(a) The probabilities for the four possibilities: $p_1, p_2, p_3, p_4$ ; (b) parameterization in terms of a single variable $x$ , where $x = p_1$ . Reproduced from Sivia and Skilling (2006). . . . .	42
2.5	Different candidates for the variational principle. See Gull and Skilling (1984). Table from Sivia and Skilling (2006) . . . . .	43
2.6	Maxent monkeys . . . . .	44
2.7	The Galton board. . . . .	45
2.8	A (biased) maxent board. . . . .	46
2.9	Heavy-tailed distributions. Exponential tails go to zero faster than heavy tails. . . . .	52
2.10	(A) Some weakly chaotic maxent boards. (B) The job market for entropic monkeys. Which $q$ ? . . . . .	54
3.1	Marginal distribution of $x$ (in thousands) for the period 2000-2016. Histogram and Box-Plot. . . . .	72
3.2	Stacked histogram plots of the educational returns variable $x$ (in thousands). 2000-2016. The stacked histograms reveal the time invariance of the statistical equilibrium distributions for the different years that make up the pooled sample. . . . .	73

3.3	Quantal responses functions for probability of household entry/exit. At near zero $T$ the quantal response function takes resembles the heaviside step function. Under a zero entropy assumption ( $T = 0$ ), we would expect households to move into districts exclusively in the case where $x - \mu > 0$ . This graph is illustrative (taking $T=0.1$ and $T = 3$ as examples). . . . .	80
3.4	Trace plots for MCMC samples obtained using the Metropolis-Hastings algorithm. 3 chains, 30,000 iterations per chain and 4,000 tuning samples. All US School Districts. 2000-2016. . . . .	87
3.5	Posterior pair plots for parameters $T$ , $S$ , $\mu$ and $\alpha$ . . . . .	88
3.6	Posterior distributions for $T$ , $S$ , $\mu$ , and $\alpha$ . All US School Districts. 2000-2016. . . . .	90
3.7	Marginal distribution of $x$ , and yearly line plot for $x$ and $\kappa$ for the period 2000-2016. The histogram excludes the upper 0.01 quantile (for visualization). Plotted in thousands. All US School Districts. 2000-2016. . . . .	91
3.8	Line histogram of observed distribution for $x$ (educational returns). Overlaid is the fitted predictive marginal density $\hat{f}_x$ . We excluded the upper 0.01 quantile (for visualization). The Soofi ID/performance fit measure is shown. All US School Districts. 2000-2016. . . . .	92
3.9	Left: Predictive entry and exit densities $f_{a,x}$ . Right: Household Quantal Response Functions $f_{a x}$ . The estimated ‘tipping point’ $\hat{\mu}$ is plotted with a dashed-dotted line. . . . .	92
3.10	Variations to individual parameter, holding all others constant. The baseline setting is $T = 5$ , $S=5$ , and $\mu = \alpha = 0$ . . . . .	94
4.1	Variations to individual parameter, holding all others constant. The baseline setting is $T = 5$ , $S=5$ , and $\mu = \alpha = 0$ . . . . .	103
4.2	Left: Plot of the difference in quantal response probabilities, with different values of $T$ . Right: A symmetric competitive feedback process centered at $\mu = \alpha$ . . . . .	106

4.3	Conceptual illustration of probabilistic competitive feedback. Departures from $\mu$ break symmetry in the process and create positive skew in the outcome variable. . . . .	106
4.4	Joint action-outcome and marginal densities with balanced behavioral/market scales and $\alpha > \mu$ . . . . .	107
4.5	Stacked histogram plots of the educational returns variable $x$ . Urban Locales: Town, Rural, City, Suburb. 2000-2016 . . . . .	111
4.6	Tripartite division of the median household income distribution for school districts. 2013-2016. . . . .	112
4.7	Stacked histogram plots of the educational returns variable $x$ . Income Groups: Low, Medium, High. 2013-2016. The stacked histograms suggest a marked difference in the statistical equilibrium distribution for the three income groups. . . . .	113
4.8	QRSE model structure and priors used for Bayesian Estimation. . .	115
4.9	Stacked histogram of grouped yearly subsamples across the four urban locales. 2000-2016. . . . .	117
4.10	Stacked histogram of urban locale subsamples and the full ensemble of school districts for the 2000-2016 period. . . . .	118
4.11	Forest Plot with 94 % HDI Intervals. The middle bar plots the interquartile range for the posterior (0.25-0.75) and the median (central point). Urban Locales: Town, Rural, City, Suburb. 2000-2016. . . .	120
4.12	Line plots for the estimated posterior means across four year groupings. Town and Rural, 2000-2016. . . . .	120
4.13	Line plots for the estimated posterior means across four year groupings. City and Suburb, 2000-2016. . . . .	121
4.14	Forest Plot with 94 % HDI Intervals. The middle bar plots the interquartile range for the posterior (0.25-0.75) and the median (central point). Income Groups: Low, Middle, High. 2013-2016. . . . .	123
4.15	Estimated quantal response and competitive feedback. Mean posterior values are used for plotting. Town and Rural, 2000-2016. . . .	125

4.16	Estimated quantal response and competitive feedback. Mean posterior values are used for plotting. City and Suburb, 2000-2016. . . .	126
4.17	Histogram of observed distributions for $x$ (educational returns). Overlaid are the fitted marginal $f_x$ , entry $f_{e,x}$ , and exit $f_{s,x}$ distributions. Urban Locales: Town, Rural, City, Suburb. 2000-2016. . . . .	127
4.18	Estimated household quantal response and competitive feedback. Mean posterior values are used for plotting. Low and Middle income groups, 2013-2016. . . . .	128
4.19	Estimated household quantal response and competitive feedback. Mean posterior values are used for plotting. High income group, 2013-2016. . . . .	129
4.20	Histogram of observed distributions for $x$ (educational returns). Overlaid are the fitted marginal $f_x$ , entry $f_{e,x}$ , and exit $f_{s,x}$ distributions. Income Groups: Low, Middle, High. 2013-2016. . . . .	129
4.21	Distribution of median property values for all districts; 2013-2016. Plotted against the income and urban locale categories. . . . .	130
4.22	Visual summary of the estimated posterior means for $\alpha$ . All locales (2000-2016) and income groups (2013-2016). . . . .	130
5.1	Deductive-Nomological (DN) Explanation Schema . . . . .	134
5.2	Inductive-Statistical (IS) Explanation Schema . . . . .	134
5.3	Key Properties of CSEs . . . . .	141
5.4	From (Feynman, 1965) . . . . .	150
5.5	From (Dragulescu and Yakovenko, 2000a) . . . . .	150
5.6	From From (Feynman, 1965) . . . . .	152
5.7	Laplace Profit Rates . . . . .	159
5.8	Densities Chart . . . . .	163
5.9	Fits for the different theories proposed. . . . .	164



# List of Tables

3.1	Summary statistics of estimated parameters $T$ , $S$ , $\mu$ , and $\alpha$ . The means, standard deviations, 94% credible intervals and the convergence statistics $\hat{R}$ from the MCMC samples are reported. All US School Districts, 2000-2016. . . . .	89
3.2	Model variables (in thousands). All US School Districts, 2000-2016.	91
4.1	Summary statistics of estimated parameters $T$ , $\mu$ , $S$ , and $\alpha$ . The mean, standard deviation, 94% credible interval and the convergence statistics $\hat{R}$ from the MCMC samples are reported. Urban Locales: Town, Rural, City, Suburb. 2000-2016. . . . .	119
4.2	Summary statistics of estimated parameters $T$ , $S$ , $\mu$ and $\alpha$ . The mean, standard deviation, 94% credible interval and the convergence statistics $\hat{R}$ from the MCMC samples are reported. Income Groups: Low, Middle, High. 2013-2016. . . . .	122

# Chapter 1

## The complexity view in economics

### 1.1 General complexity themes

Complex systems science studies how large collections of components interacting locally at various scales self-organize into distinct global patterns and behaviors. At the core of complexity science lies the idea that wholes are irreducible to their parts, and that the analysis and prediction of complex whole-part dynamics in social, biological and physical systems requires novel inter- and trans-disciplinary methods.

The first major theme that defines the complexity view is the study of *interactions*. The components of complex systems can be seen to interact in a myriad of ways with each other, and with their environment. Complexity science resists the idea that such interactions should be studied through compartmentalized and isolated reductionist frameworks. It is not just a question of eliciting the networked and hierarchical relationships that exist between a system's components, but also of understanding how patterns of strong and weak interdependence lead to distinct statistical regularities and signatures.

A second major theme is *emergence*. The non-trivial global patterns that emerge out of networked local interactions impose barriers to traditional linear and statistical thinking. Air and vapor molecules combine to form tornadoes, billions of neurons in the brain interact to produce cognition, and global but brittle networks of banking institutions and balance sheets inadvertently tip polities into financial collapse and political crisis. The study of emergence in complex systems can be approached through dynamical, statistical and computational counterparts.

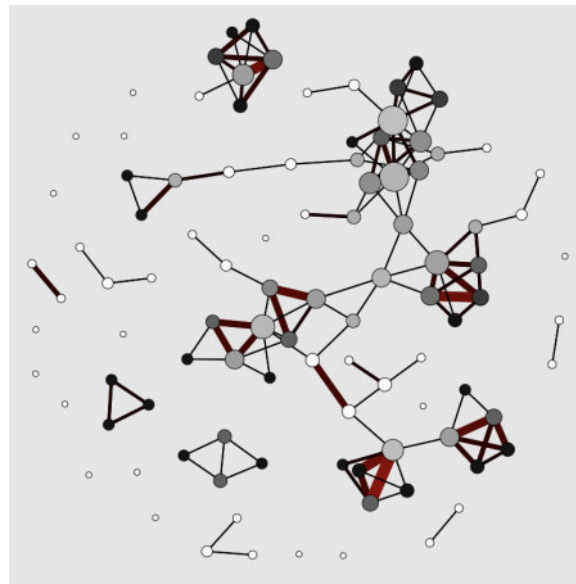


Fig. 1.1 Emergence of communities and structure in weighted networks. See the model originally proposed by Kumpula et al. (2007), which studies the relationship between network topology and interaction strengths. Image taken from Dirk Brockmann's 'Complexity Explorables' (CC BY-NC-ND 4.0). For Brockmann's interactive version of this list of examples see: <https://complexityexplained.github.io/>.

The problem of analyzing the emerging and time-dependent behavior of complex systems having more than two degrees of freedom leads to the third major theme, which is that of *dynamical systems* analysis and *chaos*. The interaction of independent linear-predictable components leads to wholes exhibiting non-linear and chaotic trajectories in a system's phase space. There are two main types of dynamical systems: *differential equations* and *iterated maps* (or difference equations). Nonlinear systems with more than two degrees of freedom are often impossible to solve analytically, due to the fact that the principle of superposition fails majorly in these contexts. This is another way of stating that in nonlinear systems the whole is precisely not equal to the sum of its parts. For systems exhibiting chaotic behavior, small errors in the measurement of initial conditions get amplified over time and the systems become effectively unpredictable. Like stable dynamical systems, chaotic systems tend to attractors, but these are often found to be complex and difficult topological spaces (Albin and Foley, 1998; Strogatz, 2018). Common examples of nonlinear and chaotic systems are the double pendulum, the weather and financial markets.

Complexity science points to the limits of analyzing such systems in terms of reduced form dynamical-law approaches. The barriers that such systems pose to



Fig. 1.2 Bird flocking is a common example of emergent collective behavior. Individual birds interact via local and inter-connected neighborhoods, which are used to propagate directional cues and achieve global coordination.

dynamicists have spawned deep methodological reflections, and have led naturally to the development of geometric/topological, computational and statistical alternatives.

A fourth major complexity theme is *self-organization*. The qualitative study of nonlinear dynamics and interaction opens up new intuitions about the processes that drive stable pattern formation in complex systems. The term ‘self-organization’ alludes to the notion that such patterns emerge without the intervention of central or external controllers (De Domenico et al., 2019); these are collective but non-teleological phenomena. The theme of self-organization provides a convenient conceptual basis for analyzing functional structures, such as the formation of crystalline patterns in matter, the evolving morphologies of living organisms (Rajapakse and Smale, 2017; Turing, 1952), swarming, and socioeconomic segregation (Schelling, 2006), to give a few examples. The dynamics of ‘self-organizing’ processes often lead to unstable equilibria and ‘critical states’ that put systems on ‘the verge of chaos’. The aggregate properties of self-organized systems are usually well characterized by fractals, and exponential or heavy-tailed distributions.

The fifth major theme is *adaptation*. Complex systems actively respond to their environment. "Living organisms display behavior that is qualitatively different from chaotic systems like geological or meteorological interactions. They are highly structured, and have powerful homeostatic mechanisms that stabilize important

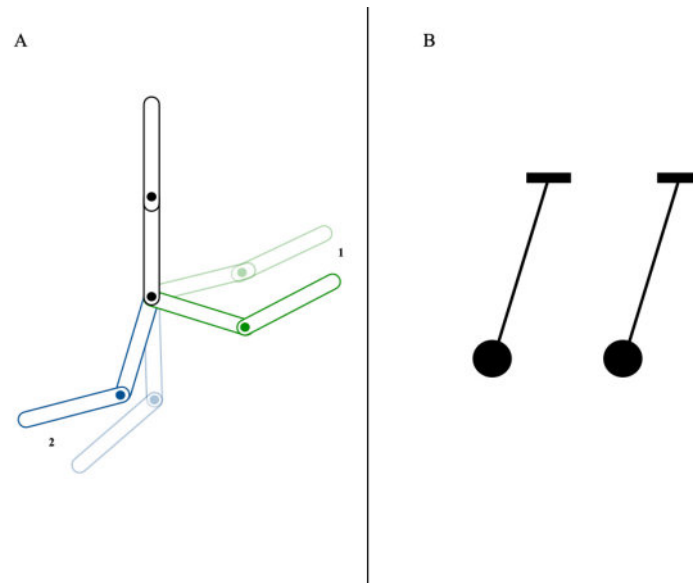


Fig. 1.3 The double pendulum is a classic example of chaotic motion in dynamical systems. Unlike single pendular systems (B), double pendulums (B) are highly sensitive to initial conditions. Two double pendulums suspended from the same initial conditions will quickly diverge in their trajectories.

aspects of their behavior ... Human beings, for example, seek solutions to problems, which may involve an exploration of the state space which does not have the repetitive features of chaotic motions. In solving a problem, a human being may pursue one approach, exhibiting one type of behavior, for a while, and then determine that this approach is a dead end, and suddenly (from the point of view of an external observer) shift to a qualitatively different type of behavior embodying an alternative approach to the problem" (Albin and Foley, 1998). Adaptation explains "the difference between a ball that rolls to the bottom of a hill and stops and a bird that adapts to wind currents while flying" (De Domenico et al., 2019). Thus, the theme of *adaptation* in complexity sciences studies processes of learning and psychological development, the formation of social ties and cooperative behavior, as well as genetic evolution and natural selection. Adaptation leads to the robustness and resilience attributes that we commonly associate with living organisms and social formations. This concept lies at the basis of the notion of *complex adaptive systems* in biology and the social sciences.

The sixth and seventh major themes are *computational complexity* and *interdisciplinarity*. Complexity science seeks for general and universal principles through cross-disciplinary mathematical and computational approaches. Schinckus (2021)

makes the important point that this cross-disciplinary perspective can be historically understood a response to the increasing balkanization of knowledge and scientific research. I will elaborate on these two themes with greater detail in the following sections.

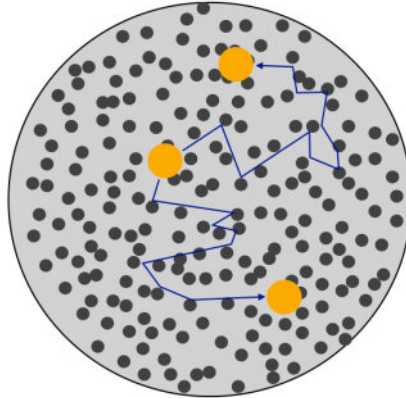


Fig. 1.4 The English botanist Robert Brown (1827) observed how pollen seeds suspended in water swarmed following irregular and motions. The complex many-body system which produces brownian motion cannot be accounted for by looking at the individual equations of motion. When faced with complex many-body problems, physicists take a census of the population's energy and reason in terms of statistics. Einstein inferred the existence of atoms through his statistical molecular theory of liquids (1905).

## 1.2 Economic complexity

### 1.2.1 Themes in economic complexity

In September 1987, twenty people came together at the Santa Fe Institute to talk about "the economy as an evolving, complex system." Ten were theoretical economists, invited by Kenneth J. Arrow, and ten were physicists, biologists, and computer scientists, invited by Philip W. Anderson. The meeting was motivated by the hope that new ideas bubbling in the natural sciences, loosely tied together under the rubric of "the sciences of complexity," might stimulate new ways of thinking about economic problems. For ten days, economists and natural scientists took turns talking about their respective worlds and methodologies. While physicists grappled with general equilibrium analysis and noncooperative game theory, economists tried to make sense of spin glass models, Boolean networks, and genetic algorithms.

(Arthur, 1997)

The three major themes that emerged out of SFI's interdisciplinary exploration of economic questions were set forth in the 1988 volume *The economy as an evolving complex system* (Anderson, 1988). David Pines notes that SFI originally established three working groups on the topics of 'Cycles', 'Webs', and 'Patterns' (see Schinckus, 2021). 'Cycles' focused on nonlinear deterministic behavior and chaos, 'Webs' focused on interaction and emergence, particularly as viewed through the lens of network science and agent-based modeling. And 'Patterns' focused on studying the statistical properties and invariants of complex economic systems. Schinckus (2021) notes that one finds a decreasing interest in *deterministic* chaos throughout the evolution of the SFI's economics research program.

### 1.2.2 Econophysics

Propelled by the vision that market-driven processes are complex, high-dimensional, dynamic and stochastic systems, econophysics began to develop in the early 1990s as a powerful alternative for conducting empirical work. The heavy influx of 'quants' and physicists into asset management firms also supported this development, along

with an increasing demand for renewed interdisciplinary research at the intersection of physics, biology, game theory and economics. The Santa Fe Institute continues to play a key role in supporting econophysics related research.

Roughly defined, econophysics is concerned with the application of "modeling and computational methods coming out of statistical physics to the empirical study of economic and financial processes" (Schinckus, 2016). As discussed by Schinckus, one can broadly identify two main traditions in the econophysics literature. The first is what Schinckus calls the "purely statistical" or "stylized facts approach", concerned with the description of the distinct statistical signatures that characterize emergent aggregate patterns in economic environments (e.g. power law, exponential and heavy-tailed distributions). This work has turned out to be immensely important, in that it began to cast light on the inadequacy of using Gaussian frameworks for modeling sources of variation around central tendencies in economic variables.

The second approach Schinckus terms "agent-based econophysics". This bottom-up simulation approach starts by specifying a theoretically or intuitively determined microeconomic environment that is then simulated to obtain emergent macro patterns and stationary distributions. Agent-based econophysics emphasizes the explicit modeling of interactive phenomena such as information diffusion in economic networks and dynamic exchange protocols with learning agents, to give a few examples (Chakrabarti et al., 2006). An important aspect of these models is that they are able to encode structural, recursive and temporal features all at once.

### **1.2.3 Maximum entropy methods and the distribution of profit rates**

Parallel to the developments in the econophysics literature highlighted in the previous section, which for the most part were carried out by physicists, a growing body of empirical research has emerged in the past thirty years at the intersection of heterodox political economy, non-walrasian economics, information theory and statistical mechanics. This literature sets itself apart in that it is not a tabula rasa attempt to apply statistical physics ideas to the study of economic processes. This line of work attempts to recover and put to empirical scrutiny, using the frameworks



of information-theoretic statistical mechanics, well-established hypotheses from the stock of the classical<sup>1</sup> and non-walrasian traditions (Scharfenaker, 2020a).

A criticism that this literature offers regarding both simulation-based and statistical approaches, arising out of econophysics, is the loose application of symmetry and ergodicity concepts. For example, as when Yakovenko treats money and financial flows as a conserved form of energy in order to derive the Pareto-Exponential laws for US income distributions (Dragulescu and Yakovenko, 2000b).

A major catalyst for the development of this line of reasoning came out of the work of Farjoun and Machover (1983). In Farjoun and Machover (1983), the authors critique the deterministic treatment of economic aggregates in classical political economy and propose a probabilistic framework capable of capturing the presence of chance fluctuations around a central tendency in the profit rate. Their research gave new (statistical) life to the Smithian and Marxian hypotheses of the long-run equalization of the profit rate, and to the classical idea that observed rates gravitate around a "natural quantity". Their theoretical considerations led them to conjecture a gamma density for the profit rate stationary distribution, but this turned out to be unconvincing since the gamma distribution is constrained to the positive domain and negative profit rates are far from being uncommon in modern capitalist economies (Scharfenaker and Semieniuk, 2017).

This led others to propose new statistical descriptions of the firm profit rate distribution. Alfarano et al. (2012a), using US data for the 1980-2006 period, propose a Subbotin (exponential-power) distribution along with a drift-diffusion process to describe the underlying firm competitive dynamics. Scharfenaker and Semieniuk (2017), building on the work of Farjoun and Machover (1983), examined profit rate data for US firms in the 1962-2014 period and found that the statistical equilibrium distribution organizes into an asymmetric Laplace distribution, a special case of the asymmetric Subbotin distribution (Yang, 2018). While Alfarano et al. (2012a) and Scharfenaker and Semieniuk (2017) agree that the statistical dispersion law for the profit rate follows a mean absolute deviation<sup>2</sup> and not a squared mean deviation<sup>3</sup>, the two approaches differ in their interpretations. Scharfenaker and Semieniuk (2017) derive the statistical equilibrium distribution via maximum entropy inference but don't provide a macro-dynamic characterization for the competitive process. They

---

<sup>1</sup>Classical political economy

<sup>2</sup> $|y - \mu|^\alpha$

<sup>3</sup> $(y - \mu)^2$

see the observed patterns as supporting evidence for the classical hypothesis of the long-run equalization of profits and the Smithian theory of competition.

### 1.2.4 Quantal response statistical equilibrium

In subsequent work, Scharfenaker and Foley (2017) propose a principle-based derivation of the observed profit rate distribution. This is the quantal response statistical equilibrium model (QRSE) which is used in this dissertation. The QRSE is a probabilistic model that aims to capture the impact of an underlying decision-theoretic context on the emergent properties of the competitive process. The model is general enough to be applied to the study of a broad set of decentralized economic environments driven by feedback and entry/exit dynamics. Its behavioral component takes the form of a quantal choice function that captures trade-offs between utility and entropic uncertainty.

An important element to note about the QRSE is that it's a theoretical model, but it's also a probability distribution. If you look at it independently of the theory that underpins its derivation, it's a Laplace-Normal type density that can be used to describe highly peaked and skewed data. So it has both theoretical and descriptive components. I treat the question of principle-based inference in more detail in chapters 2 and 5.

## 1.3 A note on econometrics

Empirical research in economics is tasked with the difficult problem of providing explanations and statistical characterizations of observed outcomes in complex and constantly evolving socio-economic environments which have a large number of degrees of freedom.

In econometrics, we find a wide array of approaches. Some approaches attempt to solve inverse causal inference problems, in which researchers attempt to identify the underlying causal structure of the data generating process from observed patterns, and others, particularly in applied microeconometrics, are more concerned with studying forward causal questions, in which one studies the "effects of causes" as opposed to the "causes of effects" (Gelman and Imbens, 2013).

Contemporary microeconometrics is predominantly concerned with solving forward causal inference problems (Imbens and Wooldridge, 2009). Some of the more notable approaches that have been used in impact evaluation and policy analysis are *randomized control trials*, *propensity score matching*, *double-difference* or *difference-in-difference* techniques, *instrumental variable* regression, *quantile regression*, as well as *structural* and *reduced-form* approaches to a lesser extent (Khandker et al., 2009). Can conditional cash-transfer programs to poor inner-city households improve educational outcomes for a well-defined cohort of students? Does new infrastructure raise welfare in a particular city? These are the sort of questions being answered by this literature, often complemented by qualitative approaches that help ground a better understanding of the communities or cohorts studied. This type of research is closer to the type of empirical and qualitative work that one finds in social psychology or in the medical sciences. The main difference being that its conclusions are largely localized to specific cohorts and as such may be of limited external validity and replicability.

In relation to the spectrum that goes from ‘reverse-causal’ to ‘forward-causal’ approaches, empirical macroeconomic research and applied general equilibrium (AGE) analysis has historically held a scientifically anomalous position. The history of the relationship between macroeconomic analysis and probability is complex, and is full of ironies and pragmatic contradictions. As Morgan (1990) points out, the so called ‘probabilistic revolution’ did not take place until the appearance of Trygve Haavelmo’s *The Probability Approach in Econometrics* (1999). Up until then, early econometricians made use of statistical methods but rejected altogether a direct application of the mathematical theory of probability; a blatant paradox, given that probability theory is the very basis upon which statistical inference and phenomenological investigations in the sciences are built.

Trading notions of causality for notions of ‘identification’, ‘exogeneity’ and ‘endogeneity’, the macroeconometric approach that emerged out of the Cowles commission in the 1950s favored the study of the invariant properties of simultaneous equation models (Hoover, 2008). It’s important to note that the Cowles econometricians were well aware of the inherent complexity of the phenomenon they were trying to tackle. Economic quantities are simultaneously determined and they are embedded in a complex network of causal relations for which ordinary least squares regressions can only provide biased estimates. In *econometrese*, this came to be

understood as the problem of ‘identifying the causal directionality of regression estimates’ (Hoover, 2005).

In retrospect, the **key historical question** that emerges is the following: how have economists historically responded to the challenges of modeling and making sense of social complexity ?

In macroeconometrics, the focus on the challenges of ‘identification’ led to convoluted statistical techniques for the estimation of parameters in simultaneous equation models, such as indirect least square methods, limited-information maximum likelihood, and full-information maximum likelihood (see Anderson and Rubin, 1949; Hood et al., 1953; Koopmans et al., 1950). Another important set of modeling approaches emerged in the 1950s which emphasized a recursive causal ordering in the structural determination of macro variables (Simon, 1977; Wold, 1954).

These approaches dominated the field until the 1970s, when Lucas and Sargent argued against the alleged invariance of the relationships estimated in SEM models. Their critique was primarily waged against Keynesian macroeconomic models with adaptive expectations. The reappraisal of dynamic concerns as being central to forecasting resurfaced old debates regarding the relationship between an unobservable microeconomic domain and observable macroeconomic aggregates (see Hoover, 2008; Lucas and Sargent, 1981). The line of reasoning proposed by real business cycle theory eventually coalesced around the development of Dynamic Stochastic General Equilibrium (DSGE) models.

In what amounts to an odd methodological ploy, the introduction of DSGE modeling transferred the analysis of macroeconomic behavior from the realm of SEM models with adaptive expectations to the realm of stochastic optimal control theory with rational expectations. Policy interventions in the economy could be better studied, so they have argued, by solving for the optimal path of a bizarre functional in a predictable dynamic programming environment<sup>4</sup> (see Kydland and Prescott, 1982; Rotemberg and Woodford, 1997; Smets and Wouters, 2007).

---

<sup>4</sup>See Klein (2015) for a historical treatment of the introduction of optimal control theory and dynamic programming into the Muthian rational expectations program.

## 1.4 Looking ahead: aggregate dynamics or information ?

Looking back at some of the historical responses that economists have given to the task of modeling economic data, one sees that a common objective of the different approaches has been that of identifying the causal structure or the dynamic behavior of economic aggregates. The use of statistical physics ideas in the social sciences can make a significant contribution to meet that challenge, but it's also worth reflecting on how such tools might be opening up previously discarded doors for conceptualizing inference problems in economics. And by this I am not referring to methods transfer or applications, but to the order of knowledge we wish to obtain from economic data. The point of view defended in this dissertation is that information-theoretic and complexity methods invite us to reconsider the relevance of certain *dynamicist* and *aggregative* conceptions for making sense of the large streams of data currently available.

Gaining information from complexity economic models also requires being able to attribute semantic content to the patterned regularities we observe and sample. The question of 'principles' is really just a question of the meaning of information. As information-theoretic approaches continue to gain ground, it's worth reflecting on what the possible space of semantics looks like for statistical physics models in economics.

Two aspects of the complexity view must be considered in order to establish that perspective. First, it's important to bear in mind that the increase in computational power, and the fact that we now have the ability to ask machines questions about our datasets, is a condition and not an idiosyncratic feature of contemporary scientific thought. The success of the SFI approach to economic thinking has largely come as a result of the rise in computational power, the ability to simulate complex dynamics and social-structural formations, and to tackle high-dimensional inference problems on personal computers. Second, information-theory, social statistical mechanics and the 'complexity view' inevitably erode on the autonomy of the economics discipline. This can be framed in a positive light. Manfred Eigen, a German biophysical chemist, stated that "life is an interplay of energy, entropy, and information to be jointly studied by physicists, biologists, chemists, and information theorists" (Eigen, 2013). The SFI program has made significant strides in having economists join that

group over the past thirty years, and in doing so it has heralded an era of complexity economics (also recently with the help of INET).

A distinguishing feature of the complexity economics approach is that it shares a common language with other disciplines, in particular dynamical systems theory, information theory, and statistical physics. A less discussed feature is that complexity economists are mining for new formulations of social and economic energy. As they do so, the question of determining the scientific and politico-economic principles that support such energy formulations seems to be ‘up for grabs’. Attributing meaning to such models naturally leads to political and philosophical divergences, which is what makes approaching these methods so much more difficult (but also interesting) in the social sciences.

The next two sections address this question by pointing to how the space of possible semantics is a function of the cognitive landscapes afforded by complexity methods, but also of the regions of mathematics in which we choose to house social thought.

## 1.5 The cognitive-historical question

There are two ways of looking at the history of complexity thinking in economics. The first one, as suggested by Arthur’s quote above, is to think of the emergence of complexity economics as a transdisciplinary phenomenon that was facilitated by the Santa Fe Institute (SFI) in the late 1980s and early 1990s. There is certainly some truth to that view, but it has the problem that it suggests casting the history of economic thought into a ‘pre-complexity’ period and a proper ‘complexity’ period that starts in the 1990s with the SFI research agenda. While this historical take may be more congenial for explaining the recent rise of econophysics and agent-based modeling, it does not give full credit to the fact that complexity is a foundational concept for economic thinking, taking roots as far back as Adam Smith and the physiocrats, and permeating at the core of modern schools of economic thought, including general equilibrium theory, Keynesianism, and Austrian ‘catallactics’ (Hayek et al., 1948).

What is true, nonetheless, is that emergence of the "Santa Fe approach" has challenged and put a considerable strain on the idea that economic phenomena can

be suitably understood through a mono-disciplinary framework. Contemporary complexity economists often have to answer the question: "Is this economics"? Upon due reflection, one should be able to comfortably answer: "Yes and No". "Yes" because complexity economics sheds new light on market processes, public goods allocation problems, financial time series, economic networks, and political-economic questions. "No" because the resulting analyses and contributions often pertain as much to the realm of economic theory as they do to sociology, anthropology, computer science, sociobiology, probability and statistics.

There are several important elements to consider with respect to the call for the de-balkanization of economic knowledge that comes out of the SFI approach (see Schinckus, 2021). The first is that the rise of new interdisciplinary modeling strategies for studying economic processes leads not only to epistemic regime shifts (in a manner that could be easily reconstructed by historians and philosophers of science), but also to the development of novel cognitive landscapes within which natural, information and social scientists are led to collectively negotiate their vision of social reality.

Nancy Nersessian has recently championed the 'cognitive-historical' method for understanding the evolution of theories and representations in science. She notes that scientists, as much as anyone else, "reason by carrying out thought experiments on internal models" (Nersessian, 1992). She proposes studying conceptual change in science by taking seriously the cognitive foundations of scientific reasoning, which is possible by analyzing the ways in which analogical, simulative (computational), and procedural reasoning both support inference and are productive of novel modes of representation. "The cognitive literature agrees with the position that analogies employed in conceptual change are not 'merely' guides to reasoning but are generative in the reasoning processes in which they are employed" (Nersessian, 2010).

Discussions of conceptual change have largely faded from the literatures on science in part because the way philosophers framed the problem led to increasingly sterile debates and because the move of historians to more social and cultural accounts of science left the mistaken impression that such facets are at odds with the cognitive dimensions of scientific practice. (Nersessian, 2010)

Developing a thorough review and treatment of the ‘cognitive-historical’ method is beyond the scope of the current paper and discussion of economic complexity, but I bring it up for two main reasons. The first is that it helps to characterize the way in which the SFI approach to economic thinking has intervened only sparsely at the level of propounding new economic theories (in the sense of building full-fledged paradigms). Rather, it can be argued that it has intervened more concretely at the level of supplying a landscape, in the form of computational representations and statistical frameworks, within which novel theory can develop, or in which long-established economic propositions can be recodified to meet the demands of contemporary scientific cognition (which is heavily data-driven). The quantal response statistical equilibrium (QRSE) model explored in this dissertation, for example, is precisely an attempt to recodify Smith’s theory of competition within a modern information-theoretic and probabilistic framework. In doing so, it brings long-established intuitions about feedback processes and statistical variation in economic variables to bear on the current landscapes of the complexity view.

It’s also important to put into historical perspective the asymmetries that exist between current and earlier attempts to theorize and model economic complexity. The use of fast digital computers, large datasets, as well as simulation and graph-theoretic (network) methods, are not mere ‘guides’ or ‘heuristics’ to deductive reasoning, but constitutive of new modes of scientific cognition. This idea fully unpacks the deeper content behind Waldrop’s comment that “scientists are beginning to think about more and more complex systems simply because they can think about them” (Waldrop, 1993).

Thus, we may think of the general complexity themes (introduced above), and the emerging themes of economic complexity, as ‘scientific-cognitive’ schemas within which new or ‘updated’ economic thinking can take place. It has often been said that to understand modern science one needs to let go of ‘common sense’ and ‘intuition’ (Chomsky, 1993). While this is true, for example, in the sense that our everyday day intuition about physical objects offers no direct guidance for understanding quantum effects or electromagnetism, this type of argument overlooks the fact there exists such a thing as ‘scientific intuition’. The core of that intuition is built collectively from the analogies, images, narratives, inference procedures, mathematical models and computational structures used to study any given object.



Nersessian argues that scientific cognition should be understood as lying on a continuum with ordinary day-to-day cognition, in the sense that they both operate by 'manipulating and simulating alternative scenarios in mental models' (Nersessian, 1992). The key difference is that in the case of scientific cognition, the analogies and models used are not stored as functional sensorimotor components in the brain, but rather need to be discovered, imagined, or abstracted from empirical observation. Furthermore, scientific cognition is often distributed amongst groups of researchers, tools, labs, and publishing networks. So it is a form of active, conscious and non-automated cognition. This picture of scientific practice lends itself well to empirical research within the field of the psychology of science (Nersessian et al., 2003).

"For example, in investigations of analogies used as mental models of a domain, it has been demonstrated that inferences made in problem-solving depend significantly upon the specific analogy in terms of which the domain has been represented. One example . . . is the study where subjects constructed a mental model of electricity in terms of either an analogy with flowing water or with swarming objects, and then specific inferences (sometimes erroneous) could be traced directly to the analogy" (Gentner and Gentner, 2014; Gentner and Smith, 2012; Nersessian, 1992).

In economics, we have a recent 'natural experiment' version of the Gentner study. As Kirman (2010) notes, the failure to predict or anticipate the 2008 financial crisis can be directly traced to the erroneous models about the financial system which were collectively held by the economics discipline, bankers, regulators, and the public at large. "The recent near-collapse of the world's banking system does not seem to correspond to the collective result of individual banks optimising in isolation and unconsciously coordinating on a disastrous solution. What is involved is a great deal of local interaction, of transmission of information, views and expectations from one actor to the other. Large systems with micro-characteristics of this sort are studied in physics, biology, and also sociology. There, it is recognised that the system may switch rapidly from one phase to another and that this will be dependent on its internal organisation and not on some exogenous shock" (Kirman, 2010).

Kirman's point is essential because it points to the fact that what can make the event of the crisis accessible, from a cognitive point of view, is building an understanding of the events that led up to the near-collapse of the banking system in terms of networks, contagion, trust, and norm-formation. The themes of contagion and networks provide a scientific basis, in the form of complexity schemas, upon

which new theories about financial crises can be developed, or in which older theories (e.g. Minsky) can be recodified using tools such as agent-based modeling and graph theory.

Economies are complex high-dimensional living processes which are notoriously difficult to theorize and fathom. The approach taken historically by 20th century economics was to rely on pure mathematics and deductive reasoning in order to reach deep knowledge about processes for which there is often no available source of scientific intuition. One of the key ideas that I would like to outline here is that the availability of such scientific intuition depends, not only on theory, but also crucially on existing data sources, computational power, compelling narratives, and analogies. From this point of view, it would be anachronistic, for example, to indict early general equilibrium theorists on the claim that they failed to acknowledge or ‘cognize’ the complexity of the economy. That said, I don’t think that holds for the rational expectations theorists, who developed their line of reasoning side by side with Herbert Simon and the artificial intelligence revolution in computer science (Klein, 2015).

## **1.6 Semantic issues in computational complexity**

### **1.6.1 Methods of approximation and partial information**

Economics and operations research sit on the boundary between the discrete and the continuous. Linear programming methods rely on convexity and divisibility conditions. When indivisibilities are taken into account (for example logical decisions, or the limited precision of agent response functions), integer programming and other combinatorial methods may be more appropriate.

In classical physics, when researchers numerically solve a partial differential equation they are ‘approximating the continuous with the discrete’, the ‘infinite with the finite’. But one can also take the opposite direction; given that "continuous structures are often cleaner, more symmetric, and richer than their discrete counterparts (for example a planar grid has a much smaller degree of symmetry than the whole euclidean space)", it is natural to consider embedding discrete structures in continuous differentiable spaces when studying complex systems (Lovász, 2010).

Researchers often have to settle for numerical solutions to a problem whose underlying structure is taken to be continuous, given that closed-form solutions are unavailable. This means that the result obtained inevitably involves **partial information**. An often quoted illustration is the common problem of computing the numerical approximation to a continuous integral (Traub, 2003). Since real or complex-valued functions cannot be directly processed by a digital computer, the effective numerical approximation proceeds by sampling the integrand at a finite number of points. Partial information is also usually contaminated by round-off errors and noise, and can be expensive to obtain in terms of computational costs. The field of information-based complexity (IBC), for example, studies the optimal algorithms and computational complexity of continuous problems for which the available information is partial, contaminated, or 'priced'. Common problems that IBC tackles in physics, finance, and engineering are path integration, partial and ordinary differential equations, nonlinear dynamics, integral equations, fixed-points, and high-dimensional integration (Traub, 2003). The statistical tools explored in this dissertation are one of such class of information-theoretic methods that may be used for inferring efficient solutions to continuous and discrete problems where partial knowledge and uncertainty are present.

Questions regarding 'methods of approximation' can and often do get muddled with semantic issues. Does it exist in a continuum or does it live in a 'quantized' discrete state-space? In many contexts, the structure is considered well known and given, so that questions of representation recede into the background and the discrete vs. continuous dilemma becomes a pragmatic modeling issue. The genetic code, for example, is a discrete biological structure. And as Lovász (2010) notes, "simple basic questions like finding matching (genetic) patterns, or tracing consequences of flipping over substrings, sound more familiar to the graph theorist than to the researcher of differential equations. Questions about the information content, redundancy, or stability of the code may sound too vague to a classical mathematician, but a theoretical computer scientist will immediately see at least some tools to formalize them".

### 1.6.2 Discrete and computational representations

The realization that the building blocks of nonlinear dynamics and chaos (in systems with many degrees of freedom) could not be simply fathomed out of traditional 'laws

of motion' has opened up deep questions regarding the mathematical foundations of complex systems analysis. "The invention of the high-speed computer", as Strogatz (2018) points out, "was a watershed in the history of dynamics. The computer allowed one to experiment with equations in a way that was impossible before, and thereby to develop some intuition about nonlinear systems. Such experiments led to Lorenz's discovery in 1963 of chaotic motion on a strange attractor. Lorenz studied a simplified model of convection rolls in the atmosphere to gain insights into the notorious unpredictability of the weather".

The recourse to simulation and computational methods for studying the qualitative behavior of nonlinear and chaotic systems has given way to the idea that the underlying systems, whether they be physical, biological or social, can be represented as the abstract computational devices capable of simulating their broad range of behavior. This follows from the fact that abstract computational devices can be seen as being able to mimic the complex mapping that takes a dynamical system from initial conditions to its finite number of point attractors. Initial conditions are the 'input' to the abstract machine, and the qualitative equilibrium behavior is the 'output' or pattern associated with that input. Thus abstract machines are studied as being capable of simulating, computing and representing the behavior of complex dynamical systems for which the underlying functions and 'laws of motion' are unstable, partially known, or unknown altogether.

A fundamental question that arises for the complexity researcher is that of defining what type of computational device that 'abstract machine' should be, and of specifying the mathematical or concrete space over which the device is taken to be performing its computations. It is at this point that a dilemma seems to emerge between continuous and discrete methods, and by the same token, between topological and combinatorial alternatives. By 'topological' I mean methods that study and leverage the geometrical properties of continuous differentiable spaces, and by 'combinatorial' I mean methods that confront ways of efficiently exploring the full state-space of systems whose variables take on a finite set of countable values. Brute-force combinatorial optimization methods can be computationally expensive and subject to the curse of dimensionality. The unavailability of the 'local properties' of differentiable functions may impose the need to compute function values for all possible combinations when searching for equilibrium points in any given problem.

From a semantics point of view, the evolution of the methods we use for simulating and approximating solutions to complex nonlinear systems open up new avenues for hypothesizing about the fundamental nature of the processes in question. If the methods are discrete and combinatorial, then their continued and successful use might invite a reevaluation of the abstract or concrete spaces over which ‘computations’ or ‘functional mappings’ are thought to be taking place.

Do information and computation play essential roles at the foundations of physics? Of biology? Of economics? If so, does this imply the fundamentally discrete and combinatorial nature of the system in question ? Not necessarily.

In the context of classical physics, the use of cellular automata<sup>5</sup> to approximate the behavior of continuous dynamical problems does not force a reinterpretation of the spacetime continuum in terms of bits and discrete graph-theoretic structures. In other fields, such as in economics, the introduction of discrete and information-theoretic perspectives do suggest a reevaluation of core theoretical assumptions. A well known example is the work of Herbert Simon, whose analysis of the effective computational planning procedures of firms led to a cognitive reevaluation of the capacities of the economic agent, and to the concept of bounded rationality which lies at the heart of behavioral economics (Simon, 1957). A lesser well-known example is the work of Peter Albin, whose detailed sociological field work on the structural and decision-making organization of the factory work-floor led him to propose a "model for the organization of computations as a model for social and economic organization" Albin (1975). For Albin, the revolutions in information-oriented occupations and technologies constituted an important regime shift in the fundamental structure of the economy (Albin and Foley, 1998).

In his concern for capturing with a high level of fidelity the structural transformations of the new ‘information economy’ we can place Albin side by side with the efforts of economists in the Schumpeterian tradition, in particular those working on models of technical innovation and its effect on aggregate growth (see Freeman, Clarke, and Soete, 1992; Dosi, 1981; Nelson and Winter, 1982; Best, 1990). We can also relate him to a broader sociological perspective that sought to characterize information-related structural changes in the composition of the post-war labor markets (see Beinger, 1986).

---

<sup>5</sup>Cellular automata "may be characterized in analogy with the behavior of dynamical systems (e.g. Ott 1981): simple rules exhibit simple point or limit cycles, while complex rules exhibit phenomena analogous to strange attractors" (Wolfram, 1983).

Other examples include the work of Edward Ames and Ariel Rubinstein. In *Automation and Group Structures in Certain Economic Adjustment Mechanisms*, Edward Ames shows how a Hurwicz type adjustment mechanism can be shown to have the structure of automata and finite semigroups (Ames, 1983). In *Finite Automata Play the Repeated Prisoner's Dilemma*, Ariel Rubinstein studies two-person games in which each player is restricted to carry out his strategies by finite automata (Rubinstein, 1986).

It frequently turns out that the automata closely resemble real-world counterparts, while the logic of their interactions corresponds closely to intuition. In such cases, one operates with a literal and direct image of the reference system—but an image, metaphor, or model that is susceptible to deep and rigorous analysis. (Albin, 1998)

## 1.7 Pooh, the tortoise and the hare

The point to take away from the previous section is that the choice of formalisms, and their underlying mathematical spaces, can end up determining in fundamental ways the space of possible semantics for our propositions. So if we choose to ground our modeling strategies in discrete, graph-theoretic, and computational structures, the knowledge to be gained will depend on our ability to map those elements into real world targets. A much simpler way to put it is simply to state that the choice of equilibrium concepts ends up determining our view of the world.

In discussing the disenchantment that came about with equilibrium concepts in economics following the 2008 crisis, Foley (2017) draws an analogy from A. A. Milne's Winnie-the-Pooh. In A. A. Milne's classic, Pooh attempts to steal the honey from a bees' nest by floating himself to the top of a tree. Helplessly suspended in the air, he is attacked by a swarm of bees. He then philosophically concludes that these must have been "the wrong sort of bees".

What is also at stake in Pooh's story is the issue of adjusting our expectations about what's to be found in the honey. Can we let social reality be the arbiter of the space of possible outcomes and the choice of equilibrium concepts? Or should we insist on tapping into the field of mathematical analysis in order to axiomatically constraint the space of possible solutions? Samuel Bowles, for example, has ad-



Fig. 1.5 Winnie-the-Pooh: The Wrong Sort of Bees, by A.A. Milne

vanced a methodological program in which he proposes to view the outcomes of social interaction as the Nash-Cournot equilibria of specific games (Bowles, 2009; Foley, 2017). It is not immediately salient that what makes this proposal powerful is the idea that we can treat the problem of game-theoretic specification as an ‘inverse problem’. By ‘inverse’ I mean the task of inferring a specification for the underlying game from observed outcomes. The outcomes that can be mapped into any given game are the emergent institutional structures and collective action problems that we observe.

Complementarily to the Bowles program, we can also ask: what are the possible ‘computational devices’ or ‘graphs’ which compute observed socio-structural outcomes? That might be important because the game-theoretic formulation still leaves open the question of causal and functional content. Within the complexity quarters, it is at the level of defining the mathematical foundations of the abstract machines that ‘solve’ nonlinear complex systems where a road split often shows up. In one direction, you find the high-road of topology and analysis. In the other, there is a combinatorial trail which is computationally expensive and messy. In the latter case, what one deems ‘computable’ requires careful attention to the social and logical properties of whatever device we take to be ‘performing the actual computations’. The analysis perspective, on the other hand, aims to provide proofs of computability which can be settled on mathematical grounds alone.

It is interesting to consider, for example, how Stephen Smale’s work on theoretical economics led him straight into foundational questions in the theory of computation. His interest in dynamic convergence to equilibrium led him to work on global Newton

algorithms, and to research on average stopping times for simplex algorithms in linear programming (see Hirsch et al., 2012; Smale, 1976). This work sought to give a ‘differentiable dynamics’ version of the combinatorial methods that had been pioneered before by Herbert Scarf and C. Eaves (Eaves, 1971; Scarf, 1967). The work of Scarf and Eaves had shown how to translate Sperner’s classical existence theorems into concrete computational procedures for approximating fixed points. In further work with M. Shub, Smale’s approach of translating numerical optimization problems into geometric ones broke new ground in the theory of *real* computational complexity (real as in taking place in  $\mathbb{R}^n$ ). "Smale always emphasized that he looks at algorithms as mathematicians do, in terms of real numbers, and not as computer scientists do, in terms of a finite number of bits of information" (Hirsch et al., 2012).

The story of the tensions that arise between combinatorial and analysis perspectives is like Zeno’s version of the race between the tortoise and the hare; the tortoise wins the race by fiat. And we should be happy that the tortoise is able to frame the race in such terms because that is ultimately what sustains progress in advanced mathematics. But in economics, it’s also worth exploring the possibility of letting the hare win for once. If anything, because we might be surprised by the results.

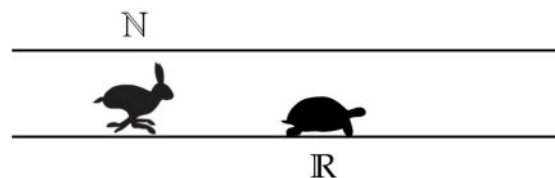


Fig. 1.6 Zeno’s paradox. The tortoise always remains slightly ahead of the hare, no matter how many times it tries to catch up.

Mathematics has infinitely many representations from which to choose from. In framing the problem of choosing equilibrium concepts and finding functional specifications as ‘inverse problems’, we can let observations, data, and scientific imagination guide our explorations for the right sort of formalisms. The next chapter discusses maximal entropy methods and how they can be used to find ‘informationally efficient’ representations in the social sciences.



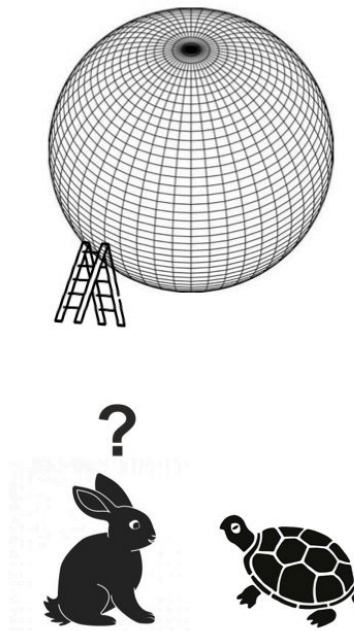


Fig. 1.7 Discrete approximation.

## Chapter 2

# Entropy, inference, and Boltzmann distributions

### 2.1 Probability distributions

Complex social, biological, and economic systems are often well described by exponential and heavy-tailed probability distributions. Some examples are the distributions of profit rates, city populations, and financial market fluctuations (see Peterson et al., 2013; Sornette, 2007). Among the reasons for distinguishing distributions is that they may be linked to different mechanisms, and examining their entire shape gives clues to the processes at play. Often, if data is not missing, finding the shape of a distribution can be as simple as looking at a histogram or fitting a suitable mathematical function. Fitting an available probability model can be followed up by thinking about why that particular function is a good fit. That is, what are some of the reasons that could lead a system to exhibit that particular shape.

The study of distributions goes back to Abraham De Moivre and Laplace. During the time of De Moivre and Laplace a central problem in science was that of finding universal "error curves". Laplace showed how fluctuations in social statistics could be described by the normal distribution, which had originally been proposed by De Moivre in his *Doctrine of Chances* (1718) as a method for approximating binomial coefficients. The problem of finding a law of errors was commonly encountered in astronomy and geodesy, where the goal was to describe variations in measurement from a true physical quantity, e.g. the distance to the moon (Geraci and Borja, 2018).

The study of Laplace’s eponymous distribution (1774), which is a non-normal error curve, has recently received renewed interest. Laplace and Normal distributions both belong to the exponential family. What distinguishes them is the specification of their error functions; the Laplace is an exponential of the numerical *magnitude* of the error, while the Normal is an exponential of the *square* of the error (see Kotz and Kozubowski, 2001).

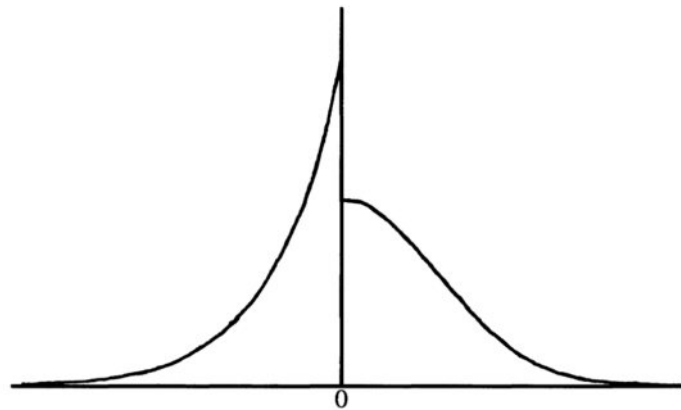


Fig. 2.1 On the left, the Laplace distribution:  $f_x = \frac{k}{2}e^{-k|x|}$ . On the right, the Normal distribution:  $f_x = \frac{1}{\sqrt{2\pi}\sigma}e^{-x^2/2\sigma^2}$ . Reproduced from Kotz and Kozubowski (2001)

One reason for using a Laplace distribution over a Normal distribution is if there is heterogeneity in the measurement of errors. Geraci and Borja (2018) give the example of astronomers in different locations trying to measure the distance to a celestial object. Every astronomer is trying to measure the same quantity, but inevitably their measurements will have different variance. So if you look at the datasets of each astronomer, they will likely be Normal. But if you’re tasked with collecting the measurements of all the astronomers into a single histogram, you will likely obtain a Laplace like curve. Observer-instrument variability is hence one possible reason for using the Laplace distribution. This distribution commonly shows up in hydrology, in finance, in speech and pattern recognition, and more recently, in statistical political economy and the industrial organization literatures (see Bottazzi and Secchi, 2005).

There are cases in which it might be of interest, not only to find the appropriate shape of a distribution, but to find out what generative principles could be leading to its distinctive statistical signature. Barabási and Albert (1999), for example, have shown how dynamic network models can lead to scale-free power law distributions.

Scharfenaker and Foley (2017) show how a principle of market feedback leads to a Laplace-Normal type curve for the distribution of profit rates.

Principle-based derivations are better understood as *inverse* problems. They are inverse because the goal is to infer the set of possible underlying mechanisms or rules that generate the data. 'Principle' simply means that we are using a description which is general enough to encompass a wide set of possible functional or dynamic specifications. Information-theoretic statistical mechanics provides a powerful framework for solving inverse problems, particularly those in which the data is exponentially distributed. The method of constrained entropy maximization is a variational principle which yields an informationally efficient solution to the inverse problem.

As an entry point, one can think of it as the task of finding the rule which yields the right "law of errors", call it  $p_k$ . Constrained optimization of the Boltzmann-Gibbs-Shannon (BGS) entropy functional  $S[\{p_k\}] = -\sum p_k \log p_k$  leads to the exponential Boltzmann distribution  $p_k \propto e^{-\beta H_k}$ . Similar to the distinction between the Normal and the Laplace distributions, what distinguishes different Boltzmann densities is what goes in the exponent. In statistical mechanics this is known as the *energy functional* or *hamiltonian*. The objective is to find out what should go in that exponent, proceeding cautiously if you will, by imposing as little structure as possible on the problem from the outset. The following sections discuss and motivate this task in detail.

## 2.2 Complex systems, statistical physics and maxent

In complexity science we confront systems which have many degrees of freedom. The behavior of a system  $S$  may be characterized as depending on a set of variables  $\{x_j\}$ , for  $j = 1, 2, \dots, N$ . When aiming for a probabilistic description of  $S$ , in which we are not describing or computing the individual trajectories of each  $x_j$ , the first natural question to ask is the following: What is the joint probability distribution  $P(\{x_j\})$  of all these variables? In the simplest view, we ask what that  $P$  is at any given point in time. That is, what is the joint probability of the variables from which we can sample a cross-sectional snapshot of the system?

From the point of view of statistical physics, the  $x_j$ s are the microscopic state variables for the system's components, such as atoms and molecules. A central goal of statistical physics is to build a bridge that takes us from an understanding of what microscopic components are doing to a macro-level description of the emergent collective behavior of a large number of interacting parts.

The micro to macro bridges of complexity science are foremost statistical and not deductive. The building materials required for the assembly of such complexity bridges are: i) some working knowledge, or hypotheses, about the principles governing micro-level interaction, and ii) probability and data.

There is an emerging simplicity and universality that comes out of the problem of describing complex systems in terms of a joint distribution  $P(\{x_j\})$ . New variables and equations emerge which have the remarkable property of being applicable to a broad range of systems across different scales, and across physical, social and biological domains. In classical physical systems, such as ideal gases, the statistical analysis of large numbers of molecules moving randomly and independently within a volume leads to the emergence of deterministic laws and constants, e.g. the ideal gas law (see Feynman, 1965). The pressure on a section of a gas volume, due to the independence and collective randomness of its particles, turns out to be the average force exerted by all the particles, divided by the relevant area.;  $P = \frac{\langle F \rangle}{A}$ . Thus, in such systems, the effects are 'additive' and unsurprising. What is interesting, however, is the notion that a physical constant emerges from the statistical aggregation of random particle collisions on a piece of wall (Bialek, 2017).

For classical deterministic systems, what does the work in guaranteeing the emergence of constants, or in determining that averages can be effectively regarded as exact numbers, is the central limit theorem (CLM) (see Fischer, 2010). In a nutshell, what the CLM says is that the mean value of a large number of *independent* observations will follow a Normal (bell-shaped) distribution. But more crucially for physicists, it states that as the number of observations tends to infinity ( $N \rightarrow \infty$ ), the variance of the distribution will shrink and tend to disappear. Since the systems considered by classical physics often contain a very large number of particles, on the order of Avogadro's number ( $10^{23}$ ), variations around the average can be effectively ignored. This is how determinism and constants emerge within a probabilistic framework, and it's what allows physicists to focus on their equations, and avoid the need to constantly amend key distributional assumptions (physical laws are stable).

In the social and biological world, where randomness is constrained by distinct principles of interaction and organization (e.g. feedback), one needs to handle the CLM with care. Furthermore, in such contexts what ends up being of fundamental interest is the type of variation we perceive around the mean. In the same way that what distinguishes a falling rock from a bird is the latter's adaptive capabilities, what distinguishes measurements of living organisms from measurements of solids is the former's non-normal statistical variance. At any limit, zero or infinite variance are not effective recipes for sustaining biological and social organization.

The most interesting systems that can be studied by a statistical physics framework are precisely those which are driven by collective behavior, and which show emergent patterns and statistical properties that turn out not to be scaled up versions of the individual components. Social and biological systems are fundamentally driven by organized collective behavior. Some common examples are herding in markets (Kirman, 1993), patterns of correlated neural firing (Schneidman et al., 2003), antibody diversity in the immune system (Mora et al., 2010), animal conflict, linguistic pattern formation, bird flocking, species abundance, and voting (Lee et al., 2015). In the next section, as a first example, I discuss the problem of understanding bird flocking from a statistical point of view. I summarize and use the discussion from Bialek et al. (2012), which provides a pedagogical entry point but is also a concrete scientific application.

### 2.2.1 Bird flocks, criticality and statistics

Birds flock to forage and migrate collectively. Local interactions between birds within a flock lead to emergent patterns of global coordination. The global emergent patterns are both highly robust and fragile; an adaptive functional trait which provides benefits for foraging and safety from predation (Bialek et al., 2012). The emerging patterns are collectively robust in the sense that they can withstand perturbations of the flock's individual parameters, and they are fragile in the sense that they are poised near critical points in their phase space. To melt an ice cube, or bring water to boil, one needs to fine-tune some sort of temperature control. Collective biological phenomena, such as bird flocks, appear to 'self-tune' in a distributed fashion towards near-critical points. "Rather than being dictated by a leader or in response to a common stimulus, the collective patterns of flock dynamics tend to be self organized, and arise from local interactions between individuals, which propagate information

through the whole group. Flocks, schools and swarms also are highly responsive and cohesive in the face of predatory threat. This balance between order and high susceptibility points to the idea of criticality" (Mora and Bialek, 2011; Munoz, 2018).

A statistical snapshot of the flock (e.g. of size  $N$ ) should give us information about, not only their average speeds, but also about variations in their directions. Assuming that the cohesive tendency of the flock is partially maintained by a persistent average speed for all the birds, we could focus (for example) on studying variations in their directionality. In this bird flocking scenario, and referring back to the problem of finding the system's joint distribution  $P(\{x_j\})$ , for two degrees of freedom, we can take the set  $\{x_j\}$  to be the set of individual bird velocities. So we appropriately redefine the set to consider as  $\{\vec{s}_j\}$ , where  $\vec{s}_j = \frac{\vec{v}_j}{|\vec{v}_j|}$  are the normalized bird velocities.

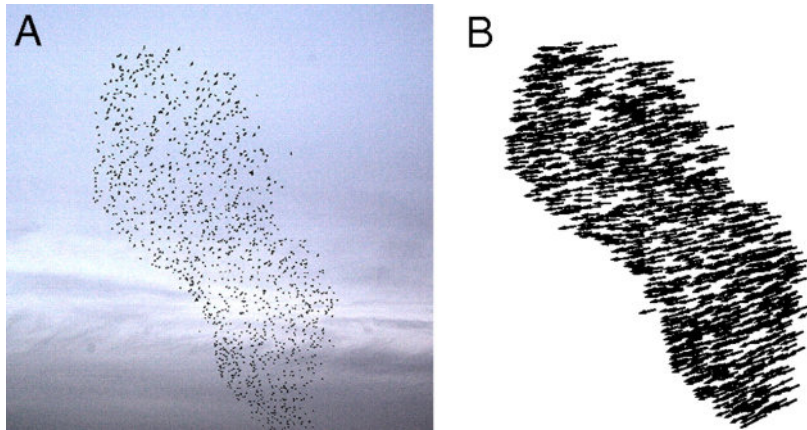


Fig. 2.2 (A) Bird flocking event. (B) Velocity vectors from snapshot. Graph from Bialek et al. (2012)

What statistical theory of directional ordering for the flock can be derived from a set of measurements taken in field observations <sup>1</sup> ?

To answer the question in a simple manner we need to define, or make an assumption about, key properties that can characterize directionality in the flock. Are birds looking to their neighbors when deciding where to move ? Or are they computing some high-dimensional and complex non-linear function in real time ? If each bird can choose amongst  $d$  possible (vector) orientations, the set of possible

<sup>1</sup>Field observations for flocks use stereometry to obtain the 3D positions of birds at different points in time (see Cavagna et al., 2008)

distributions (the lists of individual velocities) is of size  $d^N$ . Considering that real flocks can have over 1000 birds, even if we defined them to be Tetris-like in their ability to choose directions, the number of possible distributions would turn out to be quite large. Considering a simplified model of say 2-5 birds from the outset could be potentially be misleading, since what we're trying to understand is precisely how the local behavioral interactions of *many* birds lead to the observed collective coordination patterns. Thus, number (the size of the flock) matters.

Building a probabilistic model of the flock inevitably involves imparting parsimony into our description. If we believe that birds look to their neighbors to decide where to go, the simplest structuring principle that we can track (measure) is the correlation between velocities in the flock. This is given by the cross-correlation matrix  $C_{ij} = \langle \vec{s}_i \cdot \vec{s}_j \rangle$ . The symbol  $\langle \cdot \rangle$  indicates the average; if the average is taken over a specific distribution  $P$  we write  $\langle \cdot \rangle_P$ .

Out of all the possible distributions  $P(\{\vec{s}_i\})$ , which is the one that is consistent with the field measurements of the flock's bird-to-bird correlations, but otherwise as random or as unstructured as possible? As phrased by Bialek (2017), translating "as random as can be" into probability terms means maximizing the entropy of the distribution. If we impose no structure on  $P$ , in order to reduce the entropy, then we would find that  $P$  is  $\sim$  Uniform; yielding a gaseous picture of the flock. The idea of imposing a minimal structure on the distribution follows from the commitment to make our description as parsimonious as possible. This is not only motivated from an inference and tractability perspective, but theoretically from the expectation that what leads to the observed global coordination patterns is a simple principle of local interaction, and not some complex nonlinear function of the birds' trajectories.

In considering this problem, with the aim of testing the probability model on flocks of European starlings (*Sturnus vulgaris*), Bialek et al. (2012) obtain the Boltzmann-Gibbs type distribution:

$$P(\{\vec{s}_i\}) = \frac{1}{Z(\{J_{ij}\})} \exp \left[ \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N J_{ij} \vec{s}_i \cdot \vec{s}_j \right] \quad (2.1)$$

The distribution of 2.1 is a 'maximum entropy distribution' because it is the simplest one we can obtain by injecting minimal structure into what would otherwise



be a uniform measure. The term  $Z(\{J_{ij}\})$  is the normalizing factor and the parameter  $J_{i,j}$  gives a measure for the strength of coordination between any pair of birds.

Boltzmann-Gibbs distributions are ubiquitous in statistical mechanics. The key point to understand about this general distribution is that it gives the probabilities for the system's possible configurations ( $\{\vec{s}_j\}$ ) in terms of its energy functional. This can be the total energy of the system, or a potential. For a single degree of freedom  $x_j$  with energy  $\mathcal{H}(x_j)$ , the general form is given below in 2.2:

$$P(x_j) = \frac{1}{Z} e^{-\beta \mathcal{H}(x_j)} = \frac{e^{-\beta \mathcal{H}(x_j)}}{\sum_{j=1}^N e^{-\beta \mathcal{H}(x_j)}} \quad (2.2)$$

When working with systems which we know can be suitably described by a Boltzmann distribution, such as Bialek's bird flock, the direct dependence on the energy functions means that we can often amplify our assumptions about the forces driving the system simply by expanding the term in the exponent. So if we consider two potentials acting on a single degree of freedom, for example, we would obtain something like  $P(x_j) \propto e^{\beta_1 \mathcal{H}_1(x_j) + \beta_2 \mathcal{H}_2(x_j)}$ . If we are committed to an (initial) parsimonious probabilistic description, we would like to make that sum in the exponent as informationally efficient as possible. So if complex causal structure and nonlinearities need to be considered, they can be incorporated either by adding/modifying the potentials in the exponent, or via a maximum entropy derivation (amongst other methods). Hence, the subtlety lies not only in the use of the entropy functional, but also in the idea that we can regularize our knowledge about the world and obtain likelihoods for our data by a simple rule of exponentiation<sup>2</sup>.

In the flocking example considered here, keeping track of the pairwise directional correlations between birds is enough to generate the candidate distribution. Physicists will recognize this maximum entropy (henceforth *maxent*) distribution as relatable to models of spontaneous magnetization and local interaction among individual spins (equation 2.1 above). In particular, as Bialek et al. (2012) elaborates, this distribution corresponds to the classical Heisenberg model. The Heisenberg model describes a set of unit spins in  $\mathbb{R}^3$  placed on the nodes of a  $d$ -dimensional lattice and has the following Hamiltonian (Wu, 1982):

<sup>2</sup>This latter idea is what motivates the use of the Boltzmann distribution or 'softmax' function for normalizing the output of artificial neural networks in multinomial classification tasks. John S. Bridle introduced the term 'softmax' (see Bridle, 1990)>. For a review of recent work lying at the intersection of statistical mechanics and deep learning see Bahri et al. (see 2020)

$$\mathcal{H}(\{\vec{s}_i\}) = -\frac{1}{2} \sum_{i,j} \mathcal{J}_{ij} \vec{s}_i \cdot \vec{s}_j \quad (2.3)$$

As Bialek further notes, it is useful to know to which statistical physics model the maxent distribution can map to, as this allows us to leverage the well studied properties of that model to make further inferences about the flock.

Note that the conclusion that the classical Heisenberg model is the appropriate one that we can use to describe Bialek's flock of European starlings arrives as a result of imposing the simplest possible structure on the full set of possible distributions  $P(\{\vec{s}_j\})$ . So that what we are solving, following this sequence of reasoning, is an inverse problem. "The fact that equilibrium statistical mechanics is the prototype of maximum entropy models encourages us to think that the maximum entropy construction defines an effective 'energy' for the system" (Bialek, 2017). The inverse nature of the problem lies in that the problem is set up so as to 'recover' a specification of the energy functional from field measurements of the bird-to-bird velocity correlations.

However, if the distribution in 2.1 is the one to be used, this implies estimating the matrix of parameters  $J_{i,j}$ . The inference problem can be simplified by introducing a *mean-field approximation*. Bialek does this by assuming that each bird in the flock only experiences an average 'social force'  $J$  from its first  $n_c$  nearest neighbors. This reduces the distribution to:

$$P(\{\vec{s}_i\}) = \frac{1}{Z(J, n_c)} \exp \left[ \frac{J}{2} \sum_{i=1}^N \sum_{j \in n_c^i} \vec{s}_i \cdot \vec{s}_j \right] \quad (2.4)$$

With this simplified model, there are only two parameters to be estimated from the data;  $J$  and  $n_c$ . Knowledge of the inferred Hamiltonian can be used to elaborate a dynamical description and simulate the flock, using  $\frac{d\vec{s}_i}{dt}$ . This mechanical-dynamical extension for modeling the flock will generally take the form of a Brownian type process:  $\frac{d\vec{s}_i}{dt} = -\frac{\partial H}{\partial \vec{s}_i} + \text{noise}$ .

A key point to make at this juncture is that although we know the flock to be non-ergodic; the inferred Hamiltonian can be used to construct a dynamic ergodic sampler, from the trajectories  $\vec{s}_i(t)$ , which converges to the stationary maxent distribution.

Depending on the Hamiltonian, the process may or may not follow detail balance conditions, or markovian properties.

This is a rather tricky point that is still seems sparsely discussed in the literature, but my current understanding is that the process is ergodic with respect to the MaxEnt distribution. So it is an inferential and 'bootstrapping' type of device that can be used for sampling *plausible* dynamics from the stationary distribution.

It's reasonable to do this because the flock, as Bialek suggests, seems 'poised at criticality' and exhibits 'punctuated equilibria'. Meaning that if we measure the bird-to-bird correlations over several photograms of a film of the flock flying in cohesion, the estimates are unlikely to vary dramatically, up to the point where a predator shows up or something major happens. It is in this sense that we are able to build a 'statistical equilibrium' understanding of the flock, without implying that the system is in lifeless thermal equilibrium. Thus, self-organized criticality and the concept of punctuated equilibrium, or the idea that living systems experience long periods of quiescence followed by abrupt change, are two fundamental notions that motivate the application of statistical mechanics and statistical equilibrium reasoning for studying social and biological systems (see Munoz, 2018) .

### 2.2.2 Why birds ?

I chose the bird flocking example of Bialek et al. (2012) because it provides an integrated treatment of both maximum entropy and statistical physics approaches for building statistical descriptions of complex social systems driven by principles of local interaction and adaptive behavior. Furthermore, despite the mathematics involved, a flock of birds is a concrete visual object that is easier to think about than other lower or higher-dimensional examples. Also, it is aesthetically pleasing and fascinating to look at. Below I summarize some key methodological points to take away, before discussing potential social-economic applications of Bialek's bird flock model, and how those type of extensions could be supported and justified within the broader 'complexity view' that I have been motivating throughout. In the next section I provide a more exhaustive and didactic treatment of the maximum entropy principle.

**Key learnings from the starlings:**

- By imposing a minimal structure on the space of all possible bird velocity distributions, it's possible to build a parsimonious probabilistic model that can be tested with field data of bird-to-bird correlations.
- The 'efficient' energy functional recovered can be mapped to well known statistical physics models (e.g. Heisenberg, Ising, etc.) in order to explore and simulate plausible dynamics.
- The bulky inference procedure of having to estimate a large matrix of parameters, as well as measure all pairwise bird-to-bird correlations, can be greatly simplified by using a mean field approximation. This is a variational procedure that approximates a difficult hamiltonian by a more tractable one. This reduces the number of parameters to two; the global strength of interaction ( $J$ ), and the size of the bird neighborhoods driving the flock's coordination pattern ( $n_c$ ). Surprisingly, by measuring a single number (the observed correlation), it is possible to recover the entire distribution of velocities  $P(\{\vec{s}_i\})$ .
- The estimation procedure for this problem is far more involved than its conceptualization. It can be done using maximum-likelihood methods, as in Bialek et al. (2012). Or using Bayesian estimation and MCMC sampling, as is done for the local public goods model in this dissertation.

The Heisenberg model of ferromagnetic interaction inferred by Bialek and his team allowed them to study the flock as an information propagation mechanism. Their results show evidence that the starlings do not directly 'communicate' across long distances ( $n_c$  was estimated to be around 11), and that the flock achieves velocity coordination independently of how 'congested' it is; "interactions are ruled by topological order rather than metric distance" (Bialek et al., 2012). In the same way that electron spins may align without long-range interaction in a ferromagnet, the flock propagates directional order by relying on the transmissibility of correlations between neighborhoods of fixed size.

Thus, Bialek and his team find evidence for the emergence of a nontrivial topological constant which structures information propagation in the flock. Crucially for the discussion broached here, I think it's important to note just how reliant this finding

was on the commitment to impose the least structure possible on the inferential problem from the outset, and on the motivation to obtain the most parsimonious probabilistic description possible that could match the data.

### 2.2.3 Animal models and analogical transfer

One of the major themes that I have tried to motivate throughout the different discussions in this dissertation is the idea that the ‘complexity view’ provides novel schemas for thinking and making inferences about social systems. I pointed to those schemas as being ‘cognitive’ because they are not merely aides to deductive reasoning, but constitutive of the conceptual, computational and statistical territories within which new insights can emerge.

In social science, and economics in particular, where scientific intuition is extremely difficult to build and stabilize across research groups and institutions, it is worth reflecting on what social animal models can do for us in terms of grounding meaningful images of human social behavior. Animal models, as the work of Alan Kirman and others has shown (see Kalenscher and Van Wingerden, 2011; Kirman, 2010), provide rich descriptive moulds that can be adapted in order to gain knowledge about processes of herding and information diffusion in economic contexts.

Statistical physics models, and in particular models of magnetism such as the Ising model, provide a common currency for building analogical transfers between social animal and human behavior, in an empirically testable way that goes beyond a mere metaphorical assertion. Statistical physics models of social dynamics have attracted a great deal of attention in recent years, with topics as diverse as opinion dynamics, cultural dissemination, herding in financial markets, complex social networks, linguistic evolution, consensus dynamics and voting (see Castellano et al., 2009).

Considering the bird flocking example, if we wanted to build a human social-economic application, the key for building any analogical transfer would lie in specifying the space of interaction. The starlings interact in 3D space, which is what makes the bird flock phenomenon intuitive and easily graspable. Human interactions, in particular those that support fads and herding, take place in networked virtual spaces with metrics that may be difficult to define (See Kirman (2010) for a discussion on how to define network metrics for social and economic applications). If one finds

a suitable metric and underlying decision-theoretic space, it becomes possible to start thinking about how to adapt the prototype. The interpretation of the degrees of freedom will vary depending on the application: "a binary variable will indicate yes/no to a political question in opinion dynamics, two synonyms for a certain object in language evolution, two languages in competition, whether somebody has been reached by a rumor or not, etc. . . ." (Castellano et al., 2009). The set of questions that could be studied in such a model is broad : "What are the fundamental interaction mechanisms that allow for the emergence of consensus on an issue ? What favors the homogenization process? What hinders it ?" (Ibid.).

### 2.2.4 Disciplinary tensions

A point of tension for the introduction of maxent and statistical physics methods in economics, from a disciplinary perspective, comes up at the level of justifying a commitment to solving inverse problems. This point of tension is perhaps stronger than the one that arises simply from the application of frameworks and analogies drawn from the stock of physics models.

There are two layers to this tension. The first one has to do with the modeling of economic rationality and the choice of equilibrium concepts. As Kirman points out in his Ants paper;

While the economist is prepared to accept that insects may follow some mechanical rule of behavior which may have evolved so as to be close to optimal, or may be locally optimal in the space of possible strategies, if the same model is applied to economic agents, he usually requires that the behavior should be fully and consciously optimizing. (Kirman, 1993)

Thus, at one level, the requirement that analogical transfers from other fields conform to optimality conditions seems to force economists into adopting a view which may end up contradicting the model's original insights. This type of view, for example, has led some researchers to define information propagation in terms of fully optimizing behavior; "An informational cascade occurs when it is optimal for an individual, having observed the actions of those ahead of him, to follow the behavior

of the preceding individual without regard to his own information" (Bikhchandani et al., 2021; Parker and Prechter, 2005).

It's important to remember that maximum entropy is a principle, but it's also a quantity; systems appear to us in our measurements as operating at specific information-theoretic entropies. Why this should be the case is an open problem in information theory (see Landauer, 1996), but in studying social and biological systems, we leverage that fact in order to extract 'efficient' knowledge about the system.

The idea of extracting knowledge directly from the data, bringing to bear as few theoretical assumptions as possible, is difficult to sell wholesale in the economics discipline. An even less palatable proposal for many economists, and some philosophers of science, is the idea that such knowledge is a form of conjectural explanation.

This is the second, and perhaps more fundamental level, at which tensions arise. One may be able to convince some neoclassical economists of not injecting omniscient agents into the nodes of an Ising model, but it will be harder to convince researchers, across many schools of thought, of treating the challenge of defining interaction networks and microeconomic energy functionals as an open-ended, context-specific, data-driven inverse problem. This issue goes as far back as the 'measurement without theory debate' (see Morgan, 1990). An incrementalist approach to modifying model assumptions is unlikely to nudge economics into a more stable interdisciplinary equilibrium.

Social animal models and statistical physics prototypes, along with maximum entropy and other Bayesian variational methods, can be used as a common currency by researchers of *all disciplines* to pose questions and test hypotheses about economic data. The complexity view provides a steady source of analogies, concepts, jargon, as well as powerful computational and statistical methods for building general scientific intuition. In a more recent discussion, Kirman has argued for the need to adopt an ecology of methods, and to study 'noisy systems of interacting agents'. This idea is relatable to the maxent perspective; a noisy system is a system that has both probabilistic and deterministic components. The maximum entropy principle asks: What is the simplest structure that we can impose on a system's distribution, but let it otherwise be 'as random as can be'? In setting up the problem in this fashion, the *as random as can be* part is what allows us to uncover the type of noise peculiar to the system in question. And since noise can come in many shapes and forms, it

seems reasonable to want to know what its kind is, and to understand how it relates to the system's functioning.

## 2.3 Maximum entropy : a variational principle

When dealing with problems where there is uncertainty with respect to what statistical physics prototype to use, or where the microeconomic energy functionals need to be adapted or revised, it is worth considering what the least biased approach is that will lead us to a parsimonious solution. The increasing availability of numerical, analytic, and computational tools, has made it possible to solve for the maximum entropy (maxent) model that reproduces the behavior of large collective systems (Lee and Daniels, 2018). Having solved the inverse problem, the researcher can focus on finding an appropriate mapping to a prototype that can support simulation and further inference. She may either draw from the stock of statistical physics models, or decide to propose a new energy functional which may be grounded in well established behavioral, sociological and economic theories (See Aoki and Yoshikawa (2011) and Gallegati et al. (2008) for complementary perspectives in economics).

Pairwise maxent models, such as the Ising, or binomial maxent choice models, offer good starting points for exploring a wide range of economic datasets. Their conceptual simplicity and generality makes them particularly attractive for tackling questions of self-organization and statistical feedback in decentralized market environments. In the following sections I provide an overview of the classical maximum entropy inference problem. I also provide a set of concrete and didactic examples that can help get a feel for how it works.

### 2.3.1 What is information entropy ?

The concept of information entropy was introduced by Shannon in his famous 1959 paper: *A mathematical theory of communication* (1959). It provides a unique measure of uncertainty which conforms to elementary consistency principles. The Shannon entropy of the probability distribution  $p(s)$ , over possible states  $s \in \Gamma$  of a system, is given by:



$$S[p] = - \sum_{s \in \Gamma} p(s) \log p(s) \quad (2.5)$$

The states  $s$  can be the outcomes of a  $k$ -sided die, letter arrangements in a word, the orientation of spins in a ferromagnet, or the entry/exit decisions of investors in a financial asset, to name a few examples. Entropy is at a maximum when the distribution is structureless. If we consider a noisy source, for example a random number generator, this means that each output of the program will be ‘maximally surprising’ to us, and after recording a large number of trials the histogram will be uniform. At zero entropy, there is no surprise, and the observations will all pile up into a single bin. The measurements we have of social and biological systems will tend to fall somewhere in between, and by thinking in information-theoretic terms, we can link the usual ‘law of errors’ to a specific quantity; the entropy.

Depending on how much structure we are willing to impart on an otherwise uniform distribution, the entropy will vary. Thus, our state of knowledge about the source is naturally encoded by that quantity: entropy reduction results in knowledge gain. This was Shannon’s fundamental insight, and the maximum entropy procedure provides an avenue for making our knowledge gain as ‘efficient’ as possible. The classical treatment and exposition of the maxent principle is given in Jaynes (1957).

### 2.3.2 Dies and uncertainty

The simplest example of the application of the maxent principle is the problem of assigning a distribution to a six-faced die. Suppose we rolled the die a large number of times, and recorded an average of 4.5. From knowledge of that average, what distribution does the maximum entropy principle prescribe? In the absence of any constraints, our best guess would be a uniform density. To make a prediction that is consistent with the available information, but that is otherwise as least structured as possible, we maximize the entropy subject to a single moment constraint, and normalization. For  $i \in \{1, 2, 3, 4, 5, 6\}$ , and  $p_i = p(i)$ , we maximize  $S = - \sum_{i=1}^6 p_i \log_e(p_i)$  subject to  $\sum_{i=1}^6 p_i = 1$  and  $\sum_{i=1}^6 i p_i = 4.5$ . Using the method of Lagrange multipliers, this yields the distribution shown in figure 2.3.

This example provides a simple entry point, but it doesn’t really tell us why we should use the Boltzmann-Gibbs-Shannon entropy as the objective function in the

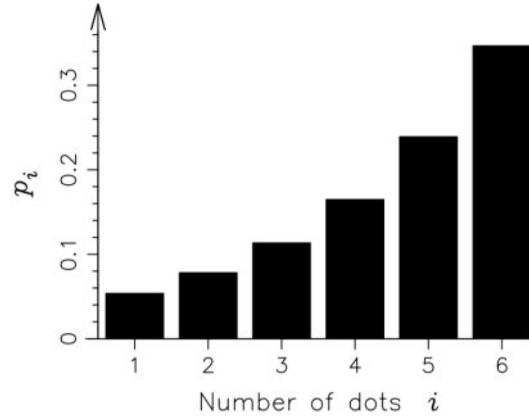


Fig. 2.3 Maxent assignment for a six-faced die with an observed average of 4.5. Reproduced from Sivia and Skilling (2006)

maximization problem, nor why the procedure should work at all (Sivia and Skilling, 2006). The following examples, found in (Gull and Skilling, 1984; Jaynes, 1986; Sivia and Skilling, 2006), along with a brief discussion on some relevant critical points, provide deeper justifications of the principle.

### 2.3.3 Kangaroos and logical consistency

Consider the following probability problem:

- **Information** : A third of all kangaroos have blue eyes, and a quarter of all kangaroos are left-handed.
- **Question** : On the basis of this information alone, what proportion of kangaroos are both blue-eyed and left-handed?

The four possible states of a kangaroo are:

$$\{(\text{blue, left}), (\text{blue, right}), (\text{not blue, left}), (\text{not blue, right})\} \quad (2.6)$$

Skilling and Sivia use a  $2 \times 2$  contingency table to represent the situation. I reproduce it below in figure 2.4.

In the table above,  $p_1 = x$ , which is the probability that the question asks us to find. The information given tells us that  $p_1 + p_2 = P(\text{blue}) = 1/3$  and  $p_1 + p_3 =$

		<i>left-handed</i>	
		True	False
<i>blue-eyed</i>	True	$p_1$	$p_2$
	False	$p_3$	$p_4$

(a)

		<i>left-handed</i>	
		True	False
<i>blue-eyed</i>	True	$0 \leq x \leq \frac{1}{4}$	$\frac{1}{3} - x$
	False	$\frac{1}{4} - x$	$\frac{5}{12} + x$

(b)

Fig. 2.4 (a) The probabilities for the four possibilities:  $p_1, p_2, p_3, p_4$ ; (b) parameterization in terms of a single variable  $x$ , where  $x = p_1$ . Reproduced from Sivia and Skilling (2006).

$P(\text{left}) = 1/4$ . The set is exhaustive:  $p_1 + p_2 + p_3 + p_4 = 1$ . If we have no reason to assume that having blue eyes and handedness are correlated, we would expect that  $P(\text{blue}, \text{left}) = P(\text{blue})P(\text{left}) = 1/12$ . In this case the preferred solution that we're seeking is independence of eye-color and handedness. If we want to make the genetic outcome of our kangaroos 'as random as can be' then we should make their traits independent. It is natural to ask, as Sivia states, what function could serve as a good *variational principle* to obtain the preferred solution. So if we cast the problem of finding  $\{p_i\}$  as a constrained optimization problem, what should be the function that we use ?

Skilling (1988) proved that the only functions which give uncorrelated assignments are those monotonically related to the Shannon entropy. I reproduce the table from Sivia and Skilling (2006) below, which shows how different functions perform in finding the optimal value for  $x$ , and the implied correlations between kangaroo traits (figure 2.5).

Thus, the use of the entropy functional can be motivated from the point of view of information theory, as was done by Shannon, but also axiomatically from the point of view of logical consistency. The canonical axiomatic treatment was given by Shore and Johnson (1980). Other possible motivations can be approached from the point of view of group theory (Jaynes, 2003), as well as through arguments of exchangeability and sufficient statistics; see De Finetti (1974); Diaconis and Freedman (1980); Porta Mana et al. (2017); Skyrms (1987).

Variational function	Optimal $x$	Implied correlation
$-\sum p_i \log[p_i]$	0.0833	None
$-\sum p_i^2$	0.0417	Negative
$\sum \log[p_i]$	0.1060	Positive
$\sum \sqrt{p_i}$	0.0967	Positive

Fig. 2.5 Different candidates for the variational principle. See Gull and Skilling (1984). Table from Sivia and Skilling (2006)

### 2.3.4 Monkeys and multiplicity

Another justification for the use of the entropy functional as a variational principle, given by Jaynes (1986), uses a combinatorial argument and a rowdy team of monkeys. Supposing we have to assign distinct probabilities to  $\{x_i\}$  possible states, and that we require those assignments to be consistent with a set of constraints, such as  $x$  being positive ( $f(x) > 0$ ), or  $x$  taking on some average value, we can ask the monkeys to help us out by having them throw balls at random into a boxed 1D grid partitioned by the  $\{x_i\}$ . The situation is displayed below in figure 2.6.

Call the set of constraints  $\mathcal{C}$ . We set up of a large number of trials, with hopefully a big crowd of monkeys, and after each trial we assess whether the assignment is consistent with  $\mathcal{C}$  and store it into memory. The ones that are not consistent we throw out, so that we have a list  $\mathcal{L}$  of recorded candidate assignments. Since the monkeys have no particular ‘axe to grind’, as Sivia says, we can be confident that they really are being arbitrary in choosing where to aim. And after a large number of trials, some assignments  $l \in \mathcal{L}$  will turn out to be more common than others. Intuitively, the preferred solution will be the assignment  $\hat{l}$  that shows up the most in our list. So the monkeys are helping us build an assignment that is, not only ‘as random as can be’, but also optimal in the sense that we can determine which one is the most frequent.

If there are  $m$  boxes in the grid, and  $n$  balls are handed to the team of ball-scattering monkeys, then there will be a large number possible ways for them to create an assignment ( as long as  $n \gg m$ ). For each trial, the frequencies of ball counts  $n_i$  in each box yields a histogram  $\{p_i\}$ ;  $p_i = n_i/n$  and  $n = \sum_{i=1}^m n_i$ . Since

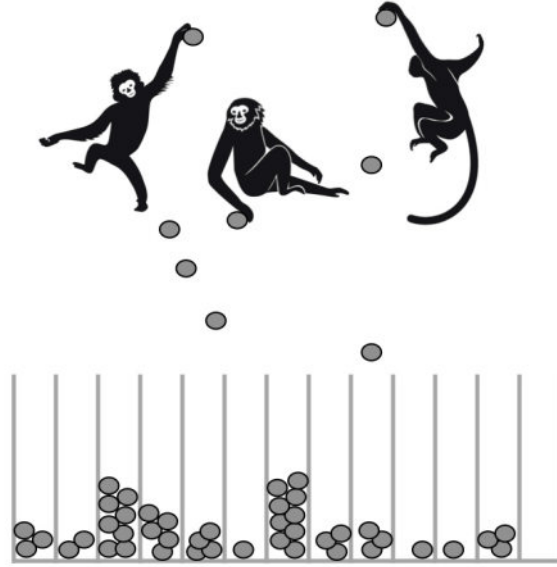


Fig. 2.6 Maxent monkeys

the assignments are distinct permutations of a multiset of  $m$  distinct elements, the number of possible assignments is given by the multinomial coefficient, which tells us the number of histograms that can be created by the monkeys from the  $n$  balls. So, for example, if we had sets of size five that could be built from the letters  $a, b, c$ ; they would look like  $\{a, a, a, b, b\}$  or  $\{a, b, a, c, b\}$ . And their histograms:  $\{3_a, 2_b, 0_c\}$  and  $\{2_a, 2_b, 1_c\}$ . The number of possible histograms  $F$  is hence given by the multinomial formula in equation 2.7 below:

$$F(\{p_i\}) = \binom{n}{n_1, n_2, \dots, n_m} = \frac{n!}{n_1! n_2! \dots n_m!} = \frac{n!}{(np_1)! (np_2)! \dots (np_m)!} \quad (2.7)$$

Taking the logarithm of  $F$  and using Stirling's approximation  $\log(n!) \approx n \log[n] - n$ , we obtain:

$$\log(F) = \log \left[ \frac{n!}{(np_1)! (np_2)! \dots (np_m)!} \right] \approx -n \sum_{i=1}^m \log[p_i] p_i = nS \quad (2.8)$$

where  $S$  is the Boltzmann-Gibbs-Shannon (BGS) entropy. Since  $S$  is monotone in the expected frequencies of the distinct histograms, we can use it as a variational principle for finding out which one will be the most common. This is then another route for showing that the use of the entropy functional makes for a good objective function to use in a constrained optimization problem, when we're interested in obtaining the *combinatorially* most likely distribution of a random variable, subject to a set of constraints  $\mathcal{C}$ .

### 2.3.5 Probability boards

The monkeys will not be happy about the derivation of a variational principle, because it puts them out of work. And perhaps there's a reason why some people choose not to stop at the monkey cage when they go to the zoo: they can be difficult to work with despite their probabilistic reliability. But it's worthwhile taking a moment to compare their ball-scattering scheme to the functioning of Galton's board, which is displayed below in figure 2.7.

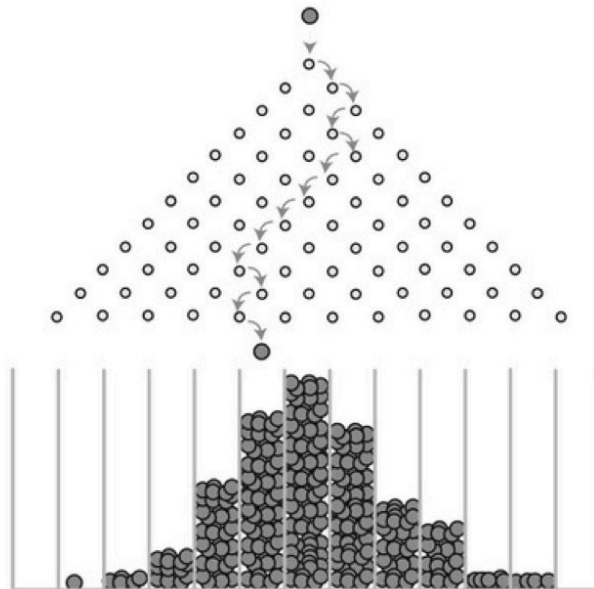


Fig. 2.7 The Galton board.

Galton was fascinated by the demonstration that the normal bell curve could be shown to emerge by dropping a large number of beads into a vertical board with interweaved rows of pegs. The beads, similar to the ball-scattering scheme, are chaotic and bounce off the pegs arbitrarily as they fall. However, they are not free to enter anywhere at the top. The beads enter at the center, and the board also has a fixed width. Thus, one should expect the mode of the histogram of beads to be at the center as well, and since the bouncing down of the beads is effectively random, they should fall symmetrically around that mode. The bell curve arises due to the chaotic movement of the beads, but more crucially, it is the work of the constraints which give it its distinct shape (unimodal-symmetric with a fixed variance).

Maximizing the entropy functional  $S[\{p_i\}]$  subject to specific mean and variance constraints, as in the Galton board, also yields a Gaussian distribution. The comparison has meaning beyond a practical demonstration of the principle, it also points to the fact that the Galton board is a model of *variance* and chance fluctuations around a central tendency. In a physical board or simulation, other distributions can be obtained by altering the shape of the beads or the board's structure, e.g. by placing two holes at the top to create a bimodal distribution.

What the information-theoretic perspective tells us is that it is enough to specify a set of constraints and their expected values in order to recover the full shape of a distribution. And this is what makes maxent, in a somewhat caricatural but practical sense, an adjustable pegboard for building probability distributions. The comparison of these two examples should also point to the fact that often, in the absence of tractable variational principles, we might have to rely on simulators and samplers to obtain the required solutions.

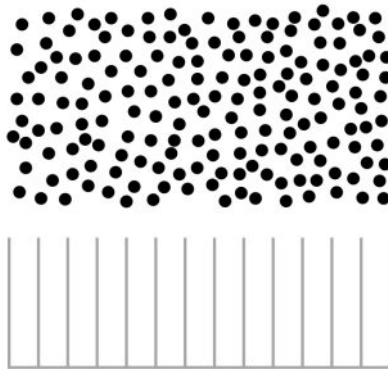


Fig. 2.8 A (*biased*) maxent board.

### 2.3.6 Critical points

In 1957, E.T Jaynes published two important papers proposing the maximum entropy principle (MEP), in which he showed that it is possible to interpret statistical mechanics in terms of information theory. This work successfully synthesized the ideas of Laplace, Bayes, Boltzmann, Gibbs, Jeffreys, Cox and Shannon, into a coherent framework for probabilistic reasoning and data analysis. Jaynes' synthesis led to the so called 'subjective turn' in statistical mechanics, in which the latter came to be viewed by supporters of the MEP as a general form of statistical inference. This newly founded perspective sparked a debate between 'objectivists' and 'subjectivists', both in statistical physics and Bayesian probability theory. Jaynes was a staunch defender of the Laplacian-Bayesian interpretation of probability, and of the maximum entropy method, but he was also a working physicist who made important contributions to quantum electrodynamics (QED) and to radiation theory. The goal of grounding quantum physics within the broader purview of information-theory and principles of logical consistency has historically been, and continues to be, a strong motivation for certain physicists to push the boundaries of our thinking about what lies at the intersection of the 'objective', the 'subjective' and the 'computable'. This latter motivation is also what makes Jaynes akin, not only to probability thinkers, but also to scientists such as Von Neumann, Lorenz, and David Deutsch, whose work has made it possible to think rigorously about the problem of reconciling theories of computation and numerical approximation with natural laws<sup>3</sup>.

Naturally, the idea of interpreting statistical mechanics in terms of information theory was not met without strong criticism. I will not deal here the physics side of the critiques to the MEP; the major arguments and debates can be found in Dias and Shimony (1981); Uffink (1995). I will focus here only on three major critical points that have been addressed at the MEP (*maxent*), and which I think are relevant for the discussions broached in this paper.

#### Insufficient reason and homogeneity

The first critical point has to do with the 'non-committal' aspect of *maxent*, and its relation to the historical *principle of insufficient reason*. In order to properly frame the critiques, I quote below Jayne's original (verbal) formulation:

<sup>3</sup>See Deutsch (1985) for his proof that led to the modern theory of quantum computation.



The principle of maximum entropy may be regarded as an extension of the principle of insufficient reason (to which it reduces in case no information is given except enumeration of the possibilities  $x_i$ ), with the following essential difference. The maximum-entropy distribution may be asserted for the positive reason that it is uniquely determined as the one which is maximally noncommittal with regard to missing information, instead of the negative one that there was no reason to think otherwise. Thus the concept of entropy supplies the missing criterion of choice which Laplace needed to remove the apparent arbitrariness of the principle of insufficient reason, and in addition it shows precisely how this principle is to be modified in case there are reasons for 'thinking otherwise'. (Jaynes, 1957)

It becomes clear reading Jayne's statement that any attempt to justify the procedure should make sure to delineate it from the historical principle of insufficient reason, so that, as Shore and Johnson wrote, the entire effort doesn't come across as an attempt to suggest that it is "a reasonable and systematic way to throw up our hands" (Beneš, 1965; Shore and Johnson, 1980).

The principle of insufficient reason (*pir*) was first articulated by Jakob Bernoulli, which states that if we are ignorant of the different ways that an event can occur, and we do not have a compelling reason to believe that anyone way will occur preferentially over the other, all events are to be regarded as being equally likely. Part of the controversy that ensued, and which persists to this day, has to do with this baggage carried from Bernoulli's inheritance. What poses a problem in Bernoulli's enunciation is the 'equally like' part, as it implies a modern reading of the inferential norm as the 'principle of the uniform distribution'. Keynes was famously opposed to Bernoulli's *pir*, renamed it the *Principle of Indifference*, and devoted an entire section of his *A Treatise on Probability* to challenge its plausibility and logical coherence;

If every probability was necessarily either greater than, equal to, or less than any other, the Principle of Indifference would be plausible. For if the evidence affords no ground for attributing unequal probabilities to the alternative predications, it seems to follow that they must be equal. If, on the other hand, there need be neither equality nor inequality between probabilities, this method of reasoning fails. (Keynes, 1921)

The issue that is at stake here is that of the metric on which the maxent probability assignment is built. Consider, once again, the example of the ‘maxent probability board’, and imagine that instead of having equally spaced bins waiting for the beads at the bottom, your board came with bins of different sizes, with some maybe complicated but nonetheless determinable ratios. What the maxent principle says is that in the absence of any further structure, the histogram  $p(x)$  that expresses complete ignorance about the possible values of  $x$  is given by the ratios of the variable bin widths. In this atypical and idealized example, the maxent assignment is then proportional to the measure structure of the board floor. If we call that non-homogenous measure  $m(x)$ , then  $p(x) \propto m(x)$ .

Considerations of how to translate the maxent procedure to variable metrics allowed Jaynes (1963) to extend the principle to the continuous setting, by modifying the entropy functional so that it takes into account the problem’s underlying geometry:

$$S_* = - \int p(x) \log \left[ \frac{p(x)}{m(x)} \right] dx \quad (2.9)$$

This is known as the Shannon-Jaynes entropy in the literature. In practice, unless there really is an unwieldy geometry to worry about, the differential entropy or the kullback-leibler divergence (with respect to uniform density) can be used without safeguard in the continuous setting to arrive at the preferred solution (see Dimitrov, 2007; Golan, 2018).

### Bayesian updating

Some supporters of maxent, Skyrms notes, "go so far as to give it the status of a principle of Bayesian logic, on a par with additivity of probability or Bayes’ rule of conditioning. Some of its detractors claim that it is almost inconsistent with Bayesian methodology. Much of the debate appears to proceed on the assumption, tacit or explicit, that maxent is an inductive rule, i.e. as a rule for updating subjective probabilities" (Skyrms, 1987). For critiques of maxent understood as an updating rule, see: Friedman and Shimony (1971); van Fraassen (2011) .

The inductive-bayesian interpretation of the maxent rule was in part driven by the introduction of the Shannon-Jaynes entropy discussed in the previous section. The

fact that the maxent optimization procedure can be cast as the problem of minimizing the distance between two distributions was also a contributing factor in making this view possible (Kullback and Leibler, 1951).

An inductive bayesian agent updates prior probabilities by using bayes' rule:  $P(H | D) = \frac{P(D|H)P(H)}{P(D)}$ . The rule defines a computation, since updating is a form of dynamics on the underlying probability space. A maxent updater would move from the prior distribution  $P(H)$  to a posterior distribution  $P(H|D)$ , not via Bayes, but by finding  $P(H, D)$  via constrained optimization, and then evaluating  $P(H|D)$ . The only simple way to get out of this historically convoluted discussion is to realize that the idea of maxent as a form of conditioning is a conflation in terms. The entropic variational principle is used to specify the probability space on which bayesian computation takes place. Thus, from the researcher's point of view, maxent enters concretely at the level of solving a *likelihood dilemma* (what probability model to choose?), in a way that is consistent with a set of knowledge-based or empirical constraints. See Skyrms (1987) and Giffin and Caticha (2007) for historical discussions and more recent attempts of casting the maxent principle as a form of 'minimal information conditioning'.

### The search for rigid priors

In decision theory, mathematical analysis shows that once the sampling distributions, loss function, and sample are specified, the only remaining basis for a choice among different admissible decisions lies in the prior probabilities. Therefore, the logical foundations of decision theory cannot be put in fully satisfactory form until the old problem of arbitrariness (sometimes called "subjectiveness") in assigning prior probabilities is resolved. (Jaynes, 1968)

The plausibility of Jaynes' deterministic hypothesis, that different people following an identically priored bayesian inference scheme will reach identical conclusions when presented with the same data and information, hinges on the existence of a canonical uniform and 'rigid' prior. For the maxent principle, this seems to come at the expense of putting homogeneity and uniformity over 'non-committalness' with respect to missing information. What defines a state of 'complete ignorance'? Attempts to give an answer to this question led to important formal explorations

on the part of Jaynes. To let the probability assignments be ‘as random as can be’, one should be willing to end up with an assignment of the form  $m(x) = \frac{\mu}{dx}$ . To take the board example again, the latter expression simply means that the proportional sizes of the bins catching the beads at the bottom vary as a function of the possible outcomes. The Shannon-Jaynes entropy functional is able to recover that assignment for Lebesgue measurable spaces. For other more extreme situations, if ‘non-committalness’ is to be preserved, new mathematical explorations might be called for. And not in the sense of asking for a measure that is ‘good for all purposes’, but in finding one that is appropriate to specific scientific problems and contexts. The path taken might involve not only revising what is meant by uniformity and ‘uninformativeness’, but also by independence.<sup>4</sup>

Certainly, there are both normative and formal components to the maxent principle. The normative part is related to the desideratum to remain non-committal and to let the probability assignments be ‘as random as can be’. The formal part has to do with the choice of the variational concept to be used in the constrained optimization problem. The good balance between the two is struck when the choice of the functional  $S$  is governed by the willingness to find ways to remain non-committal with respect to missing information in any given context, and for solving a specific problem.

### Non-exponential families and varieties of entropy

Of all the challenges that can be faced by classical maxent methods, the most piercing are those which point to issues with the optimality of the variational principle, or which question the relevance of the BGS entropy concept itself. Is there something lacking in a framework that always yields exponential solutions ? Can we use the method to infer bi-modal distributions ? What about distributions with power-law tails ? The question of how a probability assignment behaves at the tails of possible outcomes is very important, because if you’re trying to obtain a solution which is ‘as random as can be’, you don’t want to rule out extreme but likely events. The method’s ability to recover unimodal distributions which are highly peaked might give the false impression that it is generally applicable to all models,

---

<sup>4</sup>A case in point of this kind of situation was the formalization of free probability and quantum probability, where not only measures change, but also the classical axioms of the probability calculus are altered. See Deutsch et al. (2000).

but maximizing the BGS functional can lead to poor results if the data is multi-modal or heavy-tailed. Exponential assignments will effectively give vanishing and possibly inaccurate probabilities at the very far end of the range of possible outcomes;  $e^{-\text{very extreme event}} \approx 0.0$  (see Bryson, 1974). The solution could involve employing a mixture of exponential models, but that might imply that the constraints inferred are not efficient with respect to a single BGS entropy functional. So can one hope to define informational efficiency in such a model ?

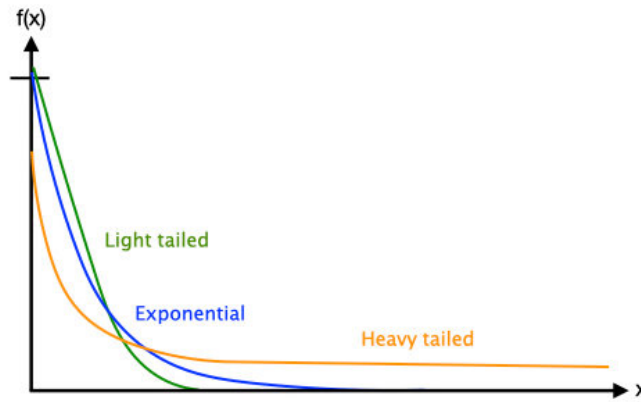


Fig. 2.9 Heavy-tailed distributions. Exponential tails go to zero faster than heavy tails.

A vast literature has emerged in recent years in statistical mechanics that focuses on tackling this issue. Nonexponential distributions can be inferred by maximizing another form of entropy, as opposed to the BGS form. Among these nontraditional entropies are those of Tsallis, Rényi, and others (see Aczél and Kannappan, 1978; Amari, 1985; Rényi, 1961; Tsallis, 1988). The Tsallis entropy gives a mathematical generalization of the BGS entropy, and takes the following form:

$$S_q(\{p_i\}) = \frac{k}{q-1} \left( 1 - \sum_i^W p_i^q \right) \quad (q \in \mathbb{R}; S_1 = S_{BG}) \quad (2.10)$$

The  $q$  parameter is an 'entropic index' which allows for that generalization, and subsumes the BGS functional under a broader (infinite) class;  $\lim_{q \rightarrow 1} S_q = S_{BG}$ .

It's important to note that Tsallis' motivation for developing this new entropy functional also stems from the need to uncover a variational principle suitable for inferring probability assignments. But the point of view from which this is done departs radically from the Jaynesian program, and poses an important challenge

to its central motivations and justifications (which were discussed in the previous sections). For readers unfamiliar with the maxent literature, their distinction could be easily missed by the fact that both programs are about entropy maximization. The key to their divergence is succinctly phrased by Peterson et al. (2013) in the following quote:

"A nonexponential distribution that is derived from a Max Ent principle requires that there be nonextensivity in either an energy-like or entropy-like term; that is, it is nonadditive over independent subsystems, not scaling linearly with system size. Tsallis and others have chosen to assign the nonextensivity to an entropy term, and retain extensivity in an energy term".

To get a more intuitive sense of what this entails, it's worth going back to the caricatural maxent board example, or to Jayne's monkeys. What Tsallis proposes, from the point of view of choosing/building a maxent probability assignment, is to tweak the beads and the pegs of the board. As for the Jaynes example, the Tsallis program suggests that when studying complex systems which exhibit *weak chaos* and important long-range interactions, we should be hiring a different team of monkeys (Tsallis and Brigatti, 2004). A second, more faithful way to portray the issue: The Tsallis program opens up a catalog for choosing the maxent board, or the monkey team, out of a possibly infinite variety. The Tsallis version of these examples is displayed below in figure 2.10.

Tsallis' proposal for 'tweaking the pegs of the board', or 'picking monkey teams from a league', is motivated by the objective of supplying dynamical-law foundations to the statistical mechanics of non-equilibrium and strongly interacting complex systems.

It is the natural (or artificial or social) system itself which, through its geometrical-dynamical properties, mandates the specific informational tool—entropy—to be meaningfully used for the study of its thermostatistical and thermodynamical properties. (Tsallis and Brigatti, 2004)

This point of view is established by Tsallis on physical grounds, as it stipulates that the  $q$  parameter should be determined *a priori* from microscopic dynamics. For

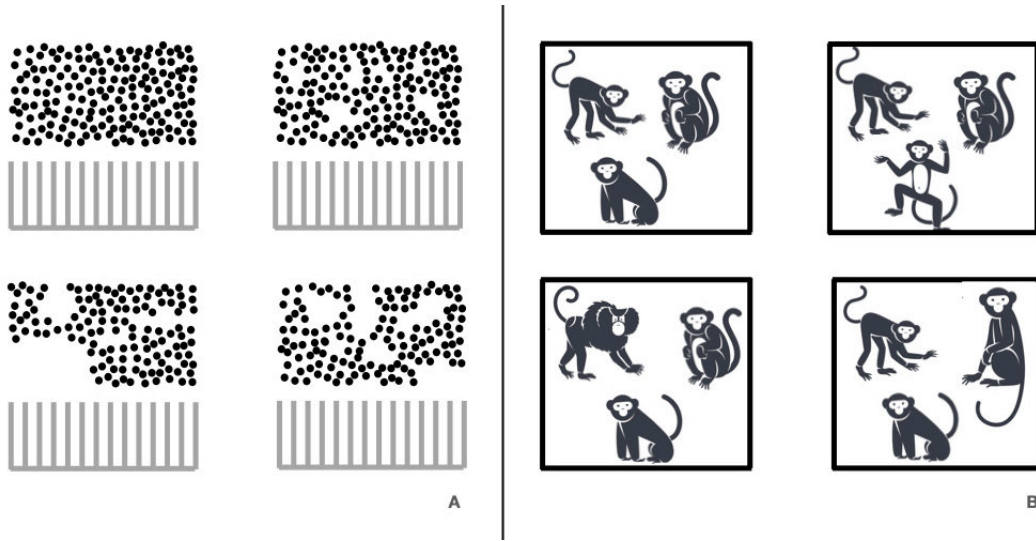


Fig. 2.10 (A) *Some weakly chaotic maxent boards.* (B) *The job market for entropic monkeys.* Which  $q$ ?

dynamicists interested in modeling non-equilibrium systems through anomalous (e.g. non-markovian) diffusion processes that converge to power-law distributions, the theory is more than compelling; it's also beautiful. Tsallis' nonextensive statistical mechanics has been used to study the fat-tailed data of options pricing, financial volumes and returns, risk aversion, among other economic phenomena. In physics settings, it has been applied to phenomena as diverse as solar winds, high energy particle collisions, black holes and quantum gravity (see Beck, 2009; Beck et al., 2006).

But the point that I want to discuss here is the extent to which this program seems to pose a direct challenge to the information-theoretic foundations of the Jaynesian ('classical' maxent program) in social and biological contexts. There is a reason to why Shannon loses credit in Tsallis' limit ( $\lim_{q \rightarrow 1} S_q = S_{BG}$ ). For Tsallis, it's important to be able to find a variational principle for inference which does not tamper with the laws of thermodynamics, and which retains the extensivity and conservation of energy. A crucial point to keep in mind is that physicists have strong and justified reasons for seeking solutions that are fully consistent with a dynamical-law approach, and to consider the idea of non-extensive energies as blasphemous. After all, it's the fact that Tsallis carves energy in stone, but lets entropy take a parametric form, which makes the program so appealing to swaths of physicists working in the most esoteric confines of non-equilibrium statistical mechanics.

But from the point of view of the Jaynesian statistical mechanic, the introduction of the  $q$  parameter runs at odds with the principle of letting things ‘be as random as can be’ and with the desideratum of remaining non-committal with respect to missing information. If the microscopic dynamics are known, then it can be said that nature ‘chooses’  $q$ , but what happens in contexts where the dynamics are not known or where we have strong reasons to believe that the deepest knowledge we can gain about a system is not dynamic but combinatorial and information-theoretic (e.g. in genetics)? The prospect of calibrating  $q$  or estimating it from the data seems to impose too much structure and to run against the logical consistency requirements established by (Shore and Johnson, 1980).

In the classical maxent program, if there is anything to be adjusted, it should be the energy functional. In social and economic settings, for example, there is no intrinsic reason to assume that quantities should be conserved or energies extensive (see Scharfenaker, 2020a). It follows that the question of how to modify the energy functional within the limits of BGS is the central and nontrivial data modeling challenge faced by those working with the classical maxent approach. It is in holding this position that the mining of probabilistic descriptions which are informationally efficient and non-biased becomes possible.

When considering how to address the issue of nonexponential and power law assignments, or when dealing with long-tailed skew in the distributions, other routes are possible which do not involve using non-extensive entropies. For nonexponentials, one route is to work with the logarithms of probabilities and to employ a nonextensive energy term. This is the strategy adopted by (Peterson et al., 2013) in deriving distributions with power-law tails from the BGS functional. They report obtaining good fits for a broad range of phenomena, such as links in social networks, protein-protein interactions, and terrorist attack severity.

When long-tailed skew seems hard to account for using a standard maxent procedure, another strategy, recently explored by Scharfenaker and Foley (2021a), is to discard the mean as a sufficient statistic for recovering the distribution. This effectively drops the assumption that the system’s quantity is ‘conserved’, and allows instead for a complex nonlinear energy functional to determine the predictive mean. This strategy turned out to be particularly important for the empirical work presented in this dissertation, as it allowed me to account for the skew in the data in terms of two distinct but interacting forces; residential sorting and competition in the local



public goods setting. Such a move would not have been ‘legal’ in the Tsallis program, and maximizing a nonextensive entropy would have led to a ‘market feedback energy’ difficult to interpret in economic terms, due to the presence of hyper-parameters.

From the perspective of analyzing the data, it could be worthwhile to compare the two approaches and see what insights could be gained. The Tsallis program provides a new plethora of anomalous and non-markovian diffusion processes that can meaningfully complement a classical maxent analysis, by providing richer dynamical prototypes and samplers. The perspective I have tried to outline throughout this paper is that it is possible to use such complements without being committed to the ontological implications of the dynamical-law perspective.

## 2.4 Boltzmann densities and microeconomic energy functionals

### 2.4.1 An integrated perspective for economic statistical mechanics

Having discussed what I think are the most relevant and important foundations for the application the Boltzmann-Gibbs-Shannon (BGS) variational principle in economic data analysis, in this section I propose a simple integrated perspective that brings all the elements together. Let me briefly recapitulate some key points.

The application of statistical equilibrium methods and statistical physics, in the context of social and biological systems, can be approached by appealing to the concepts of self-organized criticality and punctuated equilibria. While these concepts remain highly speculative, concrete evidence for their plausibility can be found by looking at the evolution of empirical histograms of complex social systems data. Data analysis can reveal time invariance in the distinctive statistical signatures of economic variables, thus opening the door to statistical equilibrium reasoning.

The BGS variational principle can be used to uncover or mine the system’s ‘informationally efficient’ energy functional. But it can also be used to hypothesize new energy functionals and to test their empirical plausibility. In both cases one is trying to solve an ‘inverse’ problem. The semantics of the *energy* concept employed have to be settled on a context-specific basis, and by appealing to relevant and well-

established social, economic or biological theories. My discussion of the complexity view has been an attempt to open up a philosophically rigorous route for liberating economic energy concepts from strict disciplinary bounds.

The energy functionals recovered can be mapped to well known statistical physics models. This can lead to dynamic analysis as a method for exploring (sampling) possible evolutions of the state variables, or to institutional/structural analysis through the use of agent based models, or even verbal models.

An added bonus that comes from the use of well established statistical physics models, is that they can serve as common currency for mediating complexity concepts across physical, biological, and economic domains. The study of animal models can spark rigorous reflections on processes such as herding, opinion dynamics, and conflict. In the mapping that connects the structural and behavioral patterns of distinct phenomena across scientific domains, what emerges is literally a social-statistical view of the natural world.

### 2.4.2 Approaching the optimization procedure

In considering then the BGS entropy as the unique measure of uncertainty which retains the consistency requirements and non-committalness that have been discussed throughout, the elementary optimization procedure looks as follows for the discrete case. The entropy  $S[p]$  of a distribution  $p(s)$ , giving the probability assignments for  $s \in \Gamma$  possible configurations of a system  $T$  is:

$$S[p] = - \sum_{s \in \Gamma} p(s) \log p(s) \quad (2.11)$$

If  $S[p] = 0$ , we can expect a single configuration  $s$  to repeat itself indefinitely. The system is ‘minimally surprising’. Conversely, at maximum entropy and in the absence of any structure on  $S$ ,  $T$  is ‘maximally surprising’. BGS entropy measures our ignorance about  $T$ , not how chaotic it is. In considering the constraints and structure that we can impose on  $\Gamma$ , we start by asking: How is the system measured ? How is it observable ? Usually we will not be dealing with raw measurements of the states  $s$  but with a set of functions or features  $f_k(s)$ . These can be, for example, assigned spins to the particles of a ferromagnet, to the firing of neurons in a temporal bin, or to the entry/exit decisions of firms in some specific economic environment.

What is of interest to us is usually a set of moments  $\{\langle f_k(s) \rangle\}$ , where the  $f_k$  are organizing functions which we use to reduce our uncertainty about  $T$  ( $\langle * \rangle$  is the expectation operator).

With this in mind, one way to motivate the procedure is to consider it as the concrete task of forming expectations about the values that the features  $f_k(s)$  can take. So we form a set of expectations  $\mathcal{E} = \{\langle f_k(s) \rangle\}$ .

The  $k$  expectations formed will naturally depend on our probability assignment  $p(s)$ , as follows:

$$\langle f_k \rangle = \sum_{s \in \Gamma} p(s) f_k(s) \quad (2.12)$$

Prior to working with the data, we can consider  $\mathcal{E}$  as taking on a set of placeholder values  $\{\mathcal{E}_k\}$  in order to optimize the BGS functional and find an informationally efficient  $p(s)$ . The standard procedure is to use the method of Lagrange multipliers, and set up the Lagrange functional  $\mathcal{L}$ :

$$\mathcal{L}[p] = - \sum_{s \in \Gamma} p(s) \log p(s) - \sum_{k=1}^K \lambda_k (\mathcal{E}_k - \langle f_k \rangle) \quad (2.13)$$

In solving for the fixed point by taking the partial derivatives with respect to  $\lambda_k$ , the resulting probability assignment  $p(s)$  is the ubiquitous Boltzmann distribution:

$$p(s) = \frac{e^{-H(s)}}{Z} \quad (2.14)$$

The entropy is a convex function of  $p$ , and the constraints are linear in  $p$ , so the optimization problem is convex and the solution is unique. The denominator is the normalizing constant or partition function  $Z = \sum_{s \in \Gamma} e^{-H(s)}$ . By taking the relative negative log-likelihood of  $p(s)$ , we obtain an informationally efficient 'energy' or Hamiltonian for  $T$ :

$$H(s) = - \sum_{k=1}^K \lambda_k f_k(s) \quad (2.15)$$

Apart from belonging to the exponential family, there's nothing peculiarly idiosyncratic about the distribution  $p(s)$ . Nonetheless, we should ask: How is this informationally optimal? How is this not an ad hoc solution? It's not obvious.

A phenomenon seldom discussed in statistics courses is our ability to build expectations about measurable quantities in order to efficiently process available information. The expectations  $\mathcal{E}_k = \langle f_k \rangle$  are the *minimally sufficient statistics* required to build a predictive distribution  $p(s)$ . They are 'sufficient' in the sense that measurement of any other average will be irrelevant and superfluous in our ability to recover the full distribution  $p(s)$  (see Porta Mana et al., 2017). For example, if we have some variable  $x$ , knowledge of a sample mean  $\bar{x}$  and its variance  $\sigma_{data}^2$  is enough to recover the full shape of the normal bell curve, under the assumption that we can interpret the data as coming from that model. So it's not only a question of obtaining sufficient statistics, but of situating the signal in the appropriate probability model, i.e. attributing meaning to a collection of data points.

The reason this isn't obvious is due to the semantic component, but also because we can recover the full distribution even if those are the only *two quantities* we have. So if you share with a friend your  $\bar{x}$  and  $\sigma_{data}^2$ , she'll be able to predict the shape of your normal dataset, and compute other statistics (for her own reasons), without you having to share the entire dataset. So there's a strong element of data compression and signal-processing.

What fully specifies  $p(s)$  is knowledge of the parameters  $\lambda_k$ . The optimal values are those that make our expectations  $\mathcal{E}_k$  as close as possible to the observed averages, taken over  $R$  samples in a dataset  $\mathcal{D}$ :

$$\langle f_k \rangle_{\text{data}} = \frac{1}{R} \sum_{s \in \mathcal{D}} f_k(s) \quad (2.16)$$

In finding the optimal  $\{\lambda_k\}$ , we are trying to minimize a divergence between our predictive distribution  $p(s)$  and the empirical distribution  $p_{data}(s)$ . This can be done via the Kullback-Leibler functional:

$$D_{\text{KL}}(p_{\text{data}} \| p) = \sum_s p_{\text{data}} \log \left( \frac{p_{\text{data}}(s)}{p(s)} \right) \quad (2.17)$$

The KL functional is minimized at the point where our expectations  $\mathcal{E}_k$  match the data (see Lee and Daniels, 2018):

$$\frac{\partial D_{\text{KL}}}{\partial \lambda_k} = \sum_s p_{\text{data}}(s) \frac{\partial (-H(s) - \log Z)}{\partial \lambda_k} = 0 \quad (2.18)$$

$$\Rightarrow \langle f_k \rangle_{\text{data}} = \langle f_k \rangle \quad (2.19)$$

We can then propagate uncertainty into our solution by using bayesian methods and MCMC sampling. Apart from recovering the optimal set  $\{\lambda_k^*\}$ , which is also the set of maximum-likelihood (MLE) estimates (see Golan, 2018), we might want to predict distributions which take input from other regions of the parameter space. We can't do that with a single parameter estimate, so we need distributions for the parameters;  $p(\lambda_k)$ . Hence, the complementarity of bayesian methods is crucial.

### 2.4.3 Principle-based search

In these last two sections I discuss the question of how to establish economic semantics for the inferred energy functional or 'economic Hamiltonian'. To start, the first thing to note is that the energy is a linear combination of the organizing functions  $f_k$ ;  $H(s) = \sum_{k=1}^K \lambda_k f_k(s)$ . This does not imply, however, that the organizing functions should be linear. They can be nonlinear functions of the state  $s$ , e.g.  $\tanh(s)$ .

Furthermore, there are instances in which we may want to consider how another system  $B$  interacts with the system in question, but we have no way of measuring directly that interaction. So  $B$ , for example, can be a microeconomic or decision-theoretic system, and we want to know how different configurations  $b \in B$  might contribute to the shape of the inferred distribution  $p(s)$ , which may refer to the probabilities of a macroeconomic state variable. If we have a working theory about  $B$ , one way to do this is to hypothesize a function  $b(s)$  in our set of expectations  $\mathcal{E}_k$ , and to optimize the joint entropy of both systems. This is the strategy followed by the quantal response model used in this dissertation (Scharfenaker and Foley, 2017). If the relationship between microeconomic configurations and the macroeconomic states were linear, we would be able to fit a regression line through a scatter plot of their features or their related parameters. But since this is usually not the case,

one way to study the complex coupling of agent systems and macro variables is to find an energy functional of the form  $H(s, b(s))$ , where  $b(s)$  might be some microeconomic nonlinear function which we use to organize our expectations about  $s$ . It is this recursive aspect which can make the inference of suitable statistical physics prototypes more challenging in macroeconomic applications. For a thorough discussion of the behavioral model used in the quantal response (QRSE) model see Foley (2020).

The key question is then that of determining what functions  $f_k(s)$  can be used to form expectations about a system's possible configurations  $s$ . The fifth chapter in this dissertation explores the question of whether reasoning in terms of general principles can provide a basis for predicting economic distributions. The main outcome of this exploratory philosophical project has been the realization that the same discussion can be broached in terms of model-based inference in cognitive neuroscience (Friston, 2009; Gottwald and Braun, 2020; Khemlani and Johnson-Laird, 2013), or in terms of semantic theories of information (Floridi, 2005).

To further motivate why the question of semantics and principle-based inference is important for statistical equilibrium reasoning, consider two simple examples: *playing football* (soccer) and *avoiding fire*. A player who scores a header during a corner kick has the remarkable ability of predicting the ball's trajectory and spin as it swerves through the air. Talented players can predict the trajectories with great accuracy from information of initial conditions (the first few seconds of the corner kick). This ability comes from years and years of training.

Now consider the situation in which a hiker infers the distal presence of fire from the proximal information that there is smoke (Sequoiah-Grayson, 2007). What allows the hiker to infer the presence of fire is the knowledge that "smoke means fire", her mental map of the forest, as well as a basic qualitative understanding of what smoke looks like (a gray and chaotic looking suspension of particles). Crucially, this is not an inferential process that involves reasoning in terms of aerodynamics or initial conditions. So clearly there are contexts where dynamic expectations may not reveal any meaningful information about an unfolding process.

This example is revealing because it helps to situate the applicability of principle-based statistical equilibrium reasoning in contexts where the use of dynamicist concepts might be misleading. The point is not to say that dynamic models are

uninformative, but to emphasize that principle-based search can be a good starting point for finding the right context in which to situate plausible and relevant dynamics.

#### 2.4.4 Mining for energy

There are other scenarios in which a mapping for the relevant statistical physics prototype may be more readily available, or in which it will can be obtained as a solution to an inverse problem by imposing only relevant statistical constraints on the procedure.

In Lee and Daniels (2018), for example, it is shown how solving an inverse maxent procedure can yield the energy functional for an Ising model. Consider a set of binary action vectors or spins  $\{s_i\}$  with two possible values: e.g. up/down, (1,-1), or entry/exit. Take as the set of constraints the mean action and the pairwise correlations;

$$\langle s_i \rangle = \langle s_i \rangle_{\text{data}} \quad (2.20)$$

$$\langle s_i s_j \rangle = \langle s_i s_j \rangle_{\text{data}} \quad (2.21)$$

Via the Lagrangian optimization procedure,

$$\frac{\partial \mathcal{L}[p]}{\partial p(s)} = -\log p(s) - 1 + \sum_{i < j}^N J_{ij} s_i s_j + \sum_i^N h_i s_i \quad (2.22)$$

,we obtain  $p(s) = e^{-H(s)}/Z$ , where

$$H(s) = -\sum_{i < j}^N J_{ij} s_i s_j - \sum_{i=1}^N h_i s_i \quad (2.23)$$

This is the well studied Ising model in statistical physics, which is also known in the literature as the pairwise maxent model (Ising, 1925). If we drop the mean action constraint, we obtain the Heisenberg model used in Bialek et al. (2012). In general, the  $h_i$  parameter measures the strength of an external field acting on  $s_i$ , and  $J_{ij}$  measures the pairwise coupling strength. So  $h_i$  gives the ‘intrinsic’ tendency of

each spin or action vector to maintain a particular direction, and  $J_{ij}$  tells us about the strength of their coordination.

In social science applications, the multinomial and interactive character of Ising-type model lends itself well for studying highly correlated decision-theoretic environments. Lee et al. (2015), for example, derive an Ising spin glass model from the observed pairwise correlations among supreme court justices' votes. Stauffer (2008) discusses similarities between the Ising model and Schelling's famous segregation model. They also make the important point that it is only until recently that we have come to observe the strong similarities between models of ferromagnetism in physics and models of residential segregation and community formation in analytical sociology.

Also inspired by the applicability of interactive physics prototypes, Vinković and Kirman (2006) study a surface tension model that explains clustering dynamics in Schelling type segregation. While not motivated from a maxent or information-theoretic perspective, one can see the complementarity between the approaches. Sylvain Barde, for example, develops a 'maxent prediction' of Schelling segregation patterns by setting up the inference problem as a bayesian image restoration task (Barde, 2015). In a nutshell, the approach taken by Barde is to take the emergent segregation pattern as the true image which is hidden by noise and to predict the outcome from minimal information about initial conditions. In another paper, Barde also builds a 'maxent prediction' for Kirman's Ant model, in which the limiting distribution of the asymmetric recruitment of ants into two food sources is given by a beta distribution (Barde, 2012). The drawback to Barde's approach is that it's not motivated by the search for an informationally efficient representation of the underlying process. In the context of herding models, it could be interesting for future work to study what set of parsimonious constraints or principles could lead to the beta distribution, either by optimizing a BGS functional or other varieties of entropy.

With respect to the research presented in the next two chapters (papers), the appropriate statistical physics prototype sought would be one capable of capturing the complex and evolving process of jurisdictional fragmentation in the US. In the US, the system of local governments is not static and evolves as a function of residential choice and local political preferences. "In 1952, the Census Bureau reported a total of 116,755 local governments. By 2017, 90,075 remained through a complex



system of dissolutions, mergers, and new additions. Nearly 50,000 independent school districts were dissolved or merged over 60 years, while almost 25,000 net special districts were created. General purpose local governments (counties, cities, towns/townships) remain relatively unchanged over time by comparison. The local government landscape is dynamic with some areas seeing very little change and others seeing complete transformations"(Goodman, 2019).

This is really a phenomenon into which we have no adequate visibility at the moment, and which is poorly captured by econometric techniques and applied general equilibrium analysis (see Nechyba, 2020). The Tiebout-QRSE model takes a first step in exploring the possible space of informationally efficient energy functionals. The empirical results point to the need to include/test additional potentials related to the housing market, to federal and state spending, and to the growing share of the private school market. I suspect that a useful starting point for finding prototypes would be to look for models of fracture in physical systems, but this is a question for future research.

## **Chapter 3**

# **A Statistical Equilibrium Model of Public School Expenditures**

### **3.1 Introduction**

#### **3.1.1 Tiebout Competition**

A central problem in the economic analysis of the provision of local public goods is the lack of incentives of voters to reveal their true demand. Tiebout (1956) proposed to study the problem of local public goods through a quasi-market model in which consumer-voters express their preferences for local public goods by moving in and out of local jurisdictions. The Tiebout hypothesis states that local jurisdictions will tend to sort into homogenous blocks with respect to demand for local public goods and tax levels, when these are taken to be a form of prices in the model. The core idea behind this hypothesis is that a Tiebout sorting equilibrium, if it exists, will eliminate inefficiencies associated with demand diversity; households will not be forced to pay higher tax levels than they would otherwise prefer, nor are they able to free-ride on neighboring households' relatively higher contributions to the local tax-service package.

Tiebout's 1956 paper was, and continues to be, an important catalyst for renewed research in the analysis of decentralized government finance. Tiebout's major contribution was to challenge the standard belief of the time that there was no market-based solution to the problem of local public goods provision. He did so by

placing geographic location and mobility at the core of the analysis, and by using the latter as a proxy for choice and preference revelation.

There is no way in which the consumer can avoid revealing his preferences in a spatial economy. Spatial mobility provides the local-goods counterpart to the private market's shopping trip. Tiebout (1956)

While Tiebout (1956) agreed with Musgrave (1939) and Samuelson (1954), that the determination of federal government expenditures could only have a political solution, he argued for explaining variations in local government expenditures in terms of decentralized sorting mechanisms (a market analogy) and not by alluding to simple majority voting schemes.

Tiebout's hypothesized mechanism of competition, when seen through the lens of neoclassical equilibrium models, may be understood as having three essential traits (Nechyba, 2020). The first is that when local communities are viewed as analogous to competing firms, decentralization will allow for the optimal provision of public services in the presence of heterogeneous household demands. The second is the notion that competition will reduce incentives for local governments to behave like 'Leviathans' (Jha, 2020). The latter notion rests on the belief that the decentralized procurement and provision of local public goods will counter the tendency of governments to arbitrarily extract higher taxes from their residents (Brennan and Buchanan, 1980; Jimenez and Hendrick, 2010). A third feature implies that in 'equilibrium', the Tiebout mechanism will lead households to sort (to some degree) on the basis of ability-to-pay and household income. This latter feature is, of course, far from being unequivocally desirable. The characterization of Tiebout sorting as an optimal outcome, possessing intrinsic merit mostly on account of its capacity to bring about productive efficiencies, turns out to be at odds with basic legal notions regarding citizens' rights to education (Jha, 2020). An equilibrium in which public school expenditures and quality are highly correlated with household characteristics presents non-negligible moral and legal challenges. The scope of these challenges has been duly evidenced by the continued legal battles and policy debates over funding inequities in the US public education system for the past 50 years (Baker, 2021; Hertert et al., 1994). The fact that the optimal outcome in a highly idealized formulation of the Tiebout hypothesis turns out to be fundamentally at odds with what may be desirable at the policy or household level (or is at the very least highly

contestable), does not rule out the possibility that sorting and the rationing of government resources are in fact shaped by Tiebout-like forces. It does however pose serious challenges to the modeling and specification of the microeconomic primitives which drive the competitive process.

One of the fundamental problems that comes out of the use of applied general equilibrium models is that they force us to consider observed economic distributions as resulting chiefly and mechanically from the interaction of optimizing agents (households and governments) whose preferences are fully satisfied. This is a modeling strategy that rules out a priori the possibility that agents' expectations will remain unfulfilled in equilibrium.

In the context of the economic analysis of the determinants of heterogeneity in public school expenditure levels and demand, where much of the theoretical and policy debates center around the explicit recognition that education markets are structured by complex political and production processes, the requirement that fully optimizing behavior be consistent with observed equilibria is hard to sustain. Furthermore, in the absence of plausible characterizations for the microeconomic and political environments, it is hard to see how any useful insights may be extracted from the study of general equilibrium forces and outcomes.

This concern has steered the Tiebout and education finance literature towards a path of building models of increased mathematical and computational complexity, where elements such as heterogeneous voting preferences and non-financial inputs are incorporated in order to provide richer descriptions that are more empirically relevant, as well as plausible from a microeconomic perspective (Kuminoff et al., 2010; Nechyba, 2003).

There has been a recent shift in the literature from building general equilibrium models to building computational equilibrium models that straddle a wide spectrum covering both purely theoretical and empirically motivated formulations. As Nechyba (2020) notes, all such models start by explicitly specifying the underlying mathematical structure of the economic environment being modeled. That is, they provide a fully structural specification for household preferences, school production functions, distributions for household characteristics in the model (such as income), as well as mathematical descriptions for the political process (voting models), the fiscal environment, and the housing and private school markets. Through simulation studies, the study of the equilibrium outcomes in these models is then expected to

yield meaningful policy insights, and to provide a sandbox for experimenting with out-of-sample policy interventions.

The problem is that the relevance of these simulation studies hinges on the empirical plausibility of the elaborate microeconomic structure that is being used to represent the underlying mechanics of the data generating process, and on the confidence we may have in the model's parameters to adequately capture empirically relevant processes. But if we consider the fact that the task of determining the empirical plausibility of any given model specification for complex social environments with large degrees of freedom may be ill-posed and underdetermined (Scharfenaker and Foley, 2017), then it is hard to see how the route of increasing model complexity in fully micro-founded general (or computational) equilibrium models is likely to yield unambiguous and normatively unbiased results. There is a very broad continuum of models and solutions that are consistent with any set of circumstantial data and evidence (Golan, 2018). And misspecification can show up either at the level of functional forms (production and preference functions), criterion or decision functions, the specification of voting models, as well as priors for stochastic inputs in the model (e.g. household characteristics).

This paper takes an alternative approach that makes use of maximum entropy methods and a statistical equilibrium framework to model and study the effect of competition in shaping the distributions of local government education expenditures for the period of 2000-2016 in the United States. The advantages of this maximum entropy/statistical equilibrium framework are plenty, but a central one that we consider here is that it allows us to study the competitive dynamics of the US public education market (a complex social systems with large degrees of freedom) without having to commit *a priori* to a heavy mathematical scaffolding of the underlying microeconomic environment. Rather, it allows is to study one plausible way in which the probabilistic structure of school district expenditures can be seen to emerge from a pair of parsimonious behavioral and institutional constraints that we place on the underlying microeconomic environment.

### 3.1.2 Unintended Outcomes, Market Efficiency and Tiebout Sorting

One of the interesting aspects of Tiebout's original 1956 formulation is that it remains non-committal with respect to any specific equilibrium model formulation, even if it highlights a set of stylized facts and features that the hypothetical competitive process should meet. But as the history of the empirical tests of the Tiebout hypothesis has shown (Edel and Sclar, 1974; Nechyba, 2020; Oates, 1969), it is not truly possible to test all of the assumptions of the larger Tiebout hypothesis at once without running into contradictions. For example, testing the assumption of residential mobility alongside the capitalization of fiscal variables into housing prices may run against Tiebout's larger efficiency hypothesis (since the presence of capitalization is evidence for the existence of excess demand for housing in the jurisdiction where taxes and local service levels are being capitalized) (Epple and Nechyba, 2004a)>. Similarly, as we pointed above, the existence of Tiebout sorting is to be better understood as a potentially unexpected macroeconomic outcome (to at least some section of households). Seen under this light, the prospect of being able to reconcile the underlying political contradictions of the education market with the assumption of fully maximizing households in a general equilibrium model seems far-fetched. That said, we believe there is need and ample room to focus on some aspects of the Tiebout hypothesis, and that it is possible to study the empirical support for the general claim that expenditures in local public goods are heavily shaped (and at least partially explained) by competitive forces and a boundedly rational arbitrage that takes place at the household level in terms of education consumption.

This paper applies the theoretical framework of the Quantal Response Statistical Equilibrium (QRSE) model developed in Scharfenaker and Foley (2017). As mentioned above, the approach taken by the paper is not fully agnostic with respect to microeconomic structure, as it utilizes an entropy constrained model of residential mobility and jurisdictional choice as the baseline characterization of household behavior. This baseline model makes the behavioral assumption that households try to maximize the rate of return on tax expenditures (considered as prices for local education services), under the constraint of a limited capacity to process market and political signals. In the context of low-income and inner-city households, we put forward the idea that this limited capacity may also be interpreted as a form of restricted economic agency. The basic outline of this behavioral model is very similar

to the one found in Sims' rational inattention program (Sims, 2003). Through the inclusion of an information-theoretic constraint on the utility maximizing program of households, this baseline specification delivers a meaningful probabilistic description of household behavior.

### 3.1.3 Sample and Paper Structure

The statistical equilibrium distribution of the QRSE model presented here is a positively skewed unimodal distribution of the household rate of return for local tax expenditures, with four parameters  $T$ ,  $S$ ,  $\mu$ , and  $\alpha$  that qualitatively predict the observed data and give insights into the possible range of variation across sub-sampling schemes. We use US public education finance data for all school districts in the period of 2000-2016. We then apply Bayesian inference and MCMC sampling to fit the observed distribution for the entire period to the theoretical QRSE model, and to recover posterior distributions for the four unknown parameters.

The paper consists of 5 sections. Section 2 provides a description of the data used and presents empirical the frequency distributions for the key fiscal and expenditure variables used in building the model. Section 3 puts our application of the QRSE model into context by discussing fiscal decentralization, Smith's theory of competition, and the measurable implications of the Tiebout hypothesis. Section 4 then develops the paper's QRSE treatment of Tiebout competition and derives the statistical equilibrium density for the *local per pupil rate of return on tax and service charges*, which we term *educational returns*. Section 5 describes the Bayesian estimation of the model, and discusses results for the four main parameter estimates.

## 3.2 School District Variables

### 3.2.1 Data

This paper uses data from the National Center for Education Statistics' Common Core of Data, the US Census Bureau Small Area Income and Poverty Estimates, and the US Department of Education's EDFacts initiative <sup>1</sup>.

<sup>1</sup>This data has been made available in a harmonized format in a publicly available API by the Urban Institute, which provides a convenient and reliable interface to all the major federal dataset. Education

We consider a sample of local expenditures in primary and secondary education, local taxes, enrollment and population estimates for all 50 US states and school districts (on average  $\sim 13,500$ ), in the 2000-2016 period. We excluded a total of 233 data points (roughly 0.1% of the dataset), 58 of which were due to extreme value observations attributable to data entry error, and the remaining 175 due to missing values in one of the outcome variables. The total number of observations for all 50 US states and school districts (on average 13,500) is  $N = 229,553$ . The outcome variable we are seeking to characterize is defined as:

$$x = \frac{\text{Total Local Education Expenditures}}{\text{Enrollment}} - \frac{\text{Total Local Taxes and Charges}}{\text{Population}} \quad (3.1)$$

The outcome variable  $x$  is the *household per pupil rate of return on tax and service charges*, which we term *educational returns*. The variable *Total Local Education Expenditures* is aggregated from a large set of expenditure categories in primary and secondary education (K-12) that include *instruction, textbooks, pupil support services, staff, transportation, administration, maintenance, food services, utilities, supplies, and technology*. We denote the variable *Total Local Education Expenditures*, scaled by enrollment, as  $\kappa$ .

The variable *Total Local Taxes and Charges* aggregates the following revenue categories: *Private contributions, fines and forfeits, property sales, rents and royalties, sales and services, individual and corporate income taxes, general fees, public utility taxes, general sales taxes, and property taxes*. We denote the variable *Total Local Education Expenditures*, scaled by school district population, as  $\tau$ .

Due to constraints from missing data or comparability across regions and years, this paper works with aggregate local revenue categories, without excluding *general fees* or *service charges*.

### 3.2.2 Empirical Distributions of $x$

Below we present the marginal empirical distribution for  $x$ , as defined in formula ??, for the period 2000 – 2016. In figure 3.2 we plot a stacked histogram with

---

Data Portal (Version 0.10.0), Urban Institute, accessed February, 2021, <https://educationdata.urban.org/documentation/>, made available under the ODC Attribution License.



the empirical density for each year in the pooled sample. The stacked histogram reveals the persistent organization of educational returns into highly peaked asymmetric distributions with positive skew. The pattern variance in the right tails is particularly revealing of disequilibrium fluctuations in the Tiebout sorting process. Fatter right tails with positive skew, we believe, might constitute a strong signal of inter-jurisdictional sorting in the Tiebout sense.

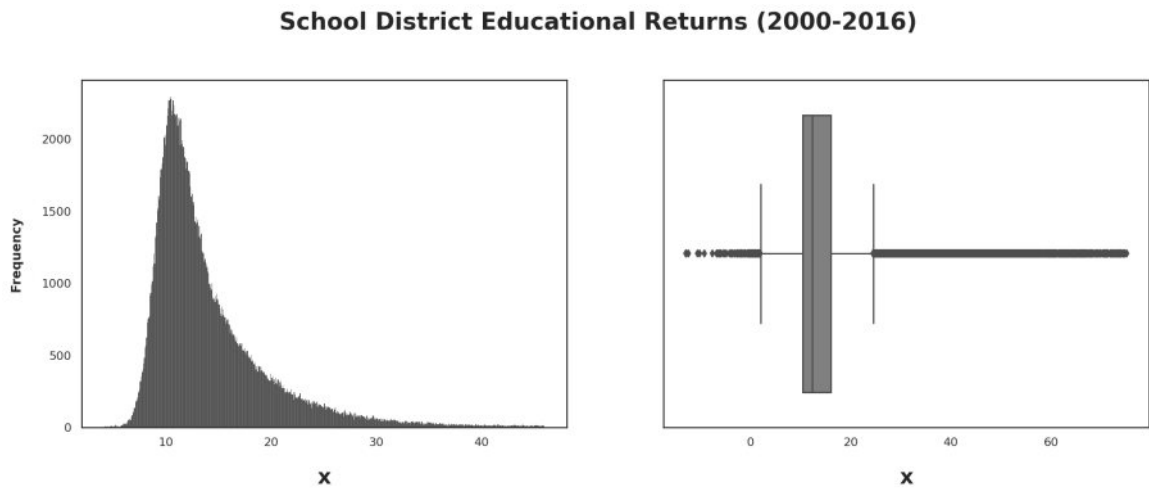


Fig. 3.1 Marginal distribution of  $x$  (in thousands) for the period 2000-2016. Histogram and Box-Plot.

A visual inspection of these distributions points to the asymmetric Subbotin or exponential power distribution (Alfarano et al., 2012b) as a potential candidate for modeling the statistical equilibrium density and for approximating the empirical frequencies of school district educational returns in the period considered (Scharfenaker and Semieniuk, 2015)>.

While these may be good candidates for characterizing highly skewed and peaked distributions, in general, their specifications would not allow us to draw straightforward theoretical conclusions from estimates of their location, scale and shape parameters. In sections 4 and 5, we show how the maximum entropy derivation of the QRSE model leads to a marginal density function  $\hat{f}_x$  whose parameter estimates can be directly linked to the impact of competition and households incentives on jurisdictional sorting and expenditure levels.

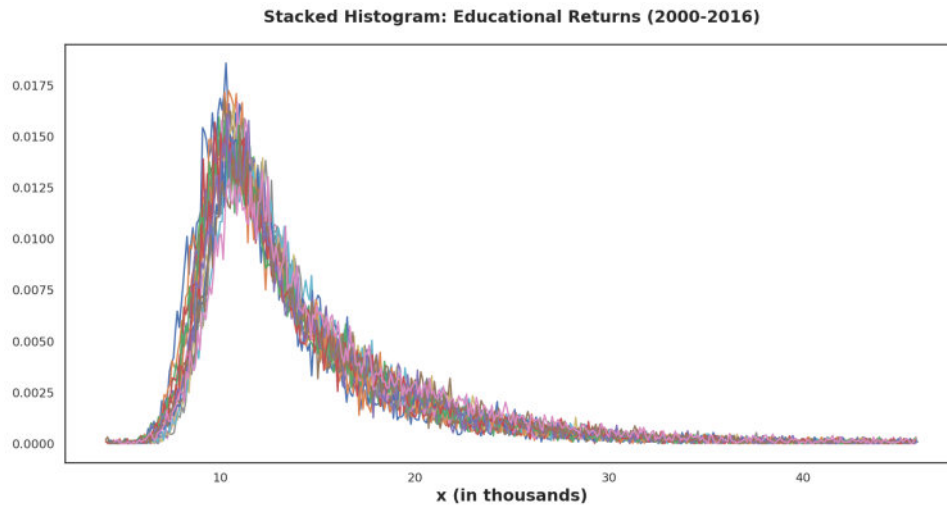


Fig. 3.2 Stacked histogram plots of the educational returns variable  $x$  (in thousands). 2000-2016. The stacked histograms reveal the time invariance of the statistical equilibrium distributions for the different years that make up the pooled sample.

### 3.3 Free Competition and The Tiebout Hypothesis

#### 3.3.1 Smith's Theory of Competition

In the classical Smithian theory of competition, profit-seeking agents make the choice to enter or exit lines of production based on the market's prevailing rate of return. Unlike general equilibrium models, where prices and rates of return are understood as static market clearing quantities in a pure exchange economy, competitive price and rate discovery in the Smithian conception follows a process governed by negative feedbacks. As producers enter profitable lines of production they tend to lower the profit rate by crowding the output and supply of that particular good. Eventually, this forces relocation of capital and resources into other sectors, pushing the rate back to attractive levels (Scharfenaker and Foley, 2017; Shaikh, 2016; Smith, 1937).

This homeostatic process, in which rates are pushed up and down as resources enter and relocate throughout sectors in the economy, has the particular advantage that it lends itself well to a probabilistic interpretation (Farjoun and Machover, 1983). Observed prices and returns can be effectively seen as gravitating around a fundamental central tendency. If we think in terms of probability distributions, we may suitably express a rate's theoretical fundamental value or its 'natural' (regulating) level as the distribution's location parameter. Similarly, the extent to which observed

rates vary and the intensity with which they respond to exit and entry decisions may be suitably captured by variance and scale parameters.

In this paper we seek to link this theory of free competition to Tiebout's original account of the role that inter-jurisdictional competition plays in determining equilibrium levels for local fiscal and expenditure variables, as well as for household choice. Our proposition departs radically from the education finance literature in that we propose a statistical understanding of equilibrium, and do not follow the requirement that parameter estimates be interpreted as operating at fully efficient margins. This leaves open the possibility that households' expectations and preferences might remain unfulfilled in equilibrium. But because the equilibrium is statistical, and not static, this does not constitute a barrier for analysis. The framework considered here allows us to study how such departures from optimality and efficiency may relate to the distinct statistical features of expenditure levels in decentralized education markets, such as positive skew and sharp pre-modal decay.

### **3.3.2 The Tiebout Hypothesis**

In his 1956 paper, Tiebout laid out a set of highly abstract assumptions for his model of local public goods competition, which he also called "a pure theory of local expenditures". The model's assumptions may be summarized as follows:

1. Consumer-voters have full mobility and knowledge of prevailing expenditure patterns in neighboring communities.
2. Mobility is a proxy for consumer-voter choice.
3. There is a large number of communities from which to choose.
4. There is an optimal community size, given demand conditions and fixed resources.
5. Tax-service packages are set according to consumer-voter preferences.

Given that Tiebout's model postulates consumer-voters as choosing tax-service packages by moving in and out of communities, and that the levels of these public goods packages are determined by local governments in response to demand (i.e

migration inflows and outflows), we can see how the Smithian framework applies. In the context of education expenditures, we assume that inhabitants are looking for high rates of return to their decision to locate or relocate to a particular community. We take these rates of return to be proportional to the difference between the *per capita local tax rate* and *per pupil local expenditures* that consumer-voters face in the local public goods market. For the purposes of this paper, we take these market units to be school districts. Competitive school districts will offer attractive per pupil expenditure rates, and low per capita tax rates and service charges. As consumer-voters crowd districts with good schools and low taxes, the rates of educational returns return will adjust accordingly, and under the assumptions of full mobility and rational incentives to fulfill expectations in a local public goods payoff, the iterative process of rate adjustment and migration flows will stabilize expenditures into the observed patterns.

The outcome variable  $x$ , the *per pupil rate of return on local tax spending and charges*, which we termed **educational returns** is defined by the difference:

$$\mathbf{x} = \kappa - \tau$$

where  $\kappa$  is the total local expenditure per pupil, scaled by the school district's enrollment, and  $\tau$  the total local tax and charge burden, scaled by the district's population. In the next section we delve deeper into our statistical treatment of Tiebout competition and derive the QRSE density  $\hat{f}_x$ .

## 3.4 Local Public School Expenditures and Household Choice

### 3.4.1 Statistical Equilibrium Modeling and Maximum Entropy Inference

The highly peaked and positively skewed patterns of the outcome variable  $x$  for the 2000-2016 period suggests the existence of a central tendency in the distribution along with non-symmetric deviations from its mean. Asymmetric Exponential Power Distributions (AEPD) and Skewed Exponential Power Distributions (SEPD) (Al-

farano et al., 2012b; Mundt and Oh, 2019; Scharfenaker and Semieniuk, 2017) are good candidates to model this kind of data. But because we need a constructive probabilistic description that is phenomenologically relevant, as well as theoretically interpretable in its parameters, we implement a Quantal Response Statistical Equilibrium (QRSE) model to fit this data.

The notion of statistical equilibrium has been widely used in physics and information theory (Jaynes, 1957, 1983). A statistical equilibrium for a quantity  $x$  takes the form of a probability density function  $f_x$ ; it represents the most likely distribution for the outcome variable given a set of theoretical and empirical conditions. It can be derived by maximizing the entropy  $H[f_x] = -\sum_x f_x \log[f_x]$  subject to constraints expressing relevant information, theory or observations. The methodology is most commonly used in the context of bayesian statistics with the purpose of deriving informative priors by feeding moment constraints and relevant background information to the maximum entropy program. An important feature of the maximum entropy program is that as long as the set of constraints provided describe a non-empty convex set in the space of distributions, the maximum entropy program will yield an optimal solution that can be used as the statistical equilibrium density of the model (Golan, 2018; Scharfenaker and Foley, 2017). For more details on the derivation of maximum entropy distributions see (Golan, 2018; Jaynes, 2003; Sivia and Skilling, 2006).

We can view single-state solutions to general equilibrium models as special cases of this statistical equilibrium model. They represent degenerate probability densities for the variable  $x$  where only the optimal solution is assigned a positive probability. Such degenerate distributions, with all the probability mass concentrated around a single point, also imply systems that operate with zero entropy.

Formally, the model we present here is a derived maximum entropy distribution for the joint density of household jurisdictional choice and educational returns  $x$ . Rather than giving full statistical content to the complete set of assumptions in the Tiebout hypothesis, we use this derived probability model to examine Tiebout's intuition regarding the role of competition and household choice in shaping the marginal distribution  $f_x$ .

### 3.4.2 A Logit Quantal Response Function for Household Choice

The general Quantal Response Statistical Equilibrium (QRSE) model presented here links a set of household quantal actions  $a \in \mathcal{A}$  to the outcome variable  $x \in \mathbb{R}$ . This could also be a vector  $\vec{x}$  in  $\mathbb{R}^n$ , but in this paper the variable  $x$  is a scalar, which corresponds to the level of educational returns at the school district level.  $\mathcal{A}$  is the binary action set  $\mathcal{A} = \{e, s\}$ —where  $e$  stands for the *entry* of households into a particular school district, and  $s$  for the *exit*. The interaction between the hidden quantal action set  $\mathcal{A}$  and the outcome variable  $x$  is modeled by the joint distribution  $f_{x,a}$ . The maximum entropy distribution  $f_{x,a}$  represents a statistical equilibrium where the inflow/outflow actions of households, represented by the set  $\mathcal{A}$ , are conditionally dependent on the educational returns rate  $x$ , but also shape it via equilibrating forces and the negative feedback process which we defined as Smithian and Tiebout-like competition. We define the payoff for the typical household by the function

$$\pi(a, x) : \mathcal{A} \times \mathcal{X} \rightarrow \mathbb{R} \quad (3.2)$$

The payoff takes as input an action from the action set  $\mathcal{A}$ , and a signal from the state space of educational returns  $x \in \mathcal{X}$ . We use linear symmetric payoffs such that  $\pi(e, x) = -\pi(s, x)$ , as shown in equation 4.5.

$$\begin{aligned} \pi(e, x) &= x - \mu \\ \pi(s, x) &= \mu - x \end{aligned} \quad (3.3)$$

The difference of the entry and exit payoffs is given by equation 3.4 below:

$$\begin{aligned} \Delta\pi(a, x) &= \\ &= \pi(e, x) - \pi(s, x) \\ &= 2(x - \mu) \end{aligned} \quad (3.4)$$

This payoff structure contains a location parameter  $\mu$  to express the fact that households will have a tipping point for moving in or out of a particular school district. Households will tend to move into districts where the level of educational

returns is above this expectation  $\mu$  and vice versa. Note that  $\mu$  is not the average rate, but the expectation that the households forms prior to relocation.

The first constraint that we impose on our statistical model of local education returns is that it be micro-founded by a probabilistic theory of behavior. In other words, we expect the entry and exit decisions of households to be non-deterministic responses to variations in local expenditure patterns for the set of communities that constitute the local public goods market. It is possible to think of this as the assumption that households follow ‘mixed strategies’ in determining whether to move in or out of a particular district. At times they will follow their payoff maximizing action, but sometimes they won’t. We expect the probabilities of observing a particular behavior to be proportional to the payoffs in equation 4.5, and exclude the degenerate case in which households choose only the payoff maximizing action with probability 1.

One way to derive the stochastic function which describes the micro-level behavioral component of the model is to impose a minimum entropy constraint on the utility maximization program of the agent. The household payoff maximization program and the associated Lagrangian take the forms shown below in equations 3.5 and 3.6.

$$\begin{aligned}
 & \max_{f_{a|x} \geq 0} \sum_{\mathcal{A}} f_{a|x} \pi(a, x) \\
 & \text{s.t:} \\
 & \sum_{\mathcal{A}} f_{a|x} = 1 \\
 & - \sum_{\mathcal{A}} f_{a|x} \log[f_{a|x}] \geq H_{\min}
 \end{aligned} \tag{3.5}$$

$$\begin{aligned}
 \mathcal{L} = & - \sum_{\mathcal{A}} f_{a|x} \pi(a, x) - \lambda \left( \sum_{\mathcal{A}} f_{a|x} - 1 \right) \\
 & + T \left( \sum_{\mathcal{A}} f_{a|x} \log[f_{a|x}] - H_{\min} \right)
 \end{aligned} \tag{3.6}$$

This maximization program introduces the behavioral parameter of the model  $T$ . In Scharfenaker and Foley (2017) it is described as a ‘behavior temperature’ parameter in analogy to statistical models of thermodynamic systems, but it can also

be understood as a bounded rationality constraint. Maximizing the payoff subject to a minimum entropy constraint is dual to the problem of maximizing the entropy of the mixed strategy  $f_{a|x}$  subject to a minimum payoff constraint. In that dual case, the Lagrangian contains the term  $\beta (\sum_a f_{a|x} \pi(a, x) - U_{\min})$ , which links the multiplier  $\beta = \frac{1}{T}$  to the minimum payoff constraint.

The solution to this programming problem yields a general logit quantal response or Gibbs density, as in equation 3.7:

$$f_{a|x} = \frac{e^{\frac{\pi(a, x)}{T}}}{\sum_{\mathcal{A}} e^{\frac{\pi(a, x)}{T}}} \quad (3.7)$$

For the case of the binary action set  $\mathcal{A}$ , the program yields the canonical QRSE logit quantal response functions in 3.8 and 3.9.

$$f_{e|x} = \frac{1}{1 + e^{-\frac{\Delta\pi(a, x)}{T}}} = \frac{1}{1 + e^{-\frac{2(x-\mu)}{T}}} \quad (3.8)$$

$$f_{s|x} = 1 - f_{e|x} = \frac{1}{1 + e^{\frac{2(x-\mu)}{T}}} \quad (3.9)$$

This pair of stochastic quantal response functions take the shape of the cumulative distribution function for the logistic distribution. The parameter  $T$  is the scale parameter that expresses the sensitivity of the household choice rule to the difference in the observed outcome from the subjective expectation  $(x - \mu)$ . The introduction of  $\mu$  allows us to model household behavior as 'chasing' a central tendency in the outcome variable  $x$ , and as having a 'tipping point' for the choice to enter or exit a particular school district. In the context of Tiebout competition, these stochastic choice rules should be conceived as representing conditional probabilities for migration inflow or outflow into the ensemble of school districts for which the parameter  $T$  is estimated. They are not 'agent-level' functions that additively aggregate to the ensemble equilibrium distribution, but rather a *meso-level* description that models the dependency of entry/exit flows on observed expenditure patterns and local fiscal variables.



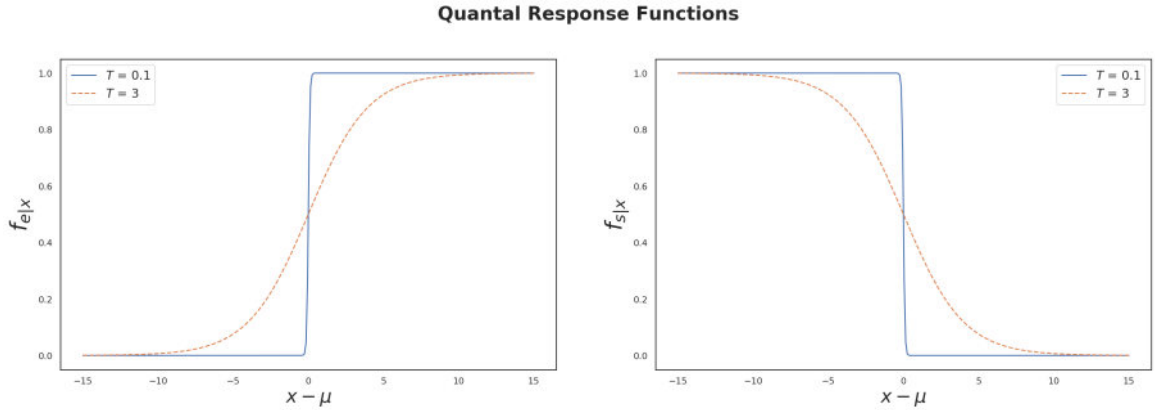


Fig. 3.3 Quantal responses functions for probability of household entry/exit. At near zero  $T$  the quantal response function takes resembles the heaviside step function. Under a zero entropy assumption ( $T = 0$ ), we would expect households to move into districts exclusively in the case where  $x - \mu > 0$ . This graph is illustrative (taking  $T=0.1$  and  $T = 3$  as examples).

### 3.4.3 The Competitive Feedback Constraint

In the canonical QRSE model from Scharfenaker and Foley (2017) which is applied here, agents respond to payoff differentials by entering or exiting a particular economic sector. In the Tiebout setting, household inflows into particular districts may cause congestion in public school services and lead to a feedback effect where the level of educational returns  $x$  will be reduced. Conversely, household outflows from school districts will tend (over time) to push the educational returns rate back up. In the QRSE model, this assumption takes the form of a congestion or ‘competitive feedback’ constraint on the outcome distribution. The competitive feedback constraint is expressed as an inequality that limits the scale of the difference between expected market rates conditional on entry and exit, as shown in 3.10.

$$0 \leq f_e E[(x - \alpha)|e] - f_s E[(x - \alpha)|s] \leq \varepsilon \quad (3.10)$$

This inequality expresses the idea that the expected jurisdictional market rates will be higher conditional on entry than on exit—but that their difference is small enough so that we wouldn’t expect an infinite inflow or crowding into a particular district. This constraint allows us to parsimoniously model the simultaneous and feedback driven relationship that exists between household choice and expenditure levels in the local education market.

The mean outcome level  $\bar{x}$  is then co-determined by a complex (non-reductive) interaction between those two layers of the economic process. In practice we tend to find that  $\alpha \neq \mu$ , which means that the market sustains unexpected outcomes and unfulfilled expectations, an assumption that seems appropriate for the case of public goods markets. The constraint in 3.10 can be unpacked in more detail using the form in equation 3.11, where it is written as an expectation of the market outcome  $(x - \alpha)$ , factored by the difference in mixed strategy probabilities  $\Delta f_{a|x}$ :

$$\begin{aligned}
 & f_e \int f_{x|e} (x - \alpha) dx - f_s \int f_{x|s} (x - \alpha) dx \\
 &= \int f_{e|x} f_x (x - \alpha) dx - \int f_{s|x} f_x (x - \alpha) dx \\
 &= \int \Delta f_{a|x} f_x (x - \alpha) dx \\
 &= \int \tanh\left(\frac{x - \mu}{T}\right) f_x (x - \alpha) dx \\
 &\leq \varepsilon
 \end{aligned} \tag{3.11}$$

The tanh function arises from the definition of the logit quantal response functions, as shown below in 3.12:

$$\begin{aligned}
 \Delta f_{a|x} &= f_{e|x} - f_{s|x} \\
 &= \left( \frac{1}{1 + e^{-\frac{2(x-\mu)}{T}}} \right) - \left( \frac{1}{1 + e^{\frac{2(x-\mu)}{T}}} \right) \\
 &= \frac{e^{2(x-\mu)/T} - 1}{e^{2(x-\mu)/T} + 1} \\
 &= \tanh\left(\frac{2(x-\mu)}{2T}\right) \\
 &= \tanh\left(\frac{x-\mu}{T}\right)
 \end{aligned} \tag{3.12}$$

Thus, the competitive feedback constraint can be written using the general form written in 3.13, noting that  $\Delta f_{a|x} = \tanh\left(\frac{x-\mu}{T}\right)$ :

$$\int \Delta f_{a|x} f_x (x - \alpha) dx \leq \varepsilon \quad (3.13)$$

In the next subsection we explain how the assumptions of entropy constrained behavior and the existence of a competitive feedback constraint determine, via the maximum entropy program, the joint distribution  $f_{a,x}$ .

### 3.4.4 QRSE Maximum Entropy Program and Density

The maximum entropy program for the QRSE model constrains the joint distribution  $f_{a,x}$  so that it is consistent with the following two propositions:

1. The behavioral property of a non-zero entropy rule for household jurisdictional choice (entry/exit decisions).
2. The competitive feedback constraint that we postulate for the Tiebout-like process in analogy with the Smithian theory of competition.

These conditions were formally defined in the previous two subsections. We also constraint the distribution so that it meets the usual normalization condition:  $\int f_x dx = 1$ . Hence, the program maximizes the joint entropy of  $f_{a,x}$  subject to normalization, competitive feedback, and bounded household choice constraints. We can express the joint entropy  $H_{x,a}$  in terms of the marginal entropy  $H_x$  and the conditional or 'binary entropy'  $H_{a|x}$  (Cover and Thomas, 2006), as shown in 3.14 and 3.15.

$$H_{x,a} = H_x + \int_{\mathcal{X}} f_x H_{a|x} dx \quad (3.14)$$

$$H_{a|x} = - \sum_{\mathcal{A}} f_{a|x} \log[f_{a|x}] \quad (3.15)$$

Using the above decomposition, we can write the maximization program for the QRSE model using the compact form shown below in equation 3.16.

$$\begin{aligned}
& \max_{f_x \geq 0} H_x + \int_{\mathcal{X}} f_x H_{a|x} dx \\
& \text{st.} \\
& \int f_x dx = 1 \\
& \int \Delta f_{a|x} f_x (x - \alpha) dx \leq \varepsilon
\end{aligned} \tag{3.16}$$

Note that the second constraint parsimoniously encodes both the behavioral and the market feedback constraints. The associated Lagrangian takes the form in equation 3.17 below:

$$\begin{aligned}
\mathcal{L}[f_x, \lambda, \gamma] &= H_x + \int_{\mathcal{X}} f_x H_{a|x} dx \\
& - \lambda \left( \int f_x dx - 1 \right) \\
& - \gamma \left( \int \Delta f_{a|x} f_x (x - \alpha) dx - \varepsilon \right)
\end{aligned} \tag{3.17}$$

The multiplier associated to the competitive feedback constraint in the program yields the  $\gamma$  parameter for the candidate statistical equilibrium density, which measures the effect of the competitive feedback process on the marginal distribution  $f_x$ .

The solution to this maximum entropy program produces a predictive density  $\hat{f}_x$  that is consistent with our description of Tiebout-like competition in local public goods markets. The distribution  $\hat{f}_x$  predicts the marginal frequencies of the outcome variable  $x$  and completes the theory by determining the conditional densities  $f_{x|e}$  and  $f_{x|s}$ , the joint density  $f_{x,a}$ , as well as the expectations  $E[x|s]$  and  $E[x|e]$ .

The solution  $\hat{f}_x$  takes the form of a Gibbs/Boltzmann distribution, shown below in equation 3.18:

$$\hat{f}_x = \frac{e^{H_{a|x}} e^{-\gamma(\Delta f_{a|x})(x-\alpha)}}{Z} \quad (3.18)$$

where  $Z$  is the partition function.

$$Z = \int_{\mathcal{X}} e^{H_{a|x}} e^{-\gamma(\Delta f_{a|x})(x-\alpha)} dx \quad (3.19)$$

By expressing the  $\gamma$  parameter as  $\gamma = \frac{1}{S}$ , we can rewrite the predictive marginal density  $\hat{f}_x$  as below in 4.8:

$$\hat{f}_x = \frac{e^{H_{a|x}} e^{-\tanh\left(\frac{x-\mu}{T}\right)\left(\frac{x-\alpha}{S}\right)}}{Z} \quad (3.20)$$

With this parametrization it is then possible to perform inference using two scale parameters  $T$  and  $S$ , and two location parameters  $\mu$  and  $\alpha$ , which all have the same dimension as the educational returns variable  $x$ . The scale parameter  $S$  accounts for the concentration of educational returns around the mode that arises from the market level process of jurisdictional competition, while the scale parameter  $T$  accounts for the concentration of values that arises from the purposive behavior of households.

In the next section we provide details on the bayesian estimation of the model for the pooled dataset using all US school districts in the period 2000-2016, and focus our discussion on theoretically interpretable results for the four unknown parameters  $T$ ,  $\mu$ ,  $S$  and  $\alpha$ .

## 3.5 Bayesian Estimation of QRSE Model

### 3.5.1 MAPs and Distance Measures

We use Bayesian inference to recover the values for the unknown parameter vector  $\Gamma = [T, S, \mu, \alpha]$ , for the full sample containing all US school districts in the 2000-2016 period. The approach we followed in our estimation procedure was to first find close to optimal values for the model by jointly minimizing the Kullback-Leibler

divergence ( $D_{KL}$ ) between the observed marginal frequency  $\bar{f}_x$  and the inferred theoretical frequency  $\hat{f}_x$ . This is equivalent to finding maximum a posteriori (MAP) point estimates for  $\Gamma$ , given that maximizing the likelihood turns out to be equivalent to minimizing the KL-Divergence (See Golan (2018)). To find these MAPs we used the available optimization packages found in the Python Scipy library (COBYLA and SLSQP). Minimizing the functional in equation 3.21 yields the MAPs that we then use as starting points for the MCMC sampler.

$$D_{KL}(\hat{f}_x \| \bar{f}_x) = \sum \hat{f}_{\Gamma;x} \log \left[ \frac{\hat{f}_{\Gamma;x}}{\bar{f}_x} \right] \quad (3.21)$$

Additionally, we use the Soofi information distinguishability statistic (Soofi ID; See Soofi and Retzer (2002) for details and theory) to evaluate fit performance. The Soofi ID is shown below in equation 4.11. Smaller values of the KL-Divergence and of the Soofi ID imply better model fits and the Soofi ID in particular gives a measure of how much informational content is explained (recovered) by the candidate distribution.

$$ID(\hat{f}_x : \bar{f}_x) = 1 - \exp \left[ -D_{KL}(\hat{f}_x \| \bar{f}_x) \right] \quad (3.22)$$

### 3.5.2 Model Specification and Markov-Chain Monte Carlo Sampling

We use the QRSE density itself as the likelihood for estimation, considering that the sampler holds the data  $D$  fixed as it explores different probabilities for the parameters in  $\Gamma$  via  $P(D|\Gamma)$ . Alternatively, to justify this, one might simply note that the likelihood is proportional to the sampling distribution ;  $L(\Gamma | x) \propto f_{x|\Gamma}$ . The QRSE log-likelihood used for the sampler is shown below in 4.9:

$$\log[\hat{f}_x] = H_{a|x} - \tanh \left( \frac{x - \mu}{T} \right) \left( \frac{x - \alpha}{S} \right) - \log(Z) \quad (3.23)$$

We directly compute the partition function  $Z$  by the sum in 4.10.

$$Z = \sum_{\mathcal{X}} e^{H_{\text{a}}|x} e^{-\tanh\left(\frac{x-\mu}{T}\right) \left(\frac{x-\alpha}{S}\right)} \quad (3.24)$$

We evaluate the log-likelihood in 4.9 by computing sequences of random samples from the joint posterior distribution of  $\Gamma$ . In this paper we use a standard Metropolis-Hastings algorithm (MCMC-MH; see Hogg and Foreman-Mackey (2018)). Our code uses PyMC3 (Salvatier et al., 2015), an open source probabilistic programming framework written in Python <sup>2</sup>.

For each parameter, we run 3 chains with 30,000 iterations and 4,000 tuning samples. All of the chains converged with  $\hat{R} = 1$ . For more details on the convergence statistic  $\hat{R}$  used see Vehtari et al. (2019). We show a plot of the chain sample traces below in figure 3.4. In figure 3.5 we show pair plots of the posterior samples for the four parameters, which do not appear to be correlated. We used truncated normal priors centered near the MAP estimates for  $T$  and  $S$ , with lower and upper bounds at 0.1 and 8 respectively. For  $\mu$  and  $\alpha$  we used normal priors centered near the MAPs and specified large variances in order to explore reasonably wide ranges of the parameter space. Given knowledge about the plausible ranges for the scale and location parameters, along with the MAP estimates, this choice of weakly informative priors seemed appropriate.

---

<sup>2</sup>Code used, data and MCMC sample traces will be made available in a public GitHub repository for review. For details on the PyMC3 library see: <https://docs.pymc.io/api/inference.html>

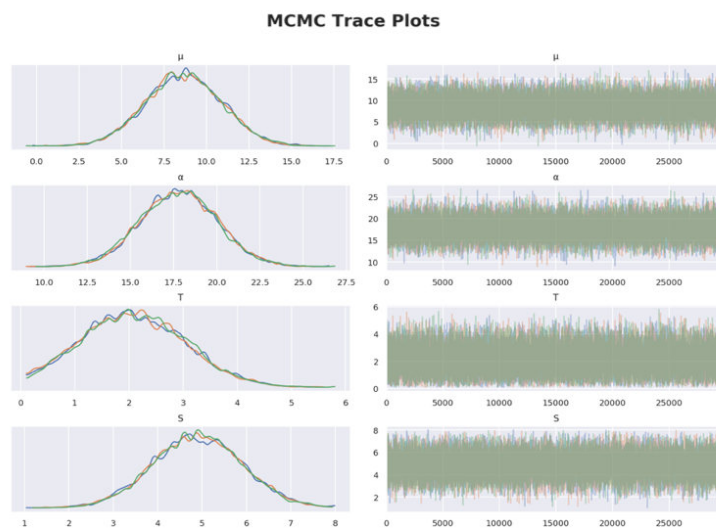


Fig. 3.4 Trace plots for MCMC samples obtained using the Metropolis-Hastings algorithm. 3 chains, 30,000 iterations per chain and 4,000 tuning samples. All US School Districts. 2000-2016.



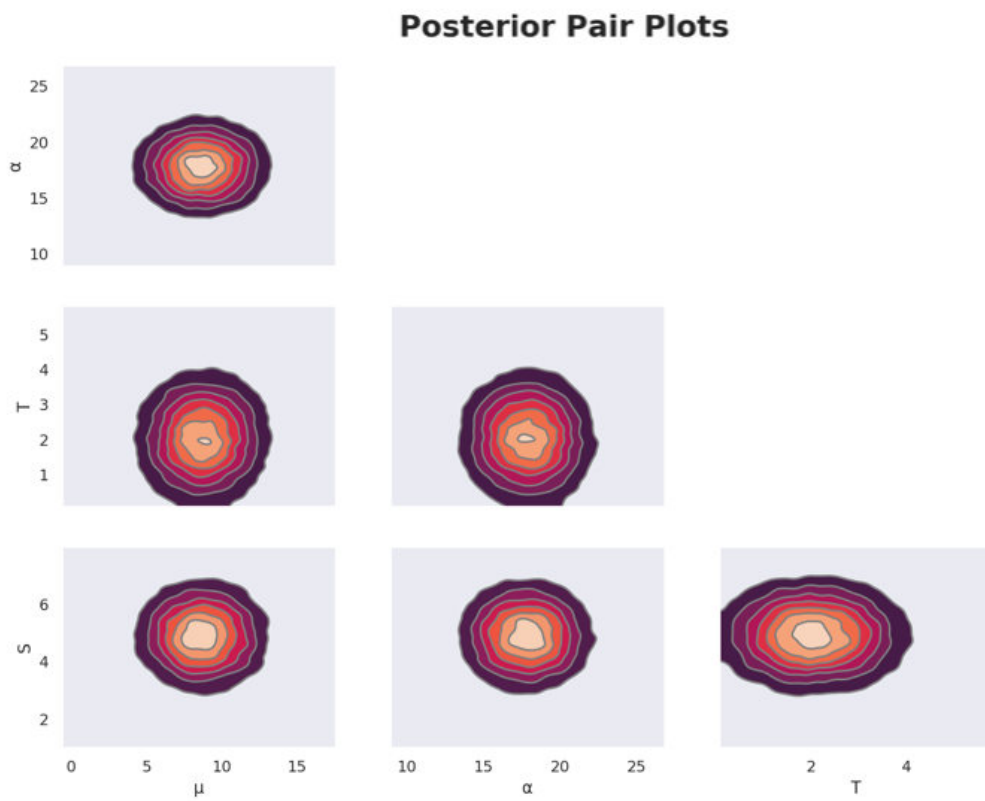


Fig. 3.5 Posterior pair plots for parameters  $T$ ,  $S$ ,  $\mu$  and  $\alpha$ .

### 3.5.3 Results

Table ?? gives the summary statistics for the estimated parameters  $T$ ,  $S$ ,  $\mu$ , and  $\alpha$ . In figure 3.6 we plot the four posterior distributions for the QRSE parameters, which are unimodal, symmetric and have relatively wide standard deviations.

**Posterior Estimates Summary**

Parameter	Mean (Sd)	Mode	94 % HDI	$\hat{R}$
$\mu$	8.66 (2.24)	7.78	[4.54, 12.9]	1.0
$\alpha$	17.8 (2.24)	19.71	[13.61, 22.05]	1.0
$T$	2.1 (0.94)	2.17	[0.24, 3.7]	1.0
$S$	4.9 (1.01)	4.69	[3.01, 6.79]	1.0

Table 3.1 Summary statistics of estimated parameters  $T$ ,  $S$ ,  $\mu$ , and  $\alpha$ . The means, standard deviations, 94% credible intervals and the convergence statistics  $\hat{R}$  from the MCMC samples are reported. All US School Districts, 2000-2016.

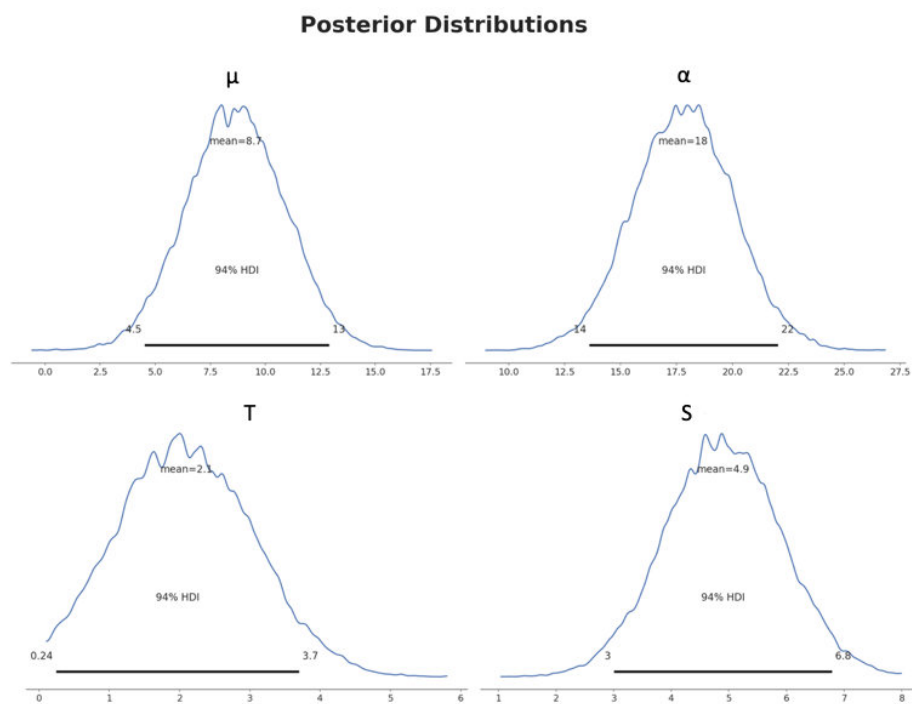


Fig. 3.6 Posterior distributions for  $T$ ,  $S$ ,  $\mu$ , and  $\alpha$ . All US School Districts. 2000-2016.

### 3.5.4 Fits

In table 3.2 we present summary statistics for the fiscal variables used in the model, and for the educational returns variable  $x$ . In figure 3.7 we plot the line histogram of  $x$  for the entire 2000-2016 period, alongside yearly time series for both mean educational returns  $x$  and mean total expenditures  $\tau$ .

#### Summary Statistics

Variable	Mean	S.D.	Min.	Max.
Educational Returns ( $x$ )	14.27	6.30	-12.83	75.08
Total District Expenditures ( $\kappa$ )	15.31	6.69	0.01	102.56
Taxes and Charges ( $\tau$ )	1.04	1.04	0.00	59.55

Table 3.2 Model variables (in thousands). All US School Districts, 2000-2016.

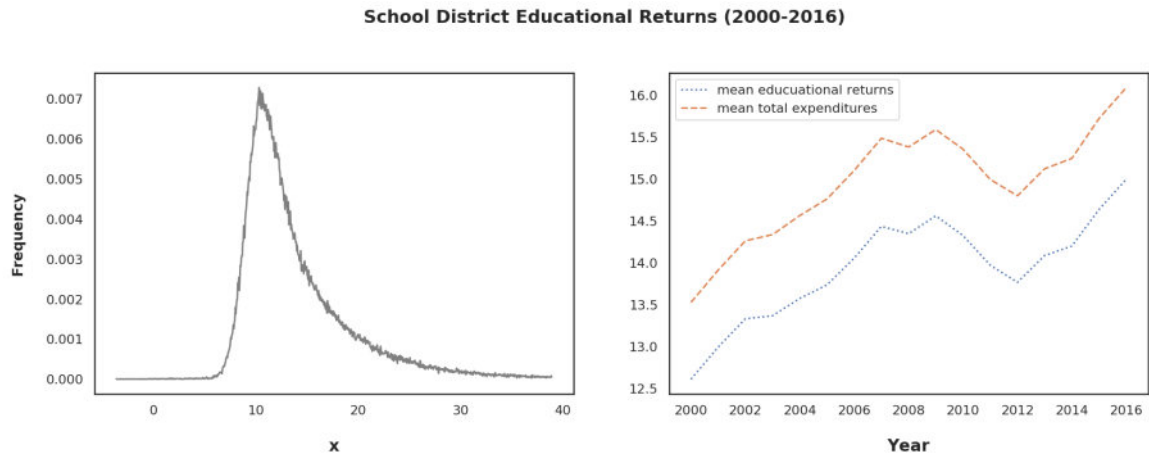


Fig. 3.7 Marginal distribution of  $x$ , and yearly line plot for  $x$  and  $\kappa$  for the period 2000-2016. The histogram excludes the upper 0.01 quantile (for visualization). Plotted in thousands. All US School Districts. 2000-2016.

In 3.8 we fit the estimated QRSE model to the histograms of the observed distribution of  $x$  for this ‘full ensemble case’, which covers all US school districts in the 2000-2016 period. In figure 3.9 we plot the predicted joint action and outcome densities  $f_{a,x}$ , alongside the estimated quantal response functions, which predict the conditional probability of entry and exit of households into districts given a certain level of educational returns  $x$ . We discuss these results in the next subsection.

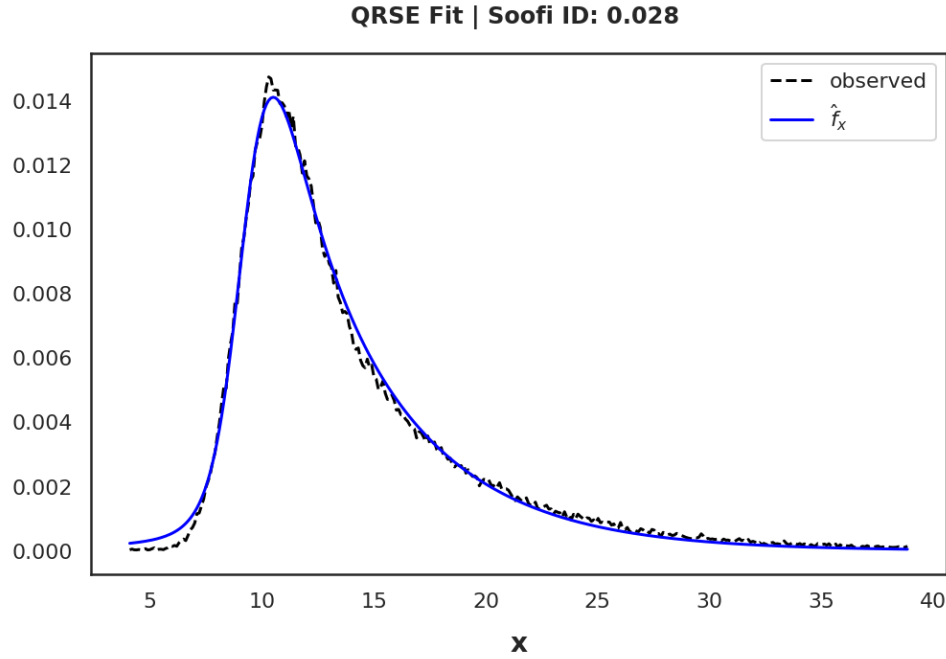


Fig. 3.8 Line histogram of observed distribution for  $x$  (educational returns). Overlaid is the fitted predictive marginal density  $\hat{f}_x$ . We excluded the upper 0.01 quantile (for visualization). The Soofi ID/performance fit measure is shown. All US School Districts. 2000-2016.

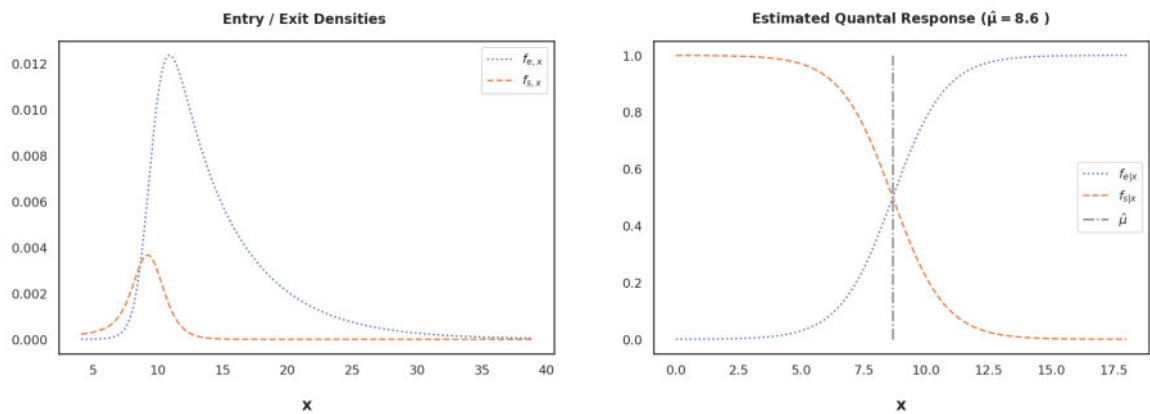


Fig. 3.9 Left: Predictive entry and exit densities  $f_{a|x}$ . Right: Household Quantal Response Functions  $f_{a|x}$ . The estimated 'tipping point'  $\hat{\mu}$  is plotted with a dashed-dotted line.

## 3.6 Discussion

Our central aim is to study the role that Tiebout-like competition may play in explaining observed educational returns across school districts in the US. To do so we implemented a quantal response statistical equilibrium (QRSE) model, which allowed us to characterize competition between local jurisdictions as a complex negative feedback process operating at both the household-level and market-level scales. The QRSE model used in this paper parsimoniously characterizes the complex interaction between household jurisdictional choice and the emergent statistical properties of decentralized education markets in terms of the parameter vector  $\Gamma = [T, S, \mu, \alpha]$ . The observed distribution of educational returns is then explained via the predictive distribution  $\hat{f}_{x;\Gamma}$ . In order to better understand the distinct role that both the scale (S & T) and location ( $\mu$  and  $\alpha$ ) parameters play in explaining observed patterns, it is useful to plot variations to the individual parameters holding all others constant. We do so below in figure 4.1.

The  $\mu$  and  $\alpha$  parameters are particularly relevant in understanding the positive skewness of the statistical equilibrium distribution that we find for the full ensemble case. The parameter  $\alpha$  estimates a market-level statistical tendency that acts as the barycenter around which the household-formed expectation  $\mu$  fluctuates. To see how this is built into the theory, note that in the competitive feedback constraint in 3.10, we write the expectation as  $E[x - \alpha]$  and not  $E[x]$ . In the case where households' expectations of the local educational returns rate matches the market level tendency, then  $\alpha = \mu$ . In that case the distribution is symmetrical and the estimated values for  $\alpha$  and  $\mu$  also match the sample mean  $\bar{x}$ . Whenever  $\mu \neq \alpha$ , then the QRSE distribution is asymmetrical, and positive values for  $\alpha - \mu$  in particular will lend the distribution a more or less sizeable positive skew. As shown in figure 4.1, both the behavioral and market scale parameters T and S predict a lesser/larger concentration of values around the mode, with lower values lending more peakedness to the distribution.

The QRSE model explains concentration around modal values as the consequence of intense competition in decentralized public education markets. Both relatively purposive households and market feedbacks work to stabilize educational returns into their statistical equilibrium distribution. This understanding of competition is consistent with the profit rate equalization hypothesis that one finds in classical political economy, and which has been given modern statistical treatments in (Al-

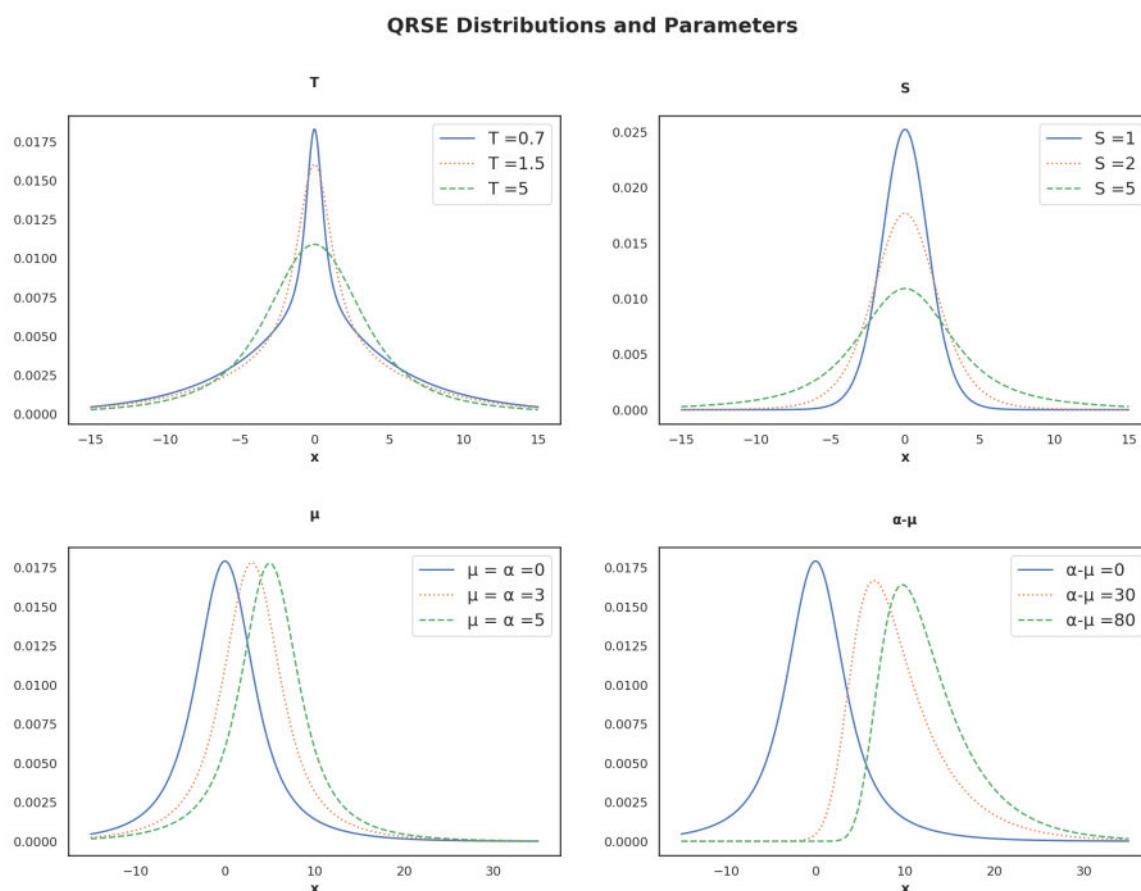


Fig. 3.10 Variations to individual parameter, holding all others constant. The baseline setting is  $T = 5$ ,  $S = 5$ , and  $\mu = \alpha = 0$ .

farano et al., 2012b; Farjoun and Machover, 1983; Scharfenaker and Foley, 2017; Scharfenaker and Semieniuk, 2017).

Our QRSE treatment shows concrete evidence that there are both sorting forces and competitive forces at play in determining the equilibrium educational returns rate. Tiebout sorting in particular educational markets might in fact be signaled by distributions with heavy right tails and positive skew. We believe this to be the case in the sense that ‘better sorted’ or more ‘balanced’ subsamples in  $x$  will undoubtedly contain a broader set of tax-service packages that households sort into via the local public goods and housing markets. In our QRSE model and estimates, the Tiebout ‘sorting forces’ are captured by the size of the difference  $\alpha - \mu$ , while the competitive forces are captured by the size and interaction of the  $T$  and  $S$  parameters. This leads to future work needing to unpack how the  $\alpha$  parameter is related to median household

---

income and property values in school districts, given that high income and property values push public school expenditures far beyond competitive or modal rates.



### 3.7 Conclusion

Educational returns in school districts across the US for the 2000-2016 period exhibit distinctively peaked, positively skewed distributions with right tails of variable width. The shaping of their statistical equilibrium distribution is the outcome of an evolving process of inter-jurisdictional competition, household residential sorting on the basis of a broad set of characteristics (such as income), and shifting policy regimes at the local, state and federal levels.

Using a statistical equilibrium framework, in this paper we sought to examine the role played by inter-jurisdictional competition and household choice in shaping the observed distribution of educational returns for a full ensemble case that covers all US school districts in the 2000-2016 period. This is a considerably larger sample than the ones found in other empirical treatments in the literature, which usually focus on single states or regions. An important aspect of our empirical findings is that it corroborates the need to divorce normative notions about market efficiency from claims about the presence of Tiebout sorting and competition. We proposed a parsimonious model that meaningfully captures the difference between competitive and sorting forces via two sets of scale and location parameters.

Our empirical analysis also corroborates previous findings in the QRSE literature regarding the use of feedback constraints as meaningful characterizations of competition in decentralized market settings. The histogram and model fit displayed in this paper are clearly suggestive of the part played by decentralized competition in sharpening modal peaks, and by elevated (far from competitive) market rates in creating positive skew.

# **Chapter 4**

## **Tiebout statistical equilibrium : sorting and inter-jurisdictional competition across urban locales and income groups**

### **4.1 Introduction**

In 1956 Charles Tiebout proposed to study the decentralized provision of local public goods by postulating the existence of an efficient competitive quasi-market allocation operating at the jurisdictional level. In the competitive equilibrium allocation hypothesized by Tiebout (1956), local governments are construed as suppliers competing with each other for the provision of local public goods, and voters as 'consumer-voters' who reveal their demand by moving in and out of jurisdictions.

Central to the Tiebout hypothesis is the idea that residential mobility can be used as a proxy for public goods demand. In the abstract formulation described by Tiebout (1956), perfect household residential sorting on the basis of preferences for local tax-service packages is hypothesized as an equilibrium condition.

The empirical education finance literature has struggled to make sense of the strong evidence pointing to the incompatibilities between household sorting and perfect inter-jurisdictional competition, along the lines defined by Tiebout. Generally,

most papers take one of two approaches in addressing this issue. A first approach focuses on measuring heterogeneity across and within jurisdictions, and on pairing that analysis of variance with an analysis of fiscal capitalization in the housing market (Dowding et al., 1994), or with any other measure that can signal the strong presence of Tiebout incentives. This approach is constantly confronted with the difficulty of distinguishing pure statistical from the type of sorting that is hypothesized to be driven by Tiebout-like competition.

A second approach, in econometrics, starts by asking whether observed patterns of local public goods expenditures satisfy the necessary conditions implied by a general equilibrium Tiebout model (Epple et al., 1999). The strong condition of satisfying necessary and deterministic equilibrium conditions has set researchers along the path of building computable equilibrium models of increasing mathematical complexity (Nechyba, 2020). The central aim of this type of approach is to predict jurisdictional fragmentation and stratification, without sacrificing neoclassical choice and allocation efficiency. The problem with generic static models is that they specify conditions that are too strict to obtain a equilibrium distribution of the outcome variable, and which are easily exchangeable for others. This exchangeability of mathematical equilibrium conditions makes the models highly idiosyncratic, and specific beyond what is needed to provide a meaningful statistical description of Tiebout sorting and competition. For recent and thorough review of recent developments in the Tiebout literature see (Epple and Nechyba, 2004b; Jha, 2020; Nechyba, 2020). For econometric approaches to studying and testing the Tiebout hypothesis see Dowding et al. (1994), and (Epple et al., 2010, 2001, 1978).

In the statistical equilibrium approach proposed here and in Melo (2021), which is an extension to the local public goods setting of the model developed in Scharfenaker and Foley (2017) and in Scharfenaker and Foley (2021a), a parsimonious model of Tiebout's quasi-market process is used to distinguish between the competitive and the sorting forces at play in the US public education market.

Educational expenditures in school districts across the US for the 2000-2016 period exhibit distinctively peaked, positively skewed distributions with right tails of variable width. The shaping of their statistical equilibrium distribution is the outcome of a complex process of inter-jurisdictional competition, household residential sorting on the basis of a broad set of characteristics (such as income), and shifting policy regimes at the local, state and federal levels. The heavy positive skew and the sharp

modal peaks observed in the empirical frequencies of school district expenditures across urban locales and income groups are suggestive of the presence of both Tiebout competition and income-based stratification.

In the complexity and political economy view taken in this paper, and in Melo (2021), the assumptions and requirements of perfect competition are dropped in favor a description of decentralized Tiebout competition as a type of probabilistic negative feedback mechanism. Demand for local public goods is studied in terms of the quantal entry/exit decisions of households into any given district, conditional on the level of public school expenditures. The in- and out-migration that results due to changes in expenditure levels or housing prices is reflected by the probabilities of entry and exit of households into school districts. The relative magnitude of these probabilities is determined by the size of the divergence between households' fundamental expectation of the distribution's central tendency and a market-determined rate which may be driven by extra-competitive forces.

The paper is structured as follows. Section 2 summarizes the Tiebout Quantal Response Statistical Equilibrium model (T-QRSE) studied in Melo (2021). Expanding on that previous contribution, section 3 discusses the main implications of entropy constrained behavior and competitive feedback in the Tiebout setting. Section 4 describes the datasets used, the subsampling strategy for urban locales and income groups, and introduces the histograms of the school district expenditures data to be explained. Sections 5 and 6 provide details on the Bayesian estimation procedure and lay out the results of the inference procedure. The results are analyzed and discussed in section 7.

## 4.2 Tiebout statistical equilibrium

### 4.2.1 The educational returns random variable

Following the development proposed in Melo (2021) this paper defines an *educational returns* random variable  $X$ . In the dataset, the observed values  $X = x$  are defined as the difference between *total district education expenditures* and *total taxes/charges*, each respectively weighed by the district's student enrollment and population counts.  $X$  can be understood as the household per pupil rate of return on the district's tax and service charges. Section 4 provides details on the census

datasets and fiscal variables used. The formula for the observed values  $x$  in the dataset used takes the following form:

$$x = \frac{\text{Total Local Education Expenditures}}{\text{Enrollment}} - \frac{\text{Total Local Taxes and Charges}}{\text{Population}} \quad (4.1)$$

The educational returns random variable is taken to be distributed according to a quantal response statistical equilibrium density with two scale and two location parameters;  $X \sim \mathcal{QRSE}(T, S, \mu, \alpha)$ . The next two subsections summarize the construction and the theoretical underpinnings of the QRSE density, and its application to the local public goods setting. For notational convenience,  $f_X(x)$  is written as  $f_x$  throughout the paper.

## 4.2.2 Maximum Entropy Inference

Maximum entropy methods provide a convenient interface for deriving distributional assumptions on the basis of partial knowledge and information. A maximum entropy distribution for a random variable  $X$  is obtained by maximizing the entropy of its density function  $f_x$  subject to a set of relevant moment constraints.

For  $x \in \mathcal{X}$ , the general maximum entropy program takes the form of constrained optimization problem:

$$\begin{aligned} \max_{f_x \geq 0} \quad & H_x = - \int_{\mathcal{X}} f_x \log[f_x] dx \\ \text{st.} \quad & \\ & \int_{\mathcal{X}} f_x dx = 1 \\ & \int_{\mathcal{X}} f_x g_c(x) dx = F_c, \quad c = 1, \dots, m \end{aligned} \quad (4.2)$$

where  $H_x$  stands for the Shannon entropy functional. A constraint enters the program in 5.11 as an expectation (over  $X$ ) of a function  $g_c(x) \in \mathcal{C}$ . The set  $\mathcal{C}$  constrains the derived distribution  $\hat{f}_x$  to take on specific higher-level moments which

are deemed empirically or theoretically relevant. If there are  $m$  constraints, then  $|\mathcal{C}| = m$ .

The solution to this program yields a maximum entropy density  $\hat{f}_x$  with the vector  $\vec{\lambda} = [\lambda_1 \cdots \lambda_m]$  of Lagrange multipliers as parameters, whose values express the marginal amount of information that the constraint contributes to reducing entropy in the distribution (Scharfenaker, 2020b).

The general form for the solution, and the partition function  $Z$ , are shown below in 5.12 and 5.13.

$$\hat{f}_x = \frac{1}{Z} e^{\lambda_1 g_1(x) + \cdots + \lambda_m g_m(x)} \quad (4.3)$$

$$Z = \int_{\mathcal{X}} \exp \left[ \lambda_1 g_1(x) + \cdots + \lambda_m g_m(x) \right] dx \quad (4.4)$$

For standard references on solving this program via the method of Lagrange multipliers see (Cover and Thomas, 2006; Golan, 2018; Sivia and Skilling, 2006) and the mathematical appendix. For detailed discussions on the derivation of maximum entropy distributions see (Golan, 2018; Jaynes, 2003; Sivia and Skilling, 2006).

### 4.2.3 T-QRSE

The QRSE model is a general probabilistic model of competition which provides a parsimonious alternative to the deterministic treatment of equilibrium outcomes in complex economic environments driven by negative feedbacks. The first development of the canonical QRSE model appears in Scharfenaker and Foley (2017), and it has been recently re-parametrized in Scharfenaker and Foley (2021a).

The Tieboutian Quantal Response Statistical Equilibrium model (T-QRSE) is a joint probability model  $f_{a,x}$  for the unobserved in- and out-migration decisions of households across jurisdictional boundaries and the marginal educational returns rate  $x$  (Melo, 2021). In the Tiebout setting, we consider a binary unobserved action set  $a \in \mathcal{A}$ , where  $\mathcal{A} = \{e, s\}$  and  $e$  and  $s$  stand for household in- and out-migration respectively. The function  $f_{a|x}$  is defined as households' probability of in- and out-migration conditional on the educational returns rate  $x$ . A household's payoff

$\pi(a, x)$  takes as input an action from the action set  $\mathcal{A}$ , and a signal from the state space of educational returns  $x \in \mathcal{X}$ . We use linear symmetric payoffs such that  $\pi(e, x) = -\pi(s, x)$ . The payoff structure takes the form:

$$\begin{aligned}\pi(e, x) &= x - \mu \\ \pi(s, x) &= \mu - x\end{aligned}\tag{4.5}$$

And  $f_{a|x}$  takes the form of a logit quantal response function:

$$f_{e|x} = 1 - f_{s|x} = \frac{1}{1 + e^{-\frac{\Delta\pi(a,x)}{T}}} = \frac{1}{1 + e^{-\frac{2(x-\mu)}{T}}}\tag{4.6}$$

For details on the derivation of this logit quantal response function see (Foley, 2020) and the mathematical appendix.

Formally, the joint probability model is obtained by maximizing the entropy of  $f_{x,a}$  subject to the moment constraint  $E[\Delta f_{a|x}(x - \alpha)] \leq \delta$ , where  $\Delta f_{a|x} = f_{e|x} - f_{s|x}$ . In the case with two actions and symmetric payoffs,  $\Delta f_{a|x} = \tanh\left(\frac{x-\mu}{T}\right)$  and the moment constraint takes the form :

$$E\left[\tanh\left(\frac{x-\mu}{T}\right)(x - \alpha)\right] \leq \delta\tag{4.7}$$

The constraint argument  $\Delta f_{a|x}(x - \alpha)$  in 4.7 encapsulates the postulate that differences in the entry/exit probabilities of households across jurisdictional boundaries have a non-zero impact on the statistical dispersion of the educational returns rate around the average. This is a non-linear moment about two location parameters  $\mu$  and  $\alpha$  which together determine the mean of the distribution. The scale of the impact of household in- and out-migration on concentrating values of the distribution around the mode is determined by  $T$ .

The maximum entropy distribution  $\hat{f}_x$  that is obtained on the basis of this moment constraint, along with the usual positivity and normalization conditions, takes the form:

$$\hat{f}_x = \frac{e^{H_{a|x}} e^{-\tanh\left(\frac{x-\mu}{T}\right) \left(\frac{x-\alpha}{S}\right)}}{Z} \quad (4.8)$$

where  $H_{a|x} = -\sum_{\mathcal{A}} f_{a|x} \log[f_{a|x}]$  and  $Z$  is the partition function or integration constant. Since  $f_{a|x} f_x = f_{a,x}$ , the derivation of the marginal density  $\hat{f}_x$  completely specifies the joint probability model<sup>1</sup>. Note that the solution has an added scale parameter  $S$ , which together with  $T$ , accounts for concentration of values around the mode. The size of  $\alpha - \mu$  determines the distribution's skew; whenever  $\alpha = \mu$  the distribution has a single location parameter and the distribution becomes symmetric. In figure 4.1 we plot variations to the individual parameters, holding all others constant.

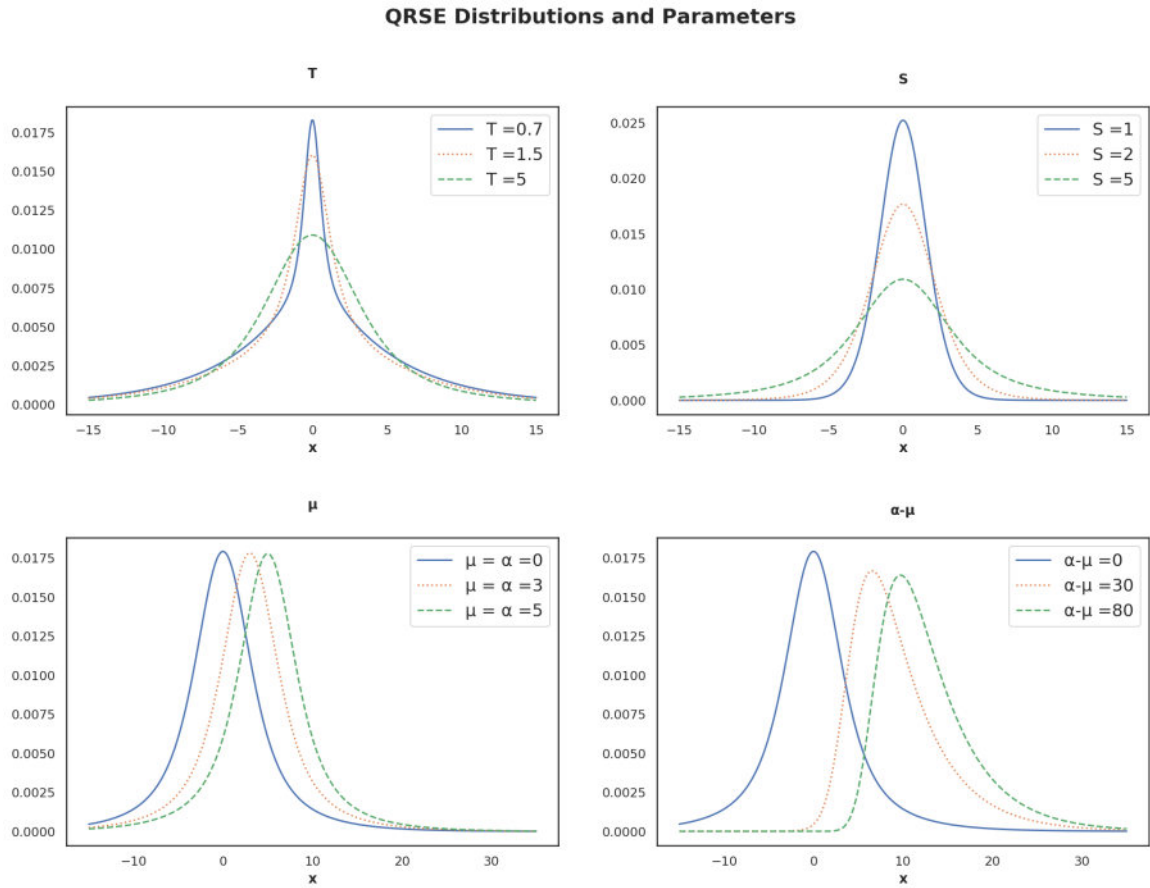


Fig. 4.1 Variations to individual parameter, holding all others constant. The baseline setting is  $T = 5$ ,  $S=5$ , and  $\mu = \alpha = 0$ .

<sup>1</sup>See mathematical appendix



The constraint in 4.7 provides a parsimonious probabilistic description of the negative feedback process that can be seen to regulate average return rates in decentralized competitive economic environments. The next section discusses how this competitive feedback process can be used to explain the positively skewed and peaked distributions of educational returns in terms of Tiebout-like competition and sorting.

### 4.3 Quantal Response & Competitive Feedback

This section illustrates the main components of the joint probability model  $f_{a,x}$ , and lays out the main inferential implications for the local public goods setting.

The plot in the left of figure 4.2 shows the difference in quantal probabilities of in- and out-migration of households across districts  $\Delta f_{a|x}$ . Whenever  $x > \mu$ ,  $f_{e|x} > f_{s|x}$ . Similarly, whenever the educational returns rate drops below the household expectation  $\mu$ ,  $f_{s|x} > f_{e|x}$ . The scale or 'sensitivity' of households to positive and negative variations around  $\mu$  is determined by the behavioral scale parameter  $T$ . In the usual bounded rationality interpretation of this logit quantal response model, larger values of  $T$  represent limited information-processing capacity; decision-makers face a trade-off between accuracy and uncertainty. In the Tieboutian context, this bounded rationality model takes on a particular meaning. Given that households' demand is expressed by moving in and out of districts, the cost of decision-making is the cost of comparing attainable public goods utility across jurisdictions and communities. Since  $\mu$  is the fundamental expectation that households form about an average level of public goods provision, suburban and high-income households inhabiting communities with positively skewed expenditures may be facing a trade-off between sorting on the basis of property values and provisioning public goods at competitive rates. Large differences between the market-level rate  $\alpha$ , which is driven by property values, and  $\mu$ , will hence leave households unfulfilled in terms of Tiebout expectations. This greater uncertainty that is implied by the tradeoff between sorting on the basis of income and competitive public goods provision is captured by larger values of  $T$ . This is a curious but inevitable consequence of the information-theoretic interpretation of Tiebout's consumer-voter, whose decision space is the housing market. Lower and middle income households whose jurisdictional choice set exhibits narrower variance in terms of property values will hence have lower values of  $T$ .

The plot on the right of figure 4.2 shows the case of a balanced competitive feedback process around  $\mu = \alpha$ . Whenever  $x > \mu$  the probability of entry on that outcome level rises. As  $x$  rises further the post-modal decay of  $f_{e,x}$  ensures the effectiveness of the market's negative feedback by lowering the *joint* probability of entry and higher returns. Symmetrically, on the other side, the decay of  $f_{s,x}$  renders improbable the possibility of constant exit at largely negative rates. The conceptual

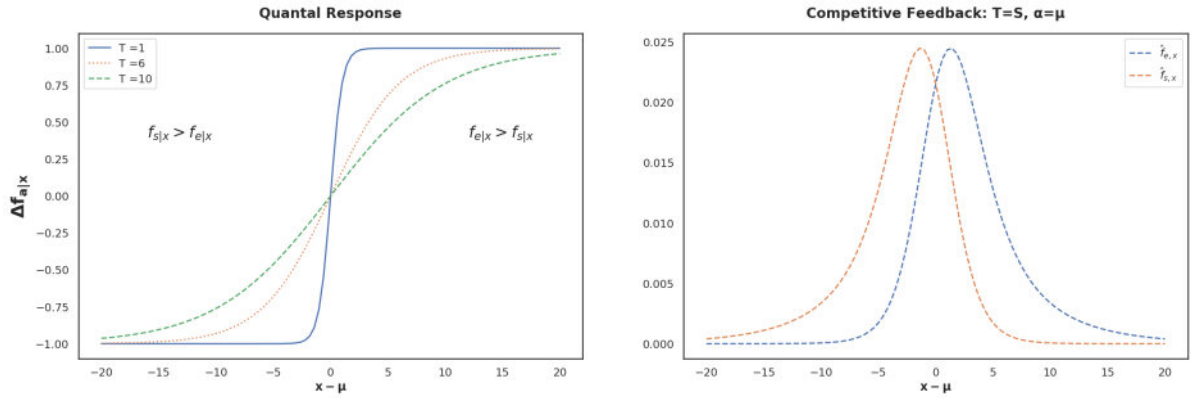


Fig. 4.2 Left: Plot of the difference in quantal response probabilities, with different values of  $T$ . Right: A symmetric competitive feedback process centered at  $\mu = \alpha$ .

schema for this probabilistic competitive feedback process is shown below in figure 4.3.

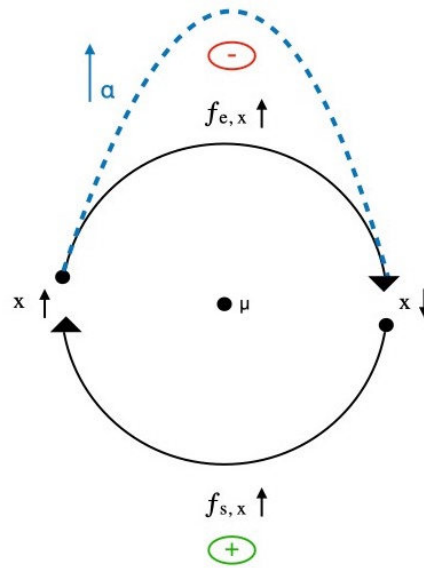


Fig. 4.3 Conceptual illustration of probabilistic competitive feedback. Departures from  $\mu$  break symmetry in the process and create positive skew in the outcome variable.

In the case of fulfilled expectations ( $\alpha = \mu$ ), the joint probabilities  $f_{e,x}$  and  $f_{s,x}$  fall and rise symmetrically around the market average. When  $\alpha > \mu$  the size of inflow probabilities above the average market rate dominates and the effectiveness of the competitive feedback process is reduced. This case is illustrated in figure 4.4.

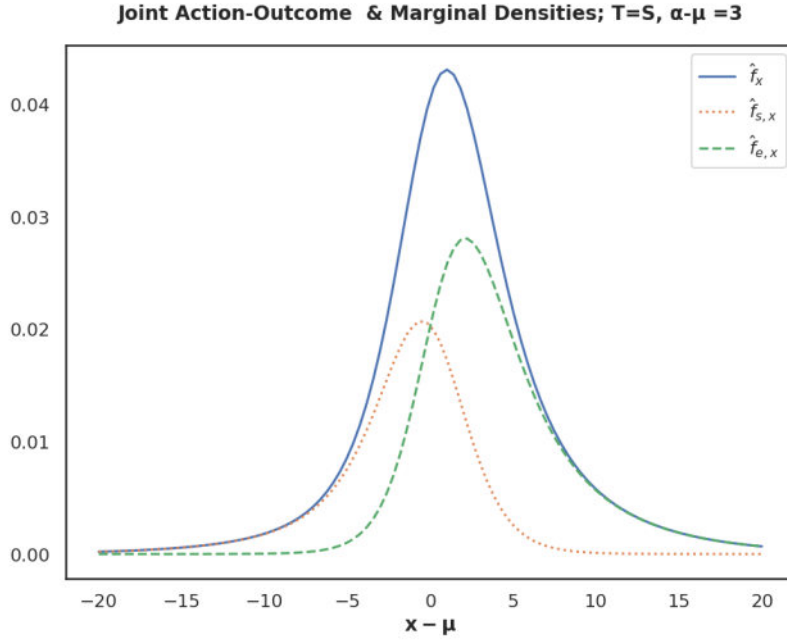


Fig. 4.4 Joint action-outcome and marginal densities with balanced behavioral/market scales and  $\alpha > \mu$ .

The dominance of large inflow probabilities above the expected Tiebout rate  $\mu$  helps explain positive skew in the distribution of educational returns. It also signals the relatively weightier importance of sorting incentives in shaping the outcome. The empirical literature has struggled to reconcile the Tiebout hypothesis with the fact that household jurisdictional choice cannot be entirely attributable to public goods incentives. The T-QRSE model proposed in this paper and in Melo (2021) gives a probabilistic measure of the relative importance of both sorting and Tiebout incentives in determining observed educational expenditures. School district ensembles where the sorting forces dominate will show large positive skew, while those where Tiebout-like competition dominates will tend to be more symmetric and peaked around the average.

In an ideal purely Tieboutian statistical equilibrium the distribution of educational returns takes the shape of a symmetric Laplace density with a location parameter at  $\mu$ , and low values for the behavioral and market scale parameters  $T$  and  $S$ . Under this statistical perspective, perfect sorting is not an equilibrium outcome of the idealized inter-jurisdictional competitive process, as this would imply a uniform distribution of educational returns. But because what is proposed is a probability measure and not a deterministic model, the two forces of competition and sorting can be statistically

reconciled. The relative dominance of either varies across school district ensembles, as the empirical results presented in this paper show. An important caveat to note is that this model explains positive skew in the distribution of educational returns as departures from an ideal Tiebout statistical equilibrium. The braking down of symmetry in the competitive feedback process points to the relative importance that other economic forces may have in shaping the equilibrium outcome.

## 4.4 School District Data

### 4.4.1 Census and NCES Data

This paper uses data from the National Center for Education Statistics' Common Core of Data, the US Census Bureau Small Area Income and Poverty Estimates, and the US Department of Education's EDFacts initiative <sup>2</sup>.

We consider a sample of local expenditures in primary and secondary education, local taxes, enrollment and population estimates for all 50 US states and school districts (on average ~ 13,500), in the 2000-2016 period. We excluded a total of 301 data points (roughly 0.1% of the dataset), 58 of which were due to extreme value observations attributable to data entry error, and the remaining 245 due to missing values in one of the outcome variables. The total number of observations for all 50 US states and school districts (on average ~ 13,500) is  $N = 272,152$ .

The outcome variable was defined as:

$$x = \frac{\text{Total Local Education Expenditures}}{\text{Enrollment}} - \frac{\text{Total Local Taxes and Charges}}{\text{Population}}$$

The variable *Total Local Education Expenditures* is aggregated from a large set of expenditure categories in primary and secondary education (K-12) that include *instruction, textbooks, pupil support services, staff, transportation, administration, maintenance, food services, utilities, supplies, and technology*. The variable *Total Local Taxes and Charges* aggregates the following revenue categories: *Private contributions, fines and forfeits, property sales, rents and royalties, sales and services, individual and corporate income taxes, general fees, public utility taxes, general sales taxes, and property taxes*.

Due to constraints from missing data or comparability across regions and years, this paper works with aggregate local revenue categories, without excluding *general fees* or *service charges*. For the imputation of the urban local classifications, we used

<sup>2</sup>This data has been made available in a harmonized format in a publicly available API by the Urban Institute, which provides a convenient and reliable interface to all the major federal dataset. Education Data Portal (Version 0.10.0), Urban Institute, accessed February, 2021, <https://educationdata.urban.org/documentation/>, made available under the ODC Attribution License.

the EdBuild dataset. This dataset contains median household income in addition to the fiscal and expenditure variables from the Urban Institute dataset <sup>3</sup>. All values have been adjusted for inflation.

#### 4.4.2 Subsamples

Using US public education finance data for the period of 2000-2016, we subsample school districts across four different urban locales: Town, Suburb, City, and Rural. Structural elements of public school markets, such as median income, mobility, demographic composition, and the varying presence of equity-based school finance reform policies, are meaningfully correlated with these urban local classifications. Additionally, for the period 2013-2016, we subsample the data using three distinct income groups based on median household estimates at the school district level: Low, Middle, and High. Details on how we defined these groups are provided in the following subsections.

#### 4.4.3 Urban Locales

Below we provide the definitions for the four urban locales studied in this paper : Town, Rural, City and Suburb. These definitions are set by the ACS (American Community Survey - Census) and the NCES (National Center for Education Statistics). We use these categories to group the data into different statistical equilibrium ensembles for the 2000-2016 period. In figure 4.5 we plot a stacked histogram of the distributions of educational returns for the four urban locales.

- **Town:** Territories inside an urban cluster at varying distances from urbanized areas.
- **Rural:** Census-defined rural territories at varying distances from urbanized areas and urban clusters.
- **City:** Territories inside an urbanized area and inside a principal city.
- **Suburb:** Territories outside a principal city and inside an urbanized area.

---

<sup>3</sup>See <http://data.edbuild.org/>

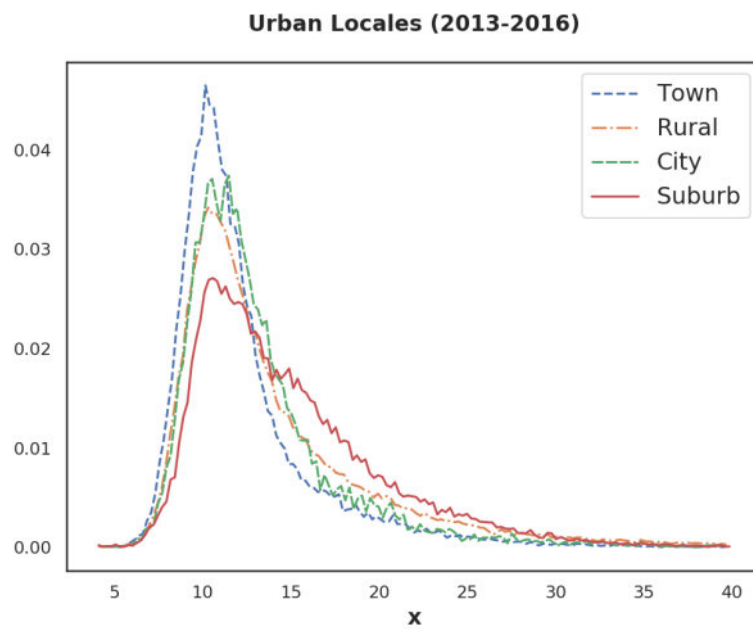


Fig. 4.5 Stacked histogram plots of the educational returns variable  $x$ . Urban Locales: Town, Rural, City, Suburb. 2000-2016



#### 4.4.4 Income Groups for the period 2013-2016

Through the EdBuild dataset we obtained estimates for school district median household income for the period 2013-2016. These estimates can also be obtained directly from the US Department of Education, the National Center for Education Statistics (NCES), and the Education Demographic and Geographic Estimates (EDGE) database.

Visual inspection and analysis of the distribution suggests a simple tripartite division of the MHI range. We use the 0.25 and the 0.95 quantiles (respectively 41320 – *USD* and 95844 – *USD*). The cumulative density and the histogram for the MHI variable, along with the category division lines, are plotted below in 4.6. In figure 4.7 we plot the stacked histogram for the three income groups.

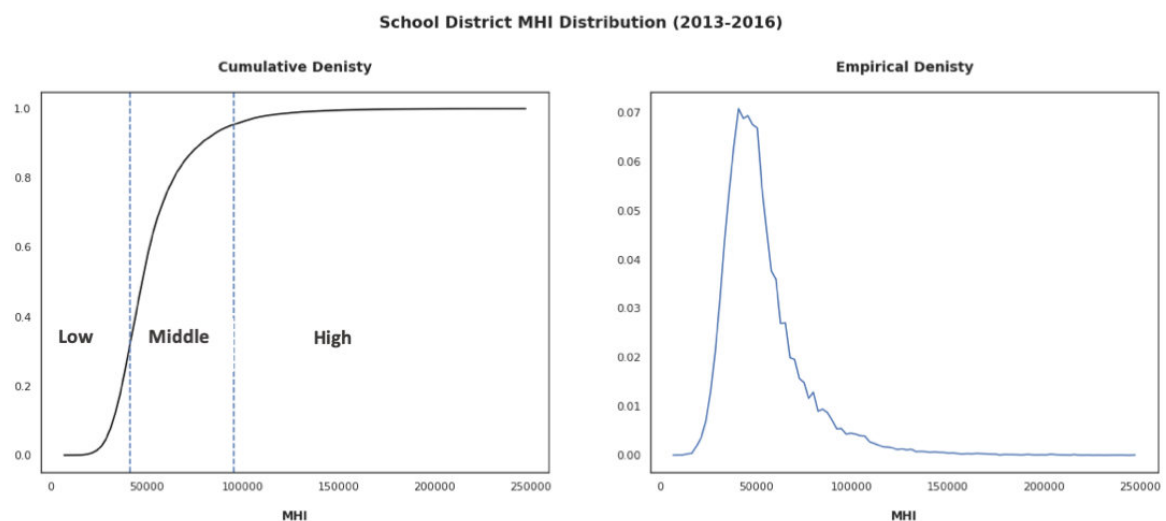


Fig. 4.6 Tripartite division of the median household income distribution for school districts. 2013-2016.

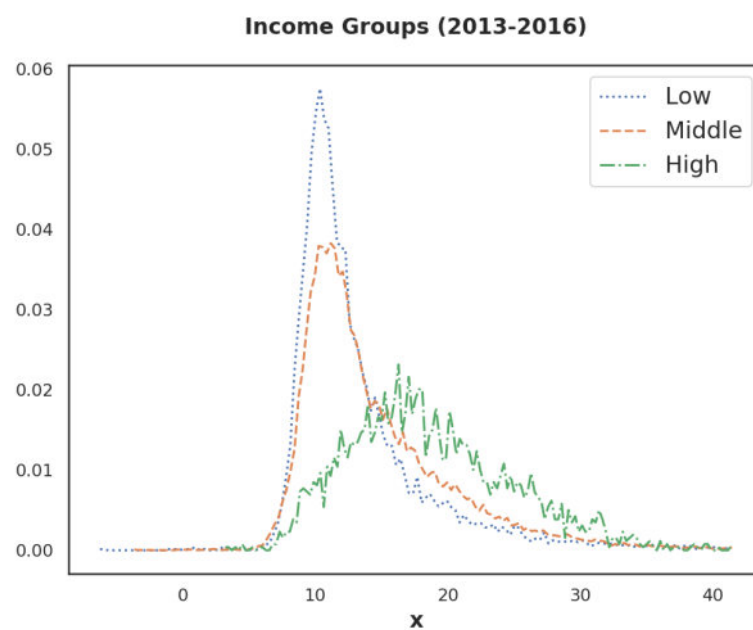


Fig. 4.7 Stacked histogram plots of the educational returns variable  $x$ . Income Groups: Low, Medium, High. 2013-2016. The stacked histograms suggest a marked difference in the statistical equilibrium distribution for the three income groups.

## 4.5 Bayesian Estimation

### 4.5.1 Model Specification and Markov-Chain Monte Carlo Sampling

The QRSE log-likelihood used for the estimation procedure is shown below in 4.9:

$$\log[\hat{f}_x] = H_{a|x} - \tanh\left(\frac{x-\mu}{T}\right) \left(\frac{x-\alpha}{S}\right) - \log(Z) \quad (4.9)$$

This follows the same approach taken in Melo (2021) for estimating the full school district ensemble case. We directly compute the partition function  $Z$  as the sum:

$$Z = \sum_{\mathcal{X}} e^{H_{a|x}} e^{-\tanh\left(\frac{x-\mu}{T}\right) \left(\frac{x-\alpha}{S}\right)} \quad (4.10)$$

We evaluate the log-likelihood in 4.9 by computing sequences of random samples from the joint posterior distribution of  $\Gamma = [T, S, \mu, \alpha]$ . In this paper we use a standard Metropolis-Hastings algorithm (MCMC-MH; see Hogg and Foreman-Mackey (2018)). Our code uses PyMC3 (Salvatier et al., 2015), an open source probabilistic programming framework written in Python <sup>4</sup>.

For each school district ensemble considered, we run 3 chains with 30,000 iterations and 4,000 tuning samples. All of the chains converged with  $\hat{R} = 1$ . For more details on the convergence statistic  $\hat{R}$  used see Vehtari et al. (2019). We used truncated normal priors centered near the MAP estimates for  $T$  and  $S$ , with lower and upper bounds at 0.1 and 8 respectively. For  $\mu$  and  $\alpha$  we used normal priors centered near the MAPs and specified large variances in order to explore reasonably wide ranges of the parameter space. Given knowledge about the plausible ranges for the scale and location parameters, along with the MAP estimates, this choice of weakly informative priors seemed appropriate. The structure used for the bayesian estimation of the model is summarized in figure 4.8.

---

<sup>4</sup>Code used, data and MCMC sample traces will be made available in a public GitHub repository for review. For details on the PyMC3 library see: <https://docs.pymc.io/api/inference.html>

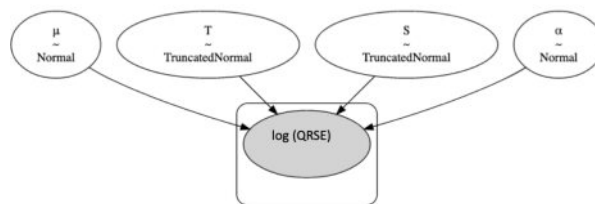


Fig. 4.8 QRSE model structure and priors used for Bayesian Estimation.

## 4.6 School District Ensembles

In the economic statistical equilibrium approach used in this paper, inference is conducted by fitting the marginal predictive density of the outcome variable  $\hat{f}_x$  to the histogram of a well defined ensemble of units of observation, which we defined as district-level educational returns. Uncertainty around the explanatory values for the parameters is recovered via Bayesian estimation of their posterior distributions. Since estimates of the posteriors in the QRSE density reveal causal information about the underlying competitive feedback process, and about departures in the data from perfect Tiebout statistical equilibrium, specific knowledge is obtained by reasoning about the ensemble-level properties which are implied by those values.

The stacked histograms in figure 4.9 reveal a persistent organization of educational returns into highly peaked asymmetric distributions with positive skew. The sharp drops in the pre-modal densities are also revealing of the effect of minimum expenditure levels and balanced district budgets in shaping the asymmetry of these distributions. In the absence of policy floors, balanced budgets, and strong sorting incentives, we would expect the distribution of educational returns to be closer to a symmetric tent-like Laplace distribution, such as the ones that are observed for firm profit rates. This is what the QRSE density predicts for the 'most decentralized' competitive economic environments (See figure 4.1).

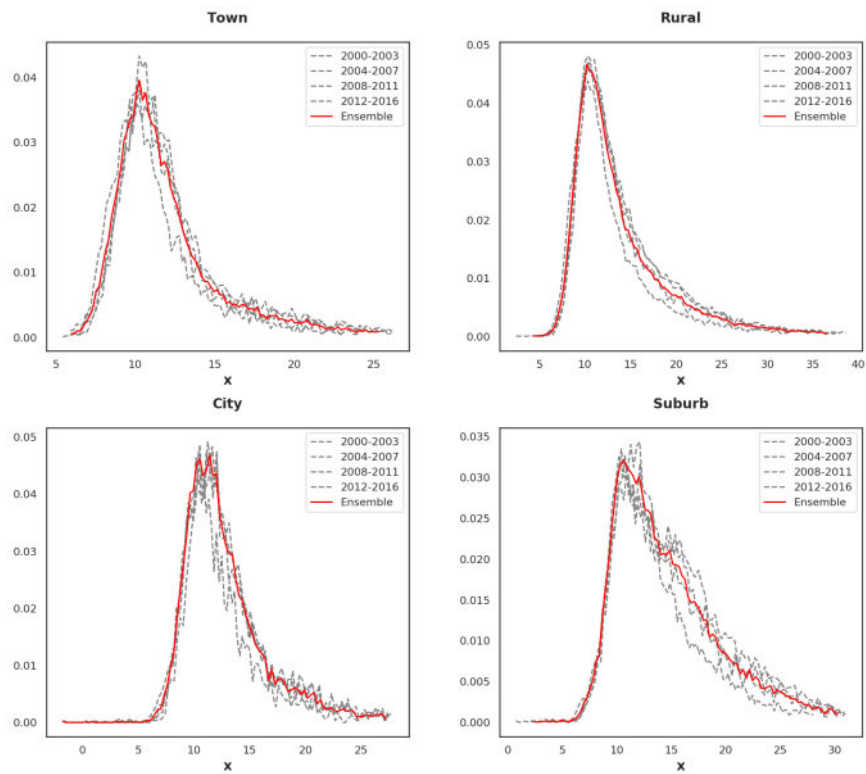


Fig. 4.9 Stacked histogram of grouped yearly subsamples across the four urban locales. 2000-2016.

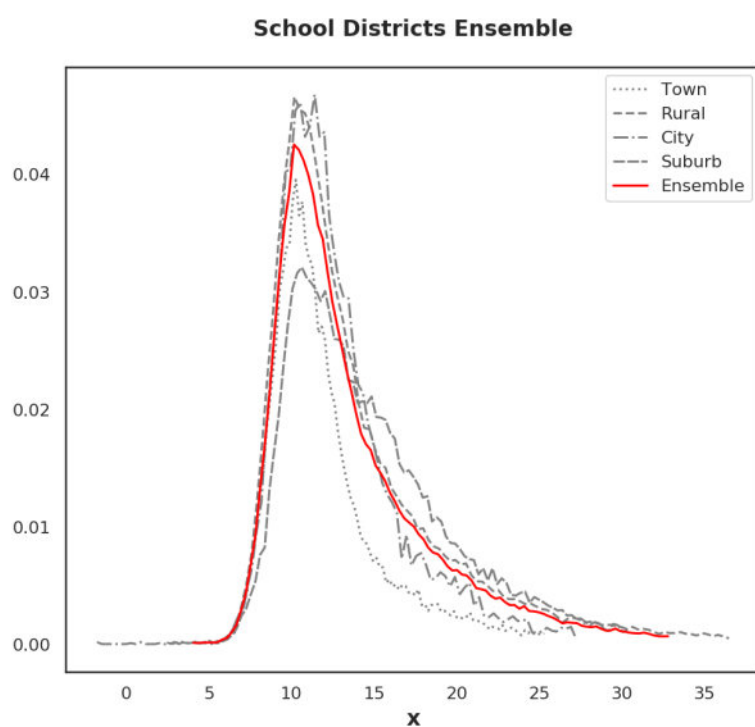


Fig. 4.10 Stacked histogram of urban locale subsamples and the full ensemble of school districts for the 2000-2016 period.

## 4.7 Results

### 4.7.1 Urban Locales

This subsection presents the results of Bayesian estimation for the four urban locales in the 2000-2016 period. The table in figure ?? gives the summary statistics for the estimated parameters  $T$ ,  $S$ ,  $\mu$ , and  $\alpha$ . Figure 4.11 shows a forest plot with the estimated 94% high-density intervals. For all urban locales, the model was also estimated across four distinct year groupings: 2000-2003, 2004-2007, 2008-2011, and 2012-2016. The line plot for the posterior means of this estimation case are shown in figures 4.12 and 4.13. The high-density intervals for this latter case can be found in the Appendix.

**Posterior Estimates Summary**

Locale	Parameter	Mean (Sd)	94 % HDI	$\hat{R}$
City	S	3.62 (0.99)	[1.67 , 5.41]	1.0
City	T	2.75 (0.98)	[0.93 , 4.62]	1.0
City	$\alpha$	18.23 (2.23)	[14.05 , 22.46]	1.0
City	$\mu$	8.36 (2.24)	[4.16 , 12.6]	1.0
Rural	S	5.02 (1.0)	[3.15 , 6.93]	1.0
Rural	T	2.09 (0.93)	[0.31 , 3.75]	1.0
Rural	$\alpha$	17.72 (2.23)	[13.51 , 21.88]	1.0
Rural	$\mu$	8.69 (2.23)	[4.48 , 12.84]	1.0
Suburb	S	6.09 (1.01)	[4.21 , 7.98]	1.0
Suburb	T	2.48 (0.97)	[0.62 , 4.27]	1.0
Suburb	$\alpha$	32.73 (2.24)	[28.51 , 36.9]	1.0
Suburb	$\mu$	7.41 (2.24)	[3.2 , 11.59]	1.0
Town	S	3.14 (0.99)	[1.25 , 4.99]	1.0
Town	T	2.28 (0.96)	[0.5 , 4.06]	1.0
Town	$\alpha$	14.17 (2.24)	[9.86 , 18.27]	1.0
Town	$\mu$	8.87 (2.23)	[4.7 , 13.06]	1.0

Table 4.1 Summary statistics of estimated parameters  $T$ ,  $\mu$ ,  $S$ , and  $\alpha$ . The mean, standard deviation, 94% credible interval and the convergence statistics  $\hat{R}$  from the MCMC samples are reported. Urban Locales: Town, Rural, City, Suburb. 2000-2016.



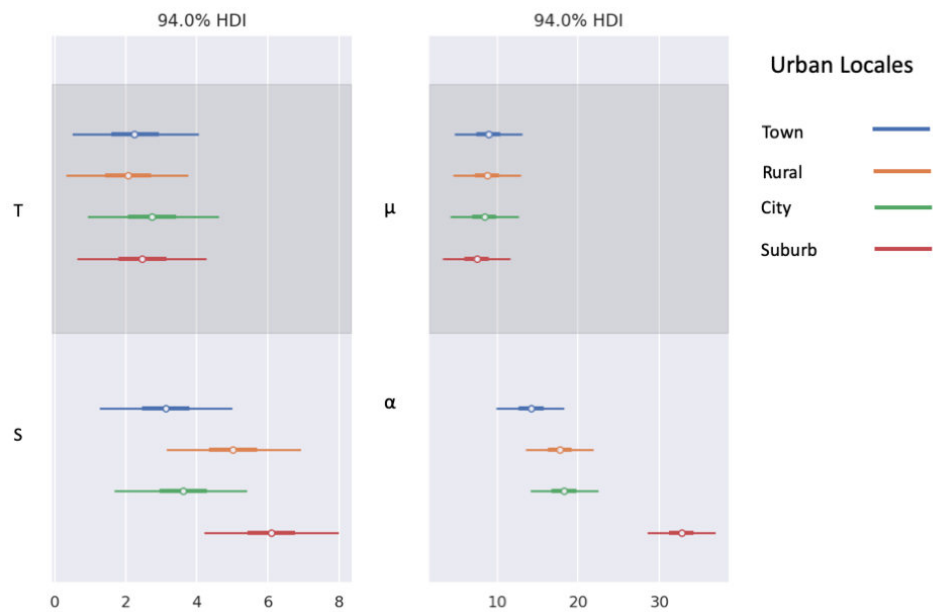


Fig. 4.11 Forest Plot with 94 % HDI Intervals. The middle bar plots the interquartile range for the posterior (0.25-0.75) and the median (central point). Urban Locales: Town, Rural, City, Suburb. 2000-2016.

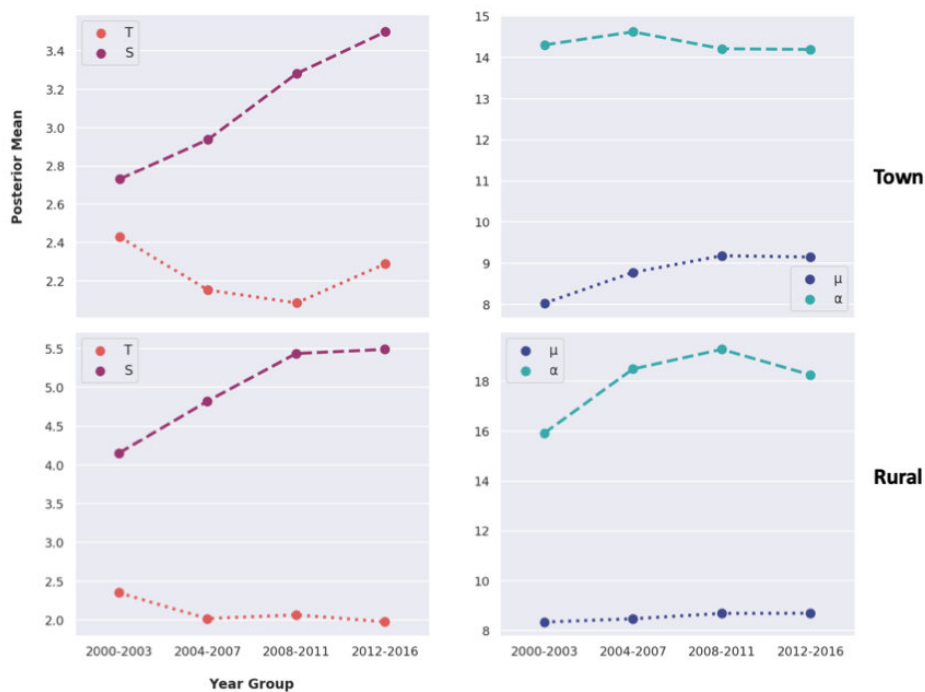


Fig. 4.12 Line plots for the estimated posterior means across four year groupings. Town and Rural, 2000-2016.

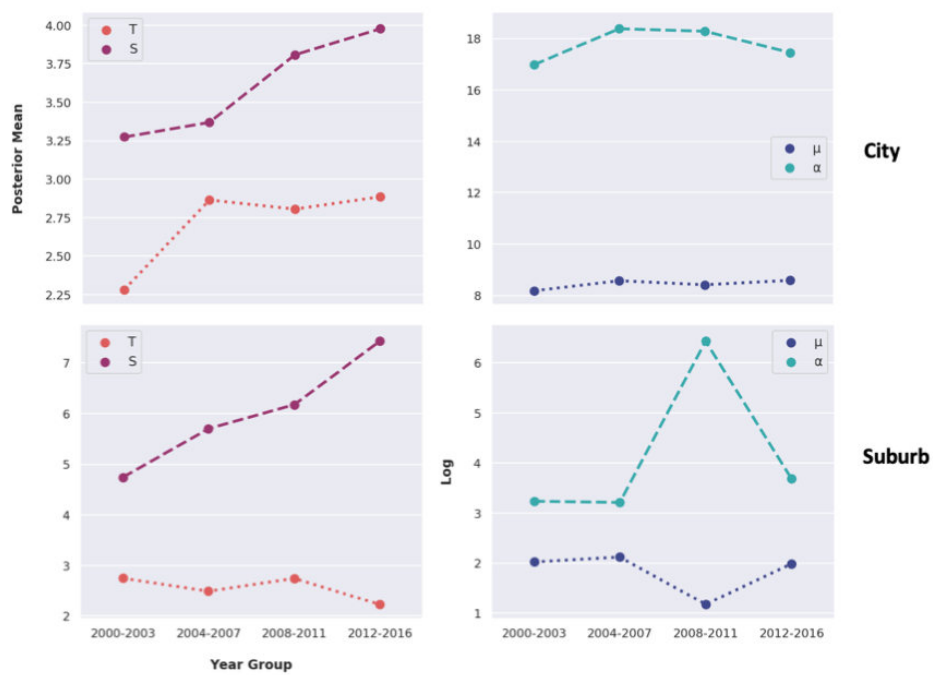


Fig. 4.13 Line plots for the estimated posterior means across four year groupings. City and Suburb, 2000-2016.

### 4.7.2 Income Groups

This subsection presents the results of Bayesian estimation for the three income groupings (Low, Middle, High) in the 2013-2016 period. The table in figure ?? gives the summary statistics for the estimated parameters  $T$ ,  $S$ ,  $\mu$ , and  $\alpha$ . Figure 4.14 shows a forest plot with the estimated 94% high-density intervals. We discuss the implications of these results in the following section.

**Posterior Estimates Summary**

Income	Parameter	Mean (Sd)	94 % HDI	$\hat{R}$
Low	S	4.04 (1.0)	[2.12 , 5.88]	1.0
Low	T	1.85 (0.9)	[0.13 , 3.34]	1.0
Low	$\alpha$	14.96 (2.24)	[10.8 , 19.2]	1.0
Low	$\mu$	8.88 (2.24)	[4.63 , 13.02]	1.0
Middle	S	5.8 (1.01)	[3.89 , 7.66]	1.0
Middle	T	2.12 (0.95)	[0.35 , 3.84]	1.0
Middle	$\alpha$	18.79 (2.23)	[14.68 , 23.04]	1.0
Middle	$\mu$	8.66 (2.23)	[4.54 , 13.01]	1.0
High	S	4.01 (0.99)	[2.13 , 5.84]	1.0
High	T	11.99 (1.0)	[10.08 , 13.85]	1.0
High	$\alpha$	60.0 (2.24)	[55.83 , 64.27]	1.0
High	$\mu$	1.01 (2.23)	[-3.12 , 5.24]	1.0

Table 4.2 Summary statistics of estimated parameters  $T$ ,  $S$ ,  $\mu$  and  $\alpha$ . The mean, standard deviation, 94% credible interval and the convergence statistics  $\hat{R}$  from the MCMC samples are reported. Income Groups: Low, Middle, High. 2013-2016.

In figures 4.17 and 4.20 we fit the estimated QRSE model to histograms of the observed distributions of  $x$  for the four different locales, and the three income groups, along with the joint action-outcome probabilities  $f_{a,x}$ . The Soofi information distinguishability statistic (Soofi ID; See Soofi and Retzer (2002)) is used to evaluate fit performance. The Soofi ID is shown below in equation 4.11. Smaller values of the KL-Divergence and of the Soofi ID imply better model fits and the Soofi ID in particular gives a measure of how much informational content is recovered by the candidate distribution. The histogram fits are shown in the next section to facilitate discussion.

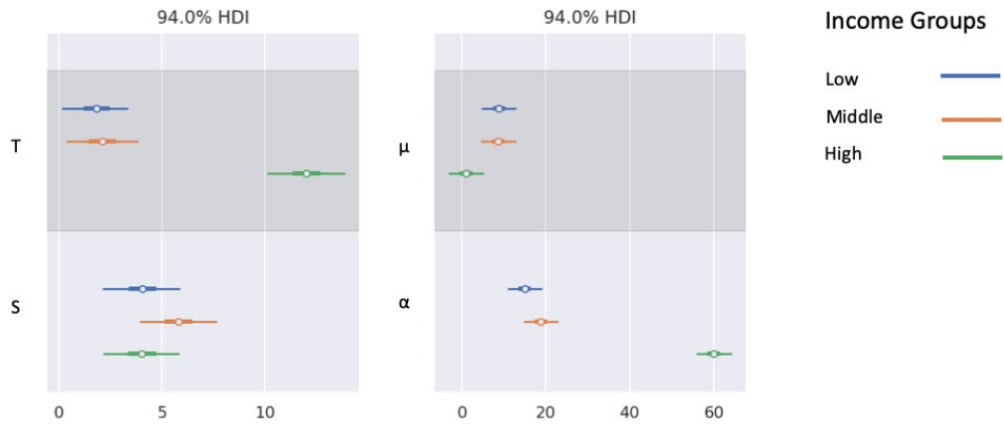


Fig. 4.14 Forest Plot with 94 % HDI Intervals. The middle bar plots the interquartile range for the posterior (0.25-0.75) and the median (central point). Income Groups: Low, Middle, High. 2013-2016.

$$ID(\hat{f} : \bar{f}) = 1 - \exp[-D_{KL}(\hat{f} | \bar{f})] \quad (4.11)$$

## 4.8 Discussion

### 4.8.1 Tiebout sorting and competition across urban locales

Figure 4.15 shows the estimated quantal response difference  $\Delta f_{a|x}$  and joint action-outcome probabilities  $f_{a,x}$  for the rural and town locales. Figure 4.16 shows the same plots for the city and suburb locales.

The estimated values of  $\hat{\mu}$  constitute the typical households' in- and out-migration indifference points, where  $\Delta f_{a|x} = 0$  and where  $f_{e,x} = f_{s,x}$ . These are then the estimated Tieboutian competitive rates for the education public good in each submarket ensemble.

Quantal response behavior and expectations appear as nearly identical across all locales, but the scale and breaking down of symmetry in the competitive feedback process is markedly different. Both the rural and suburb locales signal a relatively heavier dominance of sorting incentives in shaping the distribution of educational returns. In the suburb locale, the almost complete breakdown of competitive feedback is driven by the high value of  $\hat{\alpha}$ . As was discussed in the section 3, the large distance between  $\mu$  and  $\alpha$  can be understood in terms of unfulfilled public goods provisioning expectations for households seeking to sort on the basis of income and property values. This is a distinctive feature of the US public education market, where expenditure levels tend to be monotonic in property values. The similarities between the estimated quantal response functions reveal the role that states' minimum per pupil expenditure levels play in establishing the competitive Tiebout rate  $\mu$ .

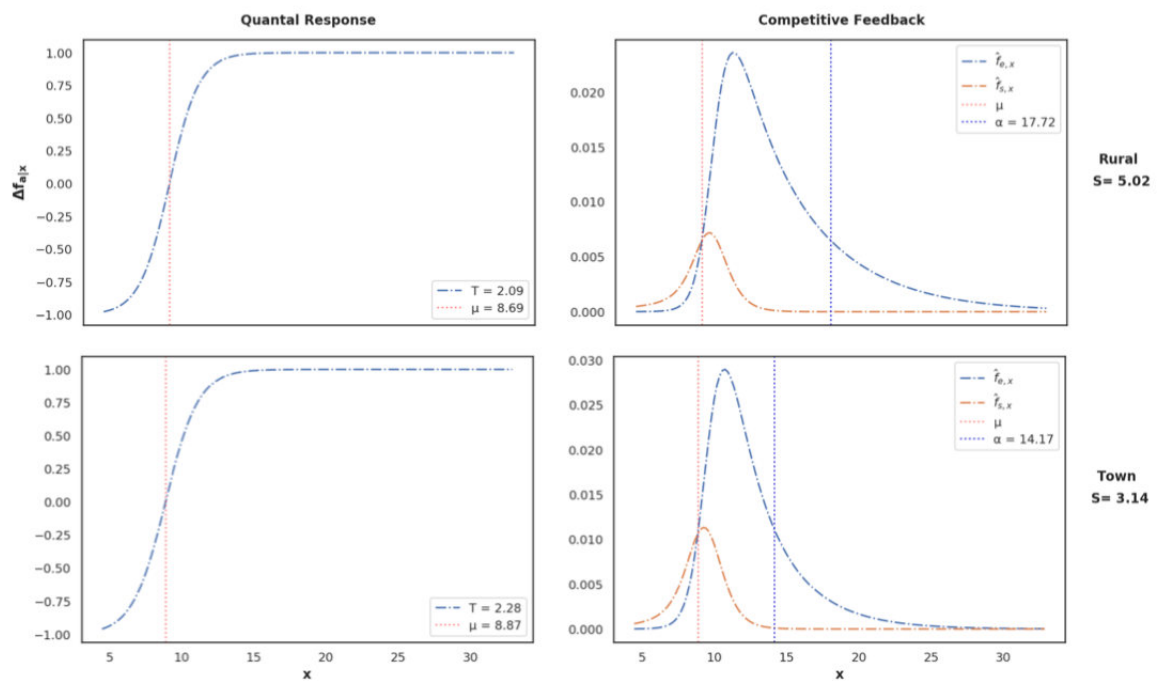


Fig. 4.15 Estimated quantal response and competitive feedback. Mean posterior values are used for plotting. Town and Rural, 2000-2016.

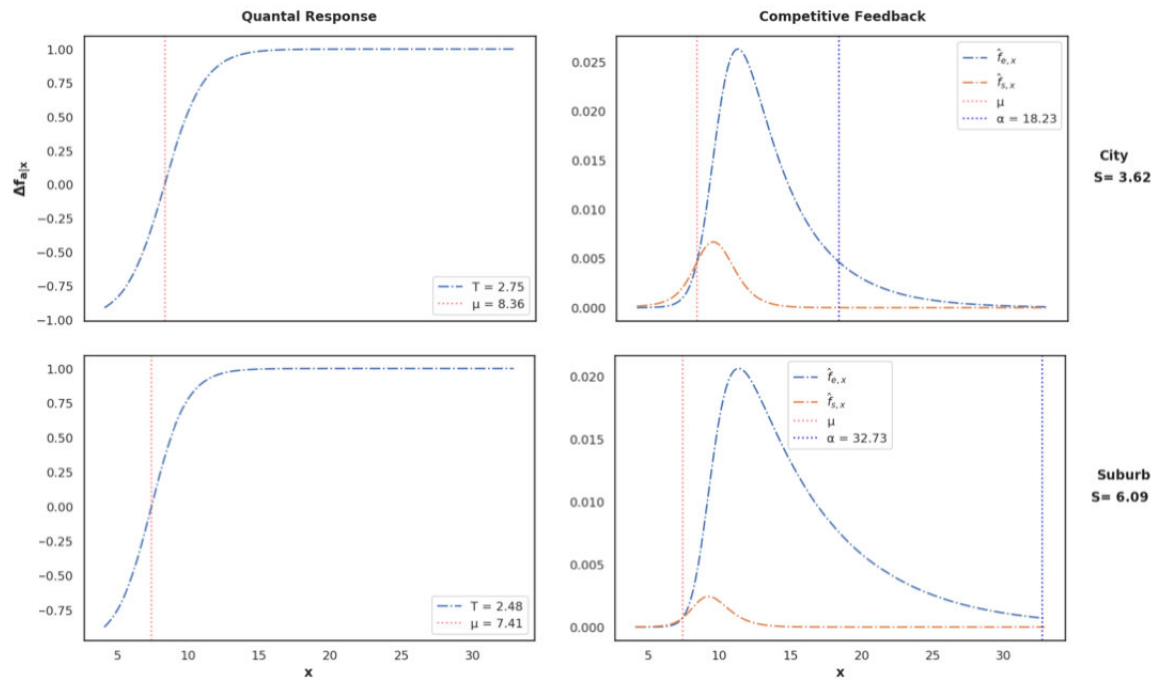


Fig. 4.16 Estimated quantal response and competitive feedback. Mean posterior values are used for plotting. City and Suburb, 2000-2016.

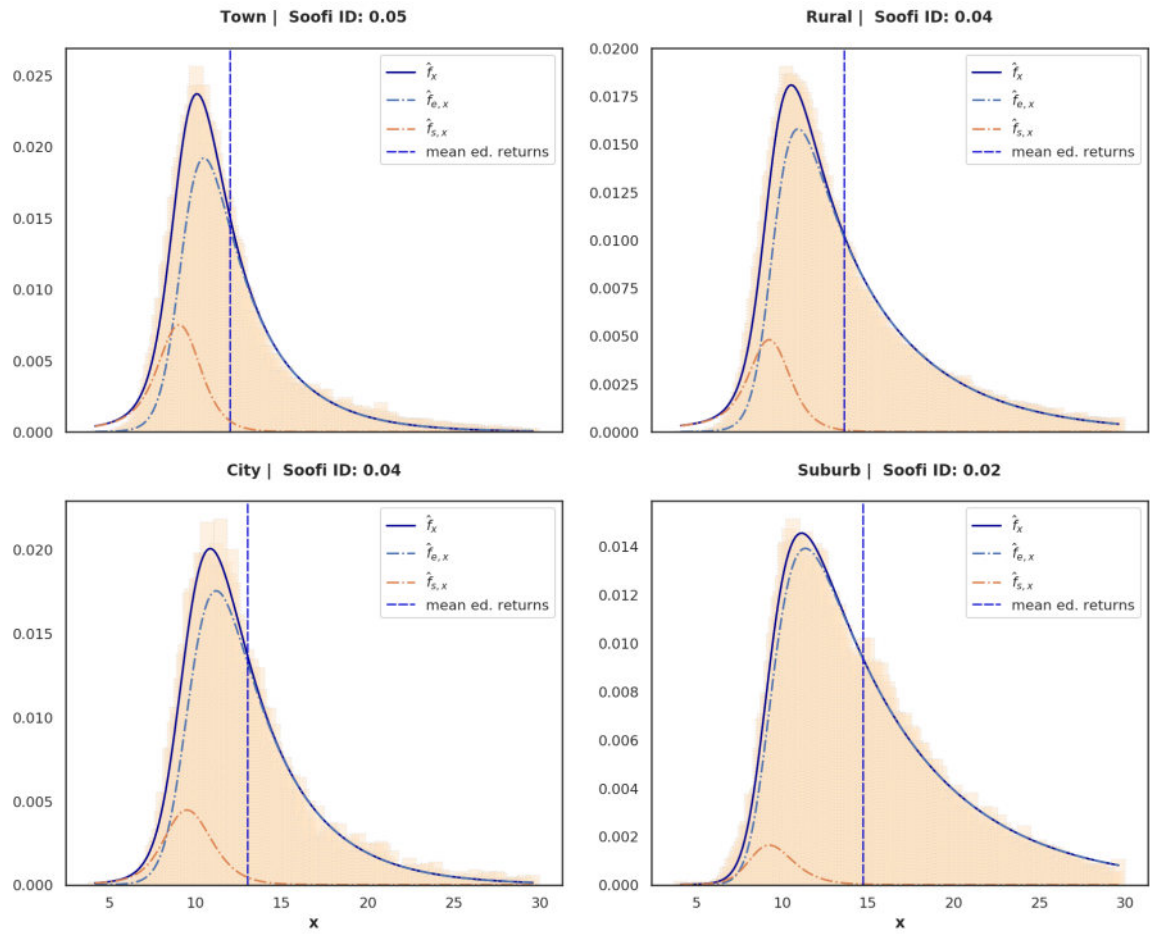


Fig. 4.17 Histogram of observed distributions for  $x$  (educational returns). Overlaid are the fitted marginal  $f_x$ , entry  $f_{e,x}$ , and exit  $f_{s,x}$  distributions. Urban Locales: Town, Rural, City, Suburb. 2000-2016.



## 4.8.2 Tiebout sorting and competition across income groups

Note that in the case of the low and middle income groups, the estimated posterior mean values for  $\hat{T}$  and  $\hat{\mu}$  fall within a similar range as that estimated for the urban locales case. This again gives supporting evidence to the ability of the T-QRSE model to infer district expenditure floors in terms of the competitive Tiebout rate  $\mu$ . In a decentralized public goods setting, households may not want to pay above this rate  $\mu$ , but they are driven to do so by the higher property taxes which are levied by districts in costlier housing markets. The basis for the interpretation of the high value of  $\hat{T}$  in the context of high-income districts was discussed in section 3. The results for this group, displayed in figures 4.19 and 4.20, point to a context in which educational returns are driven entirely by stratification incentives and independently of public competitive dynamics.

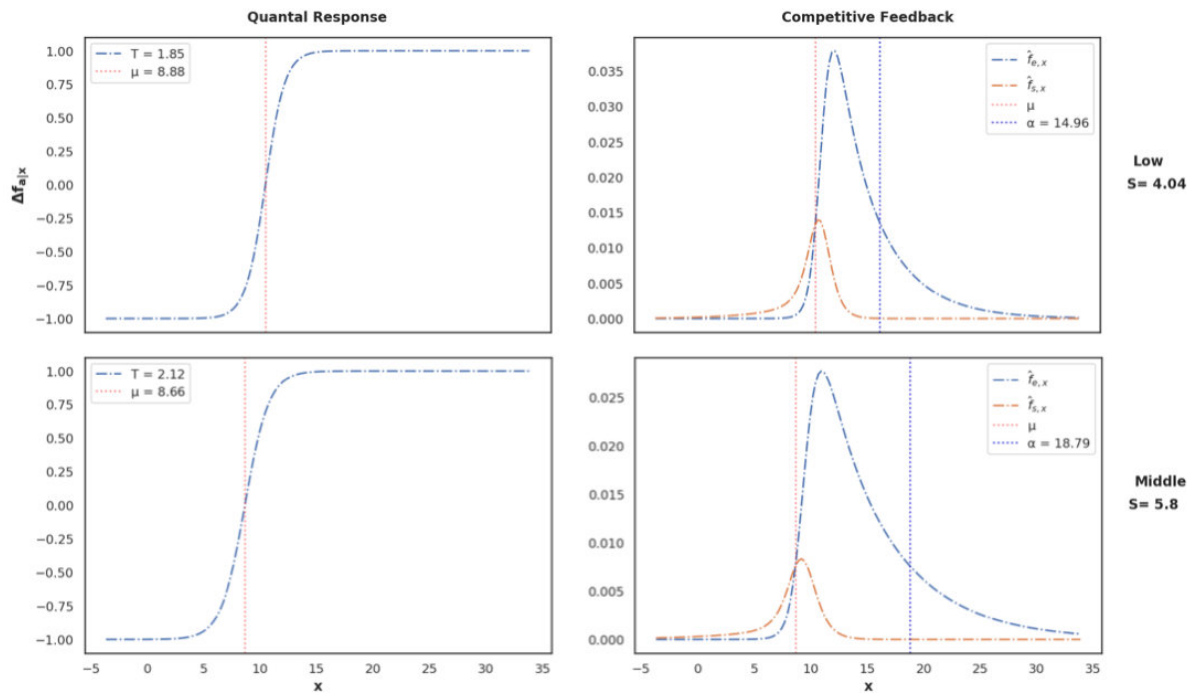


Fig. 4.18 Estimated household quantal response and competitive feedback. Mean posterior values are used for plotting. Low and Middle income groups, 2013-2016.

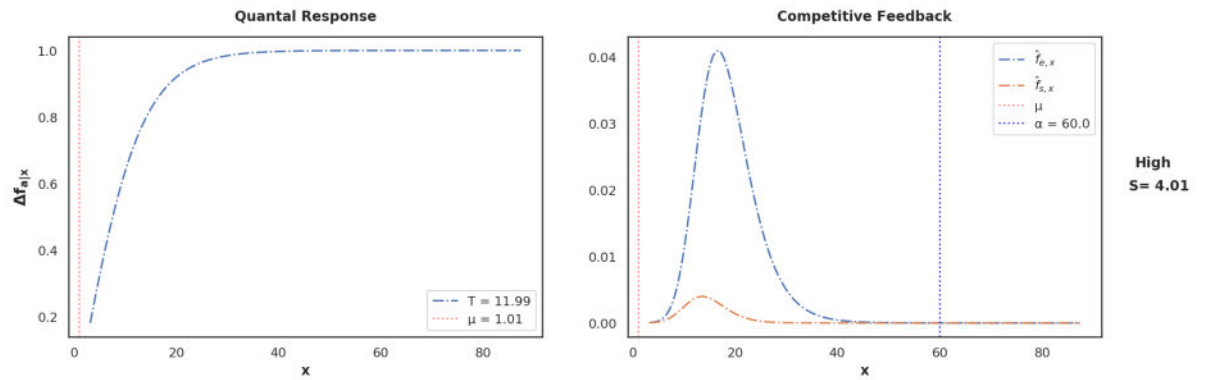


Fig. 4.19 Estimated household quantal response and competitive feedback. Mean posterior values are used for plotting. High income group, 2013-2016.

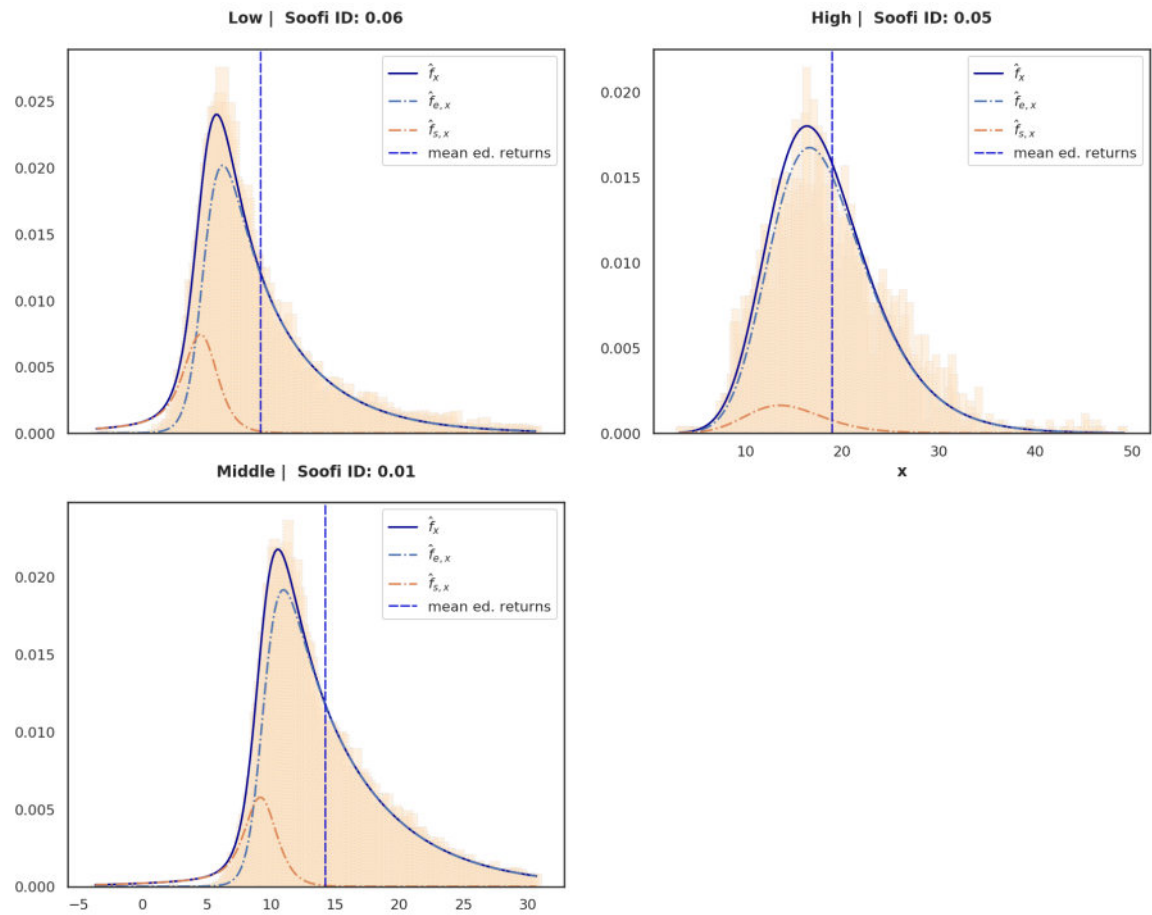


Fig. 4.20 Histogram of observed distributions for  $x$  (educational returns). Overlaid are the fitted marginal  $f_x$ , entry  $f_{e,x}$ , and exit  $f_{s,x}$  distributions. Income Groups: Low, Middle, High. 2013-2016.

### 4.8.3 Median property values and $\alpha$

This final subsection provides supporting evidence for the hypothesis that  $\hat{\alpha}$  estimates tend to be monotonic in income and property values. In figure 4.21 we plot school district median property values (MPV) against the income and locale categories used in the paper. In figure 4.22 we visually summarize the estimated posterior  $\hat{\alpha}$  means for all the school district groupings considered.

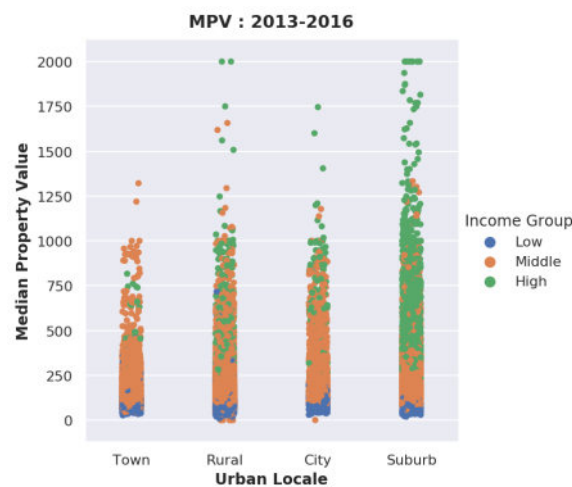


Fig. 4.21 Distribution of median property values for all districts; 2013-2016. Plotted against the income and urban locale categories.

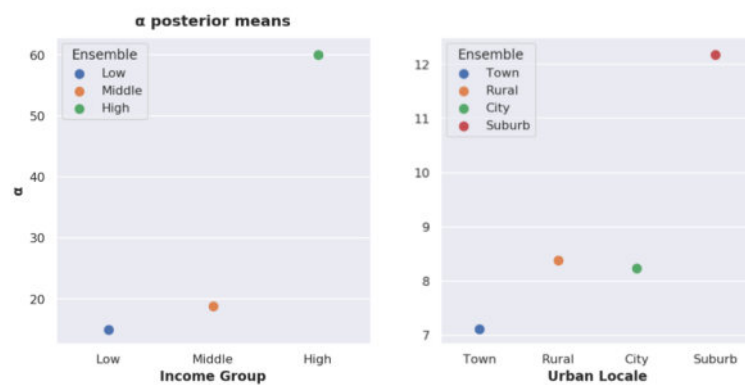


Fig. 4.22 Visual summary of the estimated posterior means for  $\alpha$ . All locales (2000-2016) and income groups (2013-2016).

## 4.9 Conclusion

The impact of Tiebout and sorting incentives on education expenditures across the urban-rural spectrum continues to be a central part of policy debates in the US. Using a quantal response statistical equilibrium model, this paper provided an inferential framework for estimating the relative importance of inter-jurisdictional competition and property-value driven sorting in shaping observed expenditure patterns across four urban locales for the period 2000-2016, and three income groups for the period 2013-2016. In Tiebout statistical equilibrium, sorting is a non-competitive outcome that can be understood as households' unfulfilled expectations with respect to a minimum expenditure floor that is determined at the state level.

## **Chapter 5**

# **Constructive statistical explanations: testing causal and non-causal principles in economic data with MaxEnt**

### **5.1 Introduction**

In everyday language we call random these phenomena where we cannot find a regularity allowing us to predict precisely their results. Generally speaking there is no ground to believe that a random phenomenon should possess any definite probability. Therefore we should have distinguished between randomness proper (as absence of any regularity) and stochastic randomness (which is the subject of probability theory). There emerges a problem of finding the reasons for the applicability of the mathematical theory of probability to the real world (...) In applying probability theory we do not confine ourselves to negating regularity, but from the hypothesis of randomness of the observed phenomena we draw definite positive conclusions. Kolmogorov (1983)

In section 1 I introduce the general model of statistical explanation as subsumption under generality, following Hempel's inductive-statistical schema. I then

translate this schema into a template that uses the more familiar language of probability models and distributions. In section 2 I summarize the key properties of constructive and non-constructive statistical explanations. Section 3 establishes the paper's working definitions of causal and non-causal explanations, and their role in supporting inference and prediction. In section 4 I build on Brian Skyrms' account of the maximum entropy principle (MEP) as a form of stochastic hypothesizing, and on Amos Golan's infometrics framework, to describe the role of MaxEnt in conducting modal inference, and in building constructive statistical explanations. This section also discusses maximum entropy methods against the backdrop of Jayne's information-theoretic and epistemic interpretation of statistical mechanics. In section 5, I take up as an economic example the problem of adducing a constructive statistical explanation for the observed distributions of firm profit rates, which are well known to be Laplace shaped. I conclude by discussing the causal and non-causal dimensions of the probabilistic competitive feedback constraint developed in Scharfenaker and Foley (2017).

## 5.2 Statistical Explanation

### 5.2.1 The Inductive-Statistical Schema

Early accounts of statistical explanation were given by Hempel et al. (1965) and Salmon (1971). The basic structure of Hempel's Inductive-Statistical (IS) model of explanation follows the general template of the well-known Deductive-Nomological (DN) model; that of explanation as subsumption under generality. To explain an event  $E$  is to provide an argument stating that on the premises that  $E$  pertains to some well specified reference class, and that the statistical regularity in question holds, we can infer with probability  $p(E)$  the occurrence of the explained event (Salmon, 1971; Sklar, 1995).

In the same way that the DN model is meant to capture explanation by adducing a deterministic (natural) law, the IS model explains by citing a 'statistical regularity' or 'statistical law' that explains the occurrence of  $E$ , with a certain level of 'inductive support'. The relationship between explanans and explanandum is not deductive, hence the name "inductive-statistical'. The general templates for the *DN* and *IS* models are displayed below in figures 5.1 and 5.2.

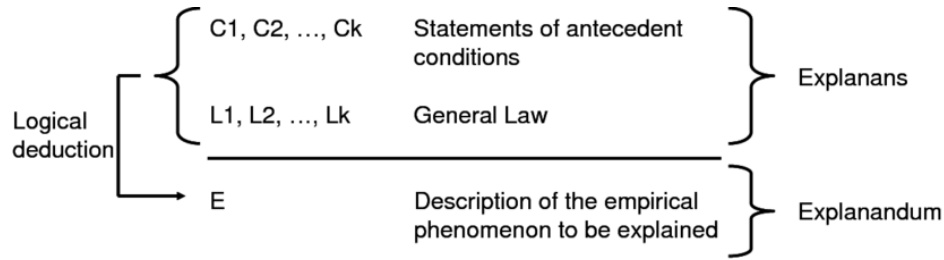


Fig. 5.1 Deductive-Nomological (DN) Explanation Schema

... a DN explanation answers the question “Why did the explanandum-phenomenon occur?” by showing that the phenomenon resulted from certain particular circumstances, specified in  $C_1, C_2, \dots, C_k$ , in accordance with the laws  $L_1, L_2, \dots, L_k$ . By pointing this out, the argument shows that, given the particular circumstances and the laws in question, the occurrence of the phenomenon was to be expected; and it is in this sense that the explanation enables us to understand why the phenomenon occurred. Hempel et al. (1965)

$$\frac{\begin{array}{l} p(G; F) = r \\ Fa \end{array}}{Ga} \quad (r) \quad \begin{array}{l} \text{(statistical law)} \\ \text{(particular facts about } a) \\ \text{(the explanandum)} \end{array}$$

Fig. 5.2 Inductive-Statistical (IS) Explanation Schema

In the *IS* template above,  $Fa$  stands for ‘facts about  $a$ ’. Thus, under the assumption of  $Fa$  and the conditional probability rule  $Pr(G; F)$ , we can infer  $Ga$  with probability  $r$ . The  $(r)$  in the entailment bar is meant to express the level of ‘inductive support’ for the explanatory schema.

### 5.2.2 Amending The IS Schema

What makes the general idea behind the *DN* and *IS* models relevant in describing explanatory schemas is their appeal to notions of ‘statistical regularity’ and ‘law-likeness’ in their explananda. Irrespective of the logical characterization used, the general concept of *subsumption under generality* turns out to be extremely useful

for exploring a host of philosophical issues related to the assertion of the 'law-like' aspect of the probability models that enter the explanans. As Salmon (1989) notes, "the essence of scientific explanation can be described as nomic expectability - that is, expectability on the basis of lawful connections".

There are naturally many issues to be raised against Hempel's account and they have been well documented in the philosophical literature (see Kitcher, 1989; Woodward, 1989). For the purposes of this paper, the following are worth mentioning. The first is that *IS* model construes statistical explanation as being inherently inductive; it takes the form an inductive syllogism. But modern probabilistic and statistical modeling, including both frequentist and bayesian statistics, can be shown to follow more closely a deductive-statistical or hypothetico-deductive framework than a strictly inductive scheme (Gelman and Shalizi, 2013). This straight-jacketing of the explanatory schema into an inductive inference framework fails to capture the separate roles that probabilistic and statistical claims play in explaining the explanandum. The second important limitation to mention in Hempel's account is that it restricts good explanations to those that are able to place high probability on the explanandum. Thus, the *IS* schema runs into trouble when attempting to account for low-probability events, such as a patient dying from the side effects of a medication that was meant to be curing. The drug is itself a probability raiser for recovery, but the statistical possibility of dying from the drug's side effects is real, even if extremely low.

These core issues are, however, easily amendable. With regards to the first issue, one can easily embed an inductive-statistical syllogism in a broader hypothetico-deductive framework or theory. The inductive-statistical schema only becomes limiting if one ignores the extra-statistical propositional base from which it is derived. Only machines can be reduced to blind inductivism; practicing empirical researchers inevitably weave together aspects of both deductive and inductive reasoning. Thus, I take the general shape of the inductive-statistical schema, and of *statistical explanation as subsumption under generality*, to be a useful characterization as long as it is embedded in a broader hypothetico-deductive framework.

With respect to the issue that the explanans has to be probability raising, we can drop the high probability requirement and simply ask that the explanans confer any positive probability ( $p > 0$ ) to the explanandum. In concrete probability terms, this means that as long as the explanans is able to account for the occurrence of



the observed event in *any quantile* of the theoretical distribution, then it counts as a possible candidate for statistical explanation (henceforth SE).

### 5.2.3 Statistical Explanation as Subsumption Under Generality

The logical shape of the schema of explanation as *subsumption under generality* can be suitably expressed using more transparent notation. We can simply take it to be a form of statistical syllogism, such as the one below:

$$\begin{array}{r}
 \text{I is an } F \\
 P \text{ proportion of } F \text{ are } G \\
 \hline
 \therefore \text{the chance that } I \text{ is a } G \text{ is } P \\
 \therefore I \text{ is a } G \text{ with frequency } P
 \end{array}
 \tag{5.1}$$

This statistical syllogism is useful in that it points to the two main essential components of a statistical explanation; the placement of the explanandum in a specific reference class, and the appeal to a probability law in the explananda that accounts for the chances or the observed frequencies of any particular event. The key then behind this standard pattern of statistical explanation, as mentioned in the previous sub-section, is the appeal to the nomic expectability of the outcome, under the assumption that the theoretical relationship between the reference class and the probability law for that reference class is suitably established. In the above example, the reference class is *F* and the probability law is *P*. The explanandum, i. e. the thing to be explained, is why *I* is *G* with frequency *P*. The core of the explanatory work is achieved by these two main components; the citing of the probability law or the statistical regularity, and the assertion that the random event belongs to a well specified reference class.

Starting out with that simple structure for SEs, an important of philosophical issues can begin to be addressed. How is the reference class established? Is it a classificatory kind? A natural kind? If the latter, are only natural kinds genuine generators of real probabilities (Sklar, 1995)? On what basis does the posited probability law explain anything about the observed outcome? Does the probability law contain non-causal and/or causal content? Or can it be posited

as being statistically autonomous ? In order to better address these questions, in the context of statistical equilibrium modeling, we can rewrite this general schema of explanation as *subsumption under generality* using the language of probability models and distributions. While it forces us to introduce some formal notation, it will make the discussion and the paper's arguments clearer.

### 5.2.4 Probability Models

Let me then briefly recall some working definitions of probability theory. A probability model or stochastic process  $M$  consists of a probability space  $S = (\Omega, \mathcal{F}, P)$  along with a set of indexed random variables  $\{X_{i \in I}\}$  defined on  $S$ .  $\Omega$  is the sample space,  $\mathcal{F}$  the  $\sigma$ -algebra of events (the collection of all subsets of  $\Omega$ ), and  $P$  a probability measure on  $\mathcal{F}$ . The random variable  $X : \Omega \rightarrow \mathbb{R}$  maps elements of  $\mathcal{F}$  to a measurable space, such as the real line, for which we can define a distribution function  $F_X(x)$  and a mass or density function  $f_X(x)$ . The distribution and density functions relate the probabilities of events to the probabilities of well-behaved subsets of  $\mathbb{R}$ . If  $B$  is such a subset, and we define an event  $E = \{\omega \in \Omega : X(\omega) \in B\}$ , then:

$$P(E) = P(X \in B) = \int_B f_X(x) dx$$

Note that  $F(x) = P(X \leq x)$ . This means that we can obtain, for continuous random variables on the real line, the density function via differentiation:

$$f_X(x) = \left. \frac{d}{du} F_X(u) \right|_{u=x} = F'_X(x) \text{ for all } x \in \mathbb{R}$$

The two fundamental properties that a density function must meet are *positivity* and *normalization*:

$$\begin{aligned} f_X(x) &\geq 0 \quad \forall x \\ \int_{-\infty}^{\infty} f_X(x) dx &= 1 \end{aligned}$$

In the case of mass functions and discrete variable we replace the above integral with  $\sum_x f_X(x) = 1$ . As will be discussed throughout the paper, the positivity and normalization constraints are only two of the many possible constraints that we may impose on  $f_X$ .<sup>1</sup> There is a broader set of constraints that we may impose on a probability model in order to derive distributions that meet certain theoretical or empirical requirements. If the constraints are theoretically motivated, then they will usually incorporate causal or modal information, i.e information about what can and can't happen in a given probability space, and about the causal hierarchies and dependencies between variables. If the constraints are empirically motivated, then they will be applied in order to restrict the density's moments (e.g. variance or kurtosis) to take on observed values. The application of such constraint rules is the main subject of maximum entropy inference.

### 5.2.5 The Probabilistic Basis of Statistical Explanation

In an effort to translate the general structure of *statistical explanation as subsumption under generality* using the standard language of probability models, we can start by thinking of explanandums (i. e. the data to be explained) as observed probability events, such as  $X = 1$  or  $X < 0$ . The event is any assignment of a value or a set of values to a random variable. Events such as "the coin lands heads", "the subject is older than 25", "the patient recovered"; any declarative statement that can be shown to be true or false or that may have a certain frequency of occurrence is then subject to statistical explanation under this general rubric. In terms of probability events, I now rewrite the explanatory schema of the statistical syllogism in 5.1 in the following form:

$$\begin{array}{l}
 T \models (\Omega, \mathcal{F}, P) \\
 E \in \mathcal{F} \\
 I := E \\
 P(E) = p \\
 \hline
 \therefore \bar{P}(I) = p
 \end{array} \tag{5.2}$$

<sup>1</sup>For ease of notation, throughout the paper when I write  $f_x$ , I am referring to  $f_X(x)$  and  $P(X = x)$

The first line in 5.2 defines the idea that the probability model is derived from some underlying theory  $T$ , or that the probability space  $S$  ‘models’ that theory. The third line  $I := E$  constitutes the reference class assumption, in that it deems the explanandum  $I$  as an observable event under the specified probability model. In practice, we are more interested in explaining the distinct patterns and shapes of distributions, and not of an isolated event or proposition. Using the same logical structure of 5.2, we can write a more detailed and relevant template in terms of probability distributions as follows:

$$\begin{array}{l}
 T \models Y \sim D(\Theta; \mathcal{R}) \\
 \{f_{\theta}; \theta \in \Theta\} \\
 f_{Y, \Theta} \\
 \hat{f}_y \\
 s(\hat{f}_y, \bar{f}_y) = \delta \\
 \hline
 \therefore \bar{f}_y
 \end{array} \tag{5.3}$$

The statistical syllogism in 5.3 provides a more complete benchmark to discuss statistical explanation as *subsumption under generality*, and captures the idea that inductive claims are embedded in a broader hypothetico-deductive framework. The first line points to the derivation of the SE’s probability model from some background theory  $T$ . The probability model here is expressed as a random variable  $Y$  that follows some distribution  $D$ , parametrized by the vector  $\Theta$ .  $T$  defines  $D$  as being appropriate or specific to a well-defined reference class  $\mathcal{R}$ . The second and third lines make the format more amenable to bayesian arguments by specifying priors for the model’s parameters, and a joint distribution for the random variable and the parameters ( $\Theta$  may also include latent variables in a bayesian formulation). The specification of the joint distribution allows for the simulation of new data instances from the underlying random process. Line 4 is the predicted marginal density of the theory  $T$  for the observed distribution  $\bar{f}_y$  (the explanandum). In practice, a bayesian model will produce a set of predictive densities  $\{\hat{f}_y\}_{\Theta|y}$  that will range over the posterior values of the estimated parameters. But for the purposes of this paper, this distinction will not be crucial. Finally, line (5) specifies a threshold value  $\delta$  for some similarity or distance measure  $s$ , which quantifies the ‘closeness’ between the empirical and

the theoretical distributions (e.g. KL-Divergence). An annotated version of this SE syllogism is displayed below:

$$\begin{array}{l}
 T \models Y \sim D(\Theta; \mathcal{R}) : \text{Distributional Assumption} \\
 \{f_{\theta}\} : \text{Priors} \\
 f_{Y, \Theta} : \text{Generative Model} \\
 \hat{f}_y : \text{Prediction} \\
 s(\hat{f}_y, \bar{f}_y) = \delta : \text{Similarity Measure} \\
 \hline
 \therefore \bar{f}_y : \text{Observed Frequencies}
 \end{array}$$

Not all SEs will be of this kind. Some SEs might not be operating under a bayesian framework, may not be ‘generative’ in the sense defined above, or may not even attempt to make explicit the link between the explanatory background theory  $T$  and the distributional assumption  $D$ . I use the sketch provided above as a benchmark to distinguish between *constructive statistical explanations* (CSE) and *non-constructive statistical explanations* (NCSE). Let me then call this benchmark sketch the constructive SE schema. I term *constructive statistical explanations* those SEs that follow this sketch and that make use of probability models whose parameters are directly interpretable in terms of the causal/non-causal content of a theory  $T$ . I will discuss the issue of causal and modal propositions in section 4, and in section 6 I will show how maximum entropy methods can be used to incorporate such propositions in the formal derivation of the predictive density function  $\hat{f}_y$ .

Those SEs that borrow their probability models from the general stock of statistical laws, or where the link between the distribution  $D$  and the observed frequencies  $\bar{f}$  relies only on moment matching and on a reference class assumption, I will call non-constructive. The next sub-section will focus on clarifying this distinction, and on giving some examples to ground the discussion. My objective here is not to approach anything resembling a definitive account of ‘statistical explanations’ by creating two mutually existing classes. There is a longstanding decline of systematic theories of explanation in philosophy of science, which nudges us to favor deflationary and

pragmatic accounts (Van Fraassen et al., 1980). Ultimately, it is more useful to speak of varieties of statistical explanation, and to pragmatically relate them to general aspects of actual scientific practice and to model-based reasoning. CSEs are one such variety, which I aim in this paper to relate to maximum entropy inference and to recent statistical equilibrium models in the social sciences.

## 5.3 Constructive Statistical Explanations

### 5.3.1 Key Properties

Following the exposition in the previous sub-section, I summarize in the table below the key properties for my proposed definition of Constructive and Non-Constructive Statistical Explanations (CSE & NCSE).

Property	Constructive SE	Non-Constructive SE
Probability Model	Derived from theory / parameters explicitly linked to causal, modal and moment constraints.	Subsidiary to existing probability model or derived from moment constraints. Uses background theory to interpret parameters.
Prediction	Predicts full distribution.	Predicts full distribution, or specific moments/ quantiles.
Generative Properties	Models joint distribution for all variables.	- Some model joint distribution - Some may rely on conditional expectations and distributions.
Criterion	Similarity or distance measure.	- Similarity or distance measure. - Maximizes conditional likelihood/ minimizes predictive error.
Causal Inference	Via constraint satisfaction on probability model.	- By estimating conditional causal effects

Fig. 5.3 Key Properties of CSEs

These proposed boundaries are soft. Some SEs will not fit strictly into these molds, and in practice they may merge aspects of both. The categories provide a benchmark against which we may compare specific examples. The main objective of my account is to put a strong emphasis on the constructive/non-constructive role of theory in specifying a probability model for the explanatory schema. The explanatory

role is ultimately fulfilled by the *proposed theory*  $T$ , not by any statistical model as a stand-alone object. The key problem is that of understanding how the distributional assumption  $Y \sim D$  is built into the explananda; whether it is explicitly derived from the background theory, or whether it is autonomously applied. This is the first distinguishing property for CSEs. The second distinguishing property operates at the level of causal inference. CSEs encode causal and modal propositions into the predictive distribution  $\hat{f}_y$ , while NCSEs may have to rely on estimating conditional causal effects.

The generative properties of an SE's probability model is not a distinguishing factor. Constructive SEs *necessarily* contain a joint distribution for all observed and unobserved variables  $f_{Y,\Theta}$ , as well as a set of predictive distributions  $\{\hat{f}_y\}_{\Theta}$  which are compared to empirical distributions via a similarity/distance measure  $s$ . Non-constructive SEs may be generative, but they may also be exclusively concerned with predicting quantiles or specific moments via conditional distributions (e.g.  $P(Y|X)$  or  $P(Y|\Theta)$ ). I take this then to be an optional property for NCSEs. The latter distinction turns out to be important in that some SEs may not be committed to explaining the full shape of an empirical distribution. Their subsidiary relationship to the underlying probability model may or not be problematic, that depends on the theory involved. Thus, the constructiveness of CSEs does not rely on the complexity of the probability model, but on the way that it accounts for model specification. Despite having a rich probabilistic structure, a large host of bayesian and structural models may fail in the task of supplying 'constructive' content to a statistical explanation if their specification is uninterpretable in terms of the underlying theory. A more succinct way to state the difference is that CSEs aim to show how probabilistic structure can emerge at the level of observed outcomes from a model's causal or modal constraints. NCSEs, in contrast, may delegate probabilistic structure to subsidiary distributional assumptions, to noise terms or 'exogenous shocks', or to complex sampling processes.

Using the inferential frameworks from statistical mechanics and information theory, along with modern probabilistic programming languages, it is becoming more approachable for social scientists to propose constructive statistical explanations that aim to show how probabilistic structure can emerge from the modal or causal principles that regulate any given system. Constructive explanatory schemas allow researchers to resolve theories about observables into distinctive statistical signatures

with interpretable parameters. If there is anything to take away from the definition of CSEs, it is this property. The resolution of theories into distinctive statistical signatures in CSEs opens up a fertile ground for phenomenological investigation and for counterfactual reasoning. A system's *observable stochasticity* may in principle be resolvable on deterministic or tychist terms, but it needs to be accounted for somehow by the proposed explanatory schema. CSEs do so by proposing a probability law whose parameters can be readily interpretable in terms of the underlying theory. Non-Constructive SEs do so via an autonomous application of a probability law, in which case the link between parameters and theory is settled outside of the derivation of the probability model. As stated earlier, this may or may not be problematic, and in practice usually constitutes a first step towards the probabilistic elaboration of a theory's core content.

## 5.4 Prediction, Causality and Modality

### 5.4.1 Prediction

I have been alluding to the role played by causal and modal propositions in building constructive SEs, whose main output is a predictive distribution  $\hat{f}_y$ , but have not properly defined these terms. In this section I will do so, in order to better understand the predictive output of the CSE schema.

Predictions are statements about a future event, or a missing piece of information, prior to the event taking place. Good predictions rely on good explanations, and there is a dynamic relationship between the two. It is often said that predictions are *conjectures*. But in what sense? What distinguishes, for example, a scientific prediction from an oracular statement? What builds up their distinctive epistemic status? Scientific predictions are *conjectural* in the sense that they rely on *plausible* explanations to state why a given event might go one way or the other. It is not just that they state a possible unfolding of future events, but that based on the premises supplied by the explanation on which they rely, they state a likely future to follow some particular pattern. Let me give an example. When in 67 AD emperor Nero visited the Oracle at Delphi, he was allegedly told :



*Your presence here outrages the god you seek . . . The number 73 marks  
the hour of your downfall.*

The oracle didn't cite any causes or principles for why that should be the case, and Nero interpreted the statement as meaning that he would live until the age of 73. The statement was eventually considered as being fulfilled after the fact when the Roman general Galba, who was 73 at the time, orchestrated a revolt and took over Nero's reign. Nero died by his own hand at age 30 after being exiled from power.

Such openly ambiguous and cryptic statements only support postdiction, or interpretation after the fact, they are not useful in any practical sense. But they are not useless because of their falsehood, nor because they may only be coincidentally fulfilled, but because they do not rely on any underlying explanation. They posit necessity *ex nihilo*, without citing causes or principles, and usually look like counterfactuals without antecedents. While the cryptic statements of the oracle of Delphi may seem remote to the modern reader, they are not too far off from the sort of predictions and forecasts that are asked of economists and pundits in policy and media outlets. Take, for example, a statement of the form 'The economy will enter into recession next year' or 'GDP is estimated to grow at a rate of 5% in Brazil'. If taken purely at face value as counterfactual possibilities with no antecedent clauses, these are purely speculative statements that carry no epistemic content. Now take a statement of the form 'If long term inflation expectations move persistently beyond our goals, we will use our tools to preserve price stability', delivered by Jerome Powell (head of the US federal reserve) earlier this year. This latter statement has a counterfactual and conditional structure that points to a reliance on an explanation of some sort; presumably the Fed has some conjectures about how inflationary and expectation dynamics work in the economy, based on which it predicts the possible stabilizing impact of its policies.

The point that I want to make is that With respect to the statements that are made about the future course of social systems, it is not always easy to distinguish between empty predictive/statistical claims and predictions that carry defeasible epistemic content. Furthermore, deciding what is deemed epistemically meaningful in the social sciences is not so easily settled, given the reign of 'as if' placeholders for theory and model-based reasoning, as well as the hard limits that complex social systems with many degrees of freedom pose to concrete causal knowledge and prediction. But if one is able to establish criteria for what constitutes a *good*

explanation, then it becomes possible to make the distinction, independently of the accuracy of the predictions which they may afford (Albin, 1998). In the account that I give here, I take explanations to be epistemically meaningful whenever they supply concrete causal and/or modal content, and if in supplying such content they provide the conditions for intelligibility of the explanandum in terms of an underlying theory. I hence take in this account *intelligibility*, and not *accuracy*, to be epistemically more relevant for scientific explanation. Predictions are hence useful whenever they rely on defeasible explanations that carry such type of content, and may turn out to possess epistemic value even when vastly inaccurate. As an example of an inaccurate prediction that turned out to be useful, Marletto (2021) cites Columbus' 15th century prediction that traveling westward from Europe one could reach the East Indies. The prediction was incomplete and imprecise, but it drew from a good explanation: that the Earth is round. Columbus' prediction relied on a rough but nonetheless causal assertion about the earth's curvature and the possible paths and endpoints that a boat of the time could reach when headed westwards from Europe.

There are, of course, the cases of spectacularly precise predictions that can come about from the type of knowledge afforded by classical mechanics. An often cited example in astronomy of a remarkable prediction that relied on causal/mechanistic knowledge is Le Verrier's discovery of Neptune. Le Verrier postulated the existence of a hidden planet as a possible explanation for the irregular orbital motions of Uranus. Using Newton's laws of motion, the hypothesis of a hidden planet allowed Le Verrier to explain Uranus' 'irregular motion' in terms of gravity-induced disturbances. The hidden planet was subsequently spotted to a high degree of accuracy by Johann Gottfried Galle in 1846, based on Le Verrier's calculations. This then established Neptune's existence as a celestial object (Bamford, 1996). Another important example worth mentioning in physics was Einstein's early work on brownian motion. In his 1905 paper "*On The Movement of Small Particles Suspended in Stationary Liquids Required By The Molecular-Kinetic Theory of Heat*", Einstein proposed to explain the irregular swarming motion of pollen seeds in water in terms of the kinetic-molecular theory of heat. He hypothesized that, if the kinetic theory was true, small but visible particles suspended in a liquid would be randomly bombarded by invisible atoms. The kinetic and statistical laws governing the atomic collisions would then explain the particle's 'irregular' brownian path and hence the macroscopic process of diffusion and drift. Jean Baptiste Perrin's experimental results subsequently confirmed Einstein's equations and allowed him to reliably estimate the size of

atoms, effectively raising their status from hypothetical to observable objects. The explanation of macroscopic laws in terms of the kinetic theory was a matter of heated debate between proponents of ‘energetics’ and ‘atomists’ in the 19th century, but Einstein’s and Perrin’s results decidedly tipped the balance in favor of the atomic theory (Sokolov and Klafter, 2005; Stachel, 2005).

Not all predictions, however, necessarily rely on causal/mechanistic knowledge. A well known example is the prediction of the neutrino, a subatomic particle first proposed Wolfgang Pauli in 1930 in order to make the experimental results obtained from studying beta decay processes consistent with the principle of energy conservation. The discovery of neutrinos relied entirely on the principles of conservation of energy and momentum, independently of their quantum/mechanical characterization which was only developed decades later (Cowan Jr et al., 1956; Marletto, 2021). When I have been referring to the, ‘modal’ content of theories above, I am referring to principles of this kind. The modal core of a theory specifies the conditions of possibility and impossibility of certain processes without exploiting details about micro-level laws. This does not mean that principle-based theories or discoveries are discordant with causal content, but only that by virtue of their greater generality, there is a broad set of causal descriptions that can plug into any given principle. The ultimate arbiter is the data; causal theories compete under the constraints of higher level principles to explain observations.

In the next sub-section I delve deeper into this distinction between causal and non-causal explanations, which is central to my account, given that I have defined CSEs as being based on theories that supply such content to their probability models.

### 5.4.2 Causal and Non-Causal Explanations

Causal explanations in physics, of the mechanistic sort, appeal to *laws of motion*. Given a set of initial conditions and a formalism for describing a physical system (e.g. the hamiltonian of a system of  $N$  classical particles), a set of dynamical laws allows for the deterministic prediction of a system’s evolution in state space. When I speak here of causal/mechanistic accounts I refer to accounts that are based on dynamic-law approaches, such as the force laws of classical mechanics. Outside of physics, in biology and in the special sciences (e.g economics), other concepts of causation may be more relevant, such as activity-based accounts, counterfactual

accounts, process theories and probabilistic causation (Pearl, 2009). In this paper I refer more generally to explanations as being causal if they build upon descriptions of a network of causal relations that is relevant to the explanandum's reference class. Thus I am steering away from focusing exclusively on reductionist, physicalist, or strictly mechanistic accounts of causation. The latter I subsume under dynamical-law approaches. I am more interested in clarifying the distinction between causal and non-causal explanations, and in understanding how they both support the predictive output of constructive statistical explanations.

The distinction between causal and non-causal explanations mirrors that between bottom-up and top-down approaches to scientific explanation (Kitcher, 1989; Salmon, 1989). "An explanation taking a bottom-up approach describes the causal processes, interactions, and (often hidden) mechanisms responsible for particular occurrences or general regularities. In contrast, an explanation taking a top-down approach subsumes the explanandum under some extremely general principles" (Lange, 2016). The general principles of the top-down approach can be understood as constraints on the causal regularities that a process may exhibit; they are modally stronger and are often described as *laws of laws*. This distinction between natural laws and principles is clearly stated by Feynman: "When learning about the laws of physics you find that there are a large number of complicated and detailed laws, laws of gravitation, of electricity and magnetism, nuclear interactions, and so on, but across the variety of these detailed laws there sweep great general principles which all the laws seem to follow. Examples of these are the principles of conservation" (Feynman, 2017). And by Einstein, with respect to the principle of relativity:

The principle of relativity . . . is not to be conceived as a 'complete system', in fact, not as a system at all, but merely as a heuristic principle. . . . It is only by requiring relations between otherwise seemingly unrelated laws that the theory of relativity provides additional statements.

For example, the theory of the motion of electrons arises in the following way. One postulates the Maxwell equations for vacuum for a system of space-time coordinates. By applying the space-time transformation derived by means of the system of relativity, one finds the transformation equations for the electric and magnetic forces. Using the latter, and applying the space-time transformation once again, one arrives at the law for the acceleration of an electron moving at arbitrary speed from

the law for the acceleration of the slowly moving electron (which is assumed or obtained from experience). Thus, we are not dealing here at all with a ‘system’ in which the individual laws are implicitly contained and from which they can be found by deduction alone, but only with a principle that (similar to the second law of the theory of heat) permits the reduction of certain laws to others. (Einstein, 1907)

What is important to note in the contrast between force laws and principles is that the distinction has to do with their degree of necessity. The explanatory power that can be derived from principles such as the conservation laws, the second law of thermodynamics, or relativity, comes from the constraints that they establish on the kinds of mechanical interactions that can and can’t happen. The added modal strength (i.e. necessity) that such principles enjoy make them true independently of the nomological details of the forces involved. This idea of ‘modal strength’ can be made more precise if we reason in terms of counterfactuals. Note that the impossibility of a system that violates the conservation of energy is stronger than the impossibility of a system, for example, that violates Newton’s gravity law. We can think of a possible world where the conservation of energy law still holds, but where the gravitational forces are repulsive rather than attractive<sup>2</sup>, or where they have different constants, or obey different inverse force laws. Lange (2020) cites the famous example of Paul Ehrenfest, who in 1917 ‘showed that had gravity been an inverse-cube force or fallen off with distance at any greater rate, then planets would eventually have collided with the sun or escaped from the sun’s gravity’ (Ehrenfest, 1917). What is interesting about the Ehrenfest example is that it requires for Newton’s second law to hold, independently of the rate at which gravity falls off with the distance. What the analysis of such counterlegals points to is that the natural impossibility of an inverse-cube gravity force is not quite the same as that of the second law, or of the conservation principles. There is a spectrum of modal strength along which we may place accidents, natural laws, physical principles and mathematical truths, each respectively ranging over a broader set of counterfactual possibilities.

The question is then in what sense can non-causal principles be used as explanations ? How do they supply content to the sort of statistical explanations and predictions that I have been alluding to throughout the paper ? The straightforward

---

<sup>2</sup>This theory was actually proposed by Le Sage in the 18th century.

answer is that they come in handy in contexts where we may ignore the nomological and causal regularities that make up a complex process, but where we have some sense of a broad and stable set of counterfactuals that the process may range over. They may not predict the specific laws in a given physical context, as pointed out by Jeans in the quote below, but in supplying a hypothetical space that bypasses and constraints the lower level details ( without undermining them), they allow us to predict what is possible and impossible in any given context.

The hypothesis of relativity predicts that a freely moving planet cannot describe a perfect ellipse around the sun as focus. This prediction is made on quite general grounds, just as the conservation of energy predicts that a stream of water cannot flow uphill. But the conservation of energy by itself is powerless to predict what will be the actual course of a stream of water, and in precisely the same way the hypothesis of relativity alone is powerless to predict what will be the orbit of a planet. Before this or any other positive gravitational predictions can be made, additional hypotheses must be introduced. (Jeans, 1921)

Non-causal explanations, however, can be useful beyond their ability to exclude counterfactuals. It is possible to ask the question of how a system's probabilistic structure can emerge from modal principles alone. To clarify this point, consider the problem in statistical mechanics of inferring the particle density per unit volume  $n_h$  of an ideal gas in a volume  $V$  as a function of height  $h$ . Suppose the gas has  $N$  molecules and is contained in a an extremely tall tube of height  $H$ , held at a constant temperature  $T$ , as illustrated in figure 5.4. This is an adaptation of Feynman's well known example in Feynman (1965).

Atoms with energy levels  $\epsilon_i$  randomly collide in the tube and  $\epsilon'_i + \Delta\epsilon$  is the energy level after each collision. If energy is conserved in this system,  $\Delta\epsilon$  constitutes an energy exchange between the atoms. This process is illustrated below in figure 5.5.

What is the probability of finding any given molecule at a height  $h \in [0, H]$  ? Without detailed knowledge of the underlying mechanics of this gas we can try to obtain the equilibrium distribution of  $h$  by using only 'higher level laws' such as energy conservation and time-reversal symmetry for the forces involved. The energy conservation principle is given by the additive identity in equation 5.4.

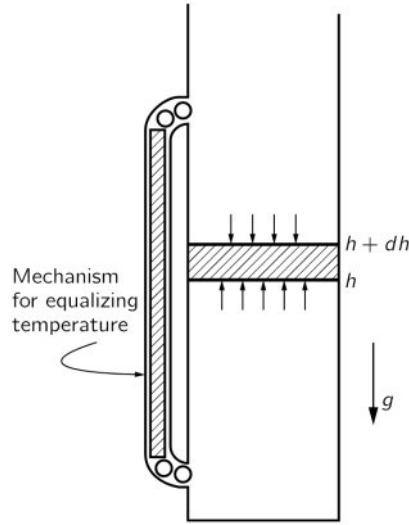


Fig. 40-1. The pressure at height  $h$  must exceed that at  $h + dh$  by the weight of the intervening gas.

Fig. 5.4 From (Feynman, 1965)

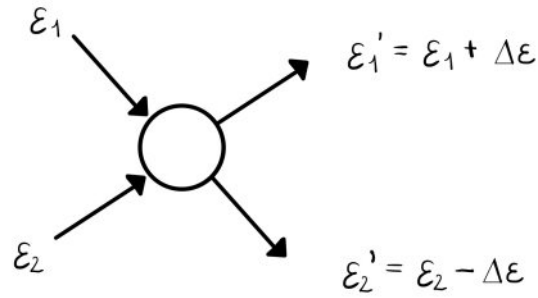


Fig. 5.5 From (Dragulescu and Yakovenko, 2000a)

$$\epsilon_1 + \epsilon_2 = \epsilon'_1 + \epsilon'_2 \quad (5.4)$$

If we reason in terms of time-reversal symmetry by considering that in equilibrium, the probabilities of transition between microstates  $l$  and  $j$  in the gas must be equal in reverse order, i.e.  $\pi_{j \rightarrow l} = \pi_{l \rightarrow j}$ , then the detailed balance equation in 5.5 simplifies to equation 5.6.

$$\pi_{12 \rightarrow 1'2'} p(\epsilon_1) p(\epsilon_2) = \pi_{1'2' \rightarrow 12} p(\epsilon'_1) p(\epsilon'_2) \quad (5.5)$$

$$p(\varepsilon_1)P(\varepsilon_2) = p(\varepsilon'_1)p(\varepsilon'_2) \quad (5.6)$$

The only stationary distribution for the energies that satisfies both the detailed balance equations and the energy conservation law in 5.4 is the Boltzmann-Gibbs distribution, an exponential function of the form:

$$p(\varepsilon) = Ce^{-\varepsilon/\langle\varepsilon\rangle} \quad (5.7)$$

where  $\langle\varepsilon\rangle$  is the mean energy level of the gas (Dragulescu and Yakovenko, 2000a; Landau and Lifshitz, 2013). Since  $h$  is a condition for the energy (the only degree of freedom we're considering), then by the previous argument the density is proportional to the Boltzmann-Gibbs law;  $n_h \propto e^{-\varepsilon/\langle\varepsilon\rangle}$ .

This leaves us with a great deal of uncertainty with respect to the underlying causal process in the gas, as it is not a *family of distributions* but a *family of functions* that is implied by this principle-based derivation. The 'additional hypotheses' that Jeans pointed to in the earlier quote would need to enter here at the level of specifying just what  $\varepsilon$  is and how it depends on the degree of freedom  $h$ , and what a constant  $n_0$  would need to be in order to fully specify the distribution. If we introduce knowledge that the gas tube stretches a very tall height from sea level to the atmosphere, and if we're only interested in studying the effect of gravitational potential energy, then we can fully specify the formula and bring it down from a family of functions to a family of distributions:  $n_h = n_0 e^{-mgh/kT}$ . If we know the weight of the molecules in different gases, we can recover the specific exponentials for each, as shown in 5.6.

In the other direction, with knowledge of the energy function and the required constants, the density can also be derived as direct consequence of the ideal gas law  $P = nkT$ , where  $n = N/V$ , independently of any higher level constraints (Feynman, 1965). But having started with the top-down derivation first, it is clear that this is not just coincidence. The modal principles tell us that independently of the causal context and atmosphere in which we place the gas tube, for any given macrostate, the probabilities of its possible microstates decrease exponentially as a function of potential energy, which is a remarkable proposition.



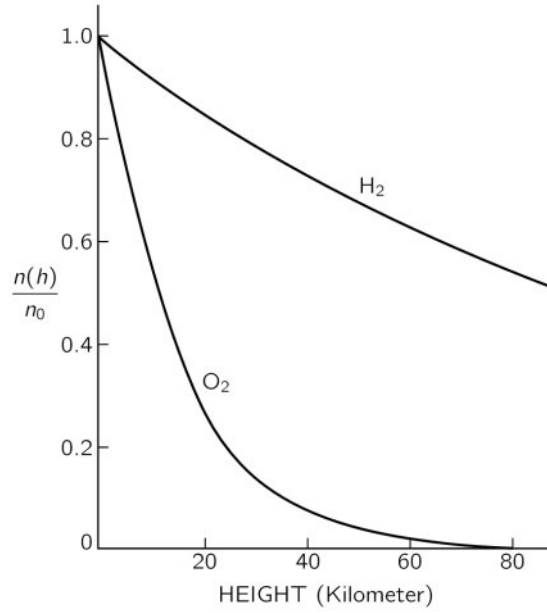


Fig. 40-2. The normalized density as a function of height in the earth's gravitational field for oxygen and for hydrogen, at constant temperature.

Fig. 5.6 From Feynman (1965)

Note, however, that the detailed balance condition in 5.5 is stronger than what is required to obtain a statistical equilibrium distribution (Lifschitz and Pitajewski, 1983). We may ask if it is possible to obtain the Boltzmann-Gibbs law in the case where we only know what the mean energy value should be, as in the usual case of  $\bar{\epsilon} = kT$ , which may be obtained using a thermometer. We can recover the density  $p(\epsilon)$ , exploiting only this measurement, by maximizing the entropy of the energy  $\epsilon$  subject to normalization and the expected value of  $\epsilon$  (Golan, 2018). This maximization program takes the form given below in 5.8:

$$\begin{aligned}
 \max_{p(\epsilon) \geq 0} H_\epsilon &= - \sum_{\epsilon} p(\epsilon) \log[p(\epsilon)] \\
 \text{st.} & \\
 \sum_{\epsilon} p(\epsilon) &= 1 \\
 \sum_{\epsilon} \epsilon p(\epsilon) &= \bar{\epsilon}
 \end{aligned} \tag{5.8}$$

The associated Lagrangian  $\mathcal{L}$  to this program is given below in 5.9:

$$\mathcal{L}_{\varepsilon, \lambda_1, \lambda_2} = H_{\varepsilon} + \lambda_0 \left( 1 - \sum_{\varepsilon} p(\varepsilon) \right) + \lambda_1 \left( \bar{\varepsilon} - \sum_{\varepsilon} \varepsilon p(\varepsilon) \right) \quad (5.9)$$

The first-order conditions to this optimization problem yield the statistical equilibrium density for  $p(\varepsilon)$ :

$$p(\varepsilon) = \frac{1}{Z} e^{-\varepsilon \lambda_1} \quad (5.10)$$

where  $Z$  is a summation constant:

$$Z = e^{-1-\lambda_0} = \frac{1}{\sum_{\varepsilon} e^{-\varepsilon \lambda_1}}$$

and  $\lambda_1 = \frac{1}{kT}$ . Thus, via the maximum entropy formalism, in equation 5.10 we recover the same distribution of energies as in 5.7 by alluding to energy conservation alone and without recourse to time-reversal symmetry, which is too strong a condition for obtaining the statistical equilibrium density. This means that if the molecules of a gas follow some reversible stochastic processes, then their energy distribution will be entropy maximizing, but a maximum entropy distribution of the energies does not imply mechanical reversibility.

The important point to make is that the same functional form (Boltzmann-Gibbs) appears in both derivations, not by mere accident or coincidence, but because of the greater generality of the conservation principles. In terms of counterfactuals, this means that the set of possible worlds that are energy conserving is broader than the set of those in which the molecules of isothermal gases can be described by reversible mechanical processes. We can continue to climb the ladder of generality and consider a situation in which we have no knowledge whatsoever about the underlying system, not even whether it is physical or not, but only that it emits some measurable and quantizable quantity  $y$  whose empirical mean is  $\bar{y}$ . If we again maximize the entropy subject only to normalization and the empirical mean constraint we will find that  $f_y \propto e^{-y\lambda}$ , where  $\lambda = \frac{1}{\bar{y}}$ . Thus  $f_y$  may as well describe the distribution of energies in an isothermal gas, the distribution of money and income (Dragulescu and Yakovenko, 2000a), or the outcome probabilities of a biased die roll (Jaynes, 2003; Sivia and Skilling, 2006). In that case, and in the absence of any other theoretical input or

explanatory framework, what explains the shape of the distribution is an information-theoretic and a statistical argument; there are no principles or causes involved. The maximum entropy framework exploits the full combinatorial range of the state space to recover the most likely distribution of states given the constraints, by finding the argument that maximizes a functional of probabilities  $F(\{p_i\})$ .

But the great generality that is afforded by this ‘distribution of distributions’, or by any other deep combinatorial argument, cannot not supply the conditions for explanation on its own; the weight is entirely on  $\lambda$ . If what is needed is a constructive statistical explanation of why a system is exponentially distributed, work must be done to make the constraints in the derivation theoretically interpretable. What are the counterfactual properties that a system must possess in order to be exponentially distributed? What modal content can we supply to the maximum entropy program in order to obtain a family of distributions that contains our observations? What causal knowledge and constants could allow us to specify the density uniquely?

By the arguments above, I have tried to show that what makes a statistical explanation constructive is the theoretical interpretability of its predictive distribution  $\hat{f}_y$ , in terms of both causal and non-causal constraints that are explicitly derived from theory. And because I defined constructive statistical explanations as supporting both causal and modal inference, a simple bottom-up/top-down distinction would have been lacking.

## 5.5 Distributional Supposition and Statistical Autonomy

### 5.5.1 Distributional Supposition

The motivation behind providing the definitions of CSEs as given in the previous sections is to make explicit the idea that statistical explanations embody a certain set of distributional assumptions which may themselves be derived from an underlying theory. Maximum Entropy (ME) methods provide one avenue for building CSEs, but there are naturally other routes, some of which are direct bottom-up derivations of frequencies. What makes ME methods central in my account is that they allow for the incorporation of both causal and non-causal content into the predictive density

$\hat{f}_y$ . I have also tried to show why ‘non-causal’ or ‘modal’ content does provide ground for building good explanations (Lange, 2016; Marletto, 2021), and why it may support meaningful probabilistic inference in complex systems with many degrees of freedom where knowledge of the causal nomological regularities is absent or ill-determined. Since the weight of explanation lies entirely on the constraints, and on how they are interpreted or constructed, the appeal to maximum entropy alone usually does not constitute sufficient grounds for building constructive statistical explanations. Thus, statistical explanations featuring maximum entropy distributions that are derived on the basis of moment matching alone are not ‘constructive’ under this account.

The work of (Jaynes, 2003; Rosenkrantz, 2012) and Golan (2018) has showed how many of the methods coming out of statistical mechanics can be distilled into general templates for conducting inference and plausible reasoning on the basis of partial information. The generality of the resulting ‘epistemic statistical mechanics’ and ‘infometrics’ frameworks is fully established on information-theoretic and statistical grounds, without appeal to physical or ergodic postulates. It is this particular feature which broadens the scope of applicability of maximum entropy methods for inverse inference problems and theory formation outside of physics.

This great generality of the ME methods affords the possibility of hypothesizing about why quantities are distinctively distributed in the way that they are, in a way that goes beyond information conditioning. The difference is subtle but non-trivial; the conclusions that I can draw about  $x$  will be different if I use  $p(x|I)$  or  $q(x|I)$ . What reasons do I have to suppose  $p$  or  $q$ ? Bayesian updating alone will not answer this question, as the answer requires the introduction of extra-statistical hypotheses. Brian Skyrms has rightly termed MaxEnt a kind of stochastic hypothesizing. In Skyrms (1987) he explains succinctly the distinction between the maximum entropy principle and the bayesian updating rule, and uses the Stalnaker logic of subjunctive conditionals to frame the discussion in terms of possible world semantics (Stalnaker, 1968). The procedure of bayesian updating takes us from one point in the interior of the convex set of a probability distribution to another. MaxEnt supposition, in contrast, moves us to a different chance distribution altogether.

A Warrenite asked to update on the piece of information that Oswald didn’t kill Kennedy would come to the conclusion that someone else did;

but when asked to suppose what the world would be like had Oswald not killed Kennedy will not suppose that someone else would have. The difference is often marked in ordinary language by the distinction between indicative and subjunctive mood. (Skyrms, 1987)

While the use of Bayes' rule, alongside the product and marginalization rules of inference, tell us how to update probabilities given evidence, they do not tell us how to assign prior probabilities or specify the likelihood. The distinction is not one of methodological opposition. In practice, MaxEnt supplies the distributional assumptions and Bayes conducts the inference. Or rather, MaxEnt provides a systematic interface for plugging causal, modal and factual content into the inductive scheme. In the absence of such content, the MaxEnt framework will yield uniform priors and likelihoods; it is the information-theoretic version of the 'principle of insufficient reason' (Hacking, 2006; Jaynes, 1957; Keynes, 1909; Stigler, 1986).

### 5.5.2 Statistical Autonomy

A subset of non-constructive statistical explanations have been described in the philosophical literature as being 'statistically autonomous' (Ariew et al., 2015) or 'really statistical explanations' (Lange, 2013). What characterizes the statistical autonomy of NCSEs in the account I give here is a subsidiary relationship to the existing stock of probability models, and the explanation of specific events or regularities in terms of 'statistical fallout', or as the consequence of a sampling process. I may state that  $y$  is normally distributed, or that it follows a power law, without citing the counterfactual properties that the data generating process must meet in order to lead to that specific distribution. To give an example: Why in a class of students do those with the lowest scores in the first term tend not to be the students with the lowest scores in the second term? A possible explanation could be that those who performed poorly in the first term dropped out. Or that a new substitute teacher came in and was able to motivate them to work better (Lange, 2013). But without alluding to such counterfactuals, we may simply explain this phenomenon by pointing to a mean regression process in the population of students. Thus I may think that grades or ages in a population of students are normally distributed and allude to the properties of the normal distribution in order to explain deviations from a presumably population-invariant mean.

Galton famously used the normal curve to explain the mean regression tendencies of hereditary traits (Hacking, 2006). He saw the properties of the normal curve as giving him the tools necessary to solve the puzzle of intergenerational heredity. Central to Galton's distinctively statistical explanation was the idea that the normal curve would hold generation after generation, i. e. that it was invariant. That would explain why exceptional members would not typically have exceptional descendants. But the principle of natural selection is presumably the causal factor that acts to shape the probability of genetic traits in a population.

Statistically autonomous explanations thus provide one side of the story that can be told about the explanandum, and can be supplemented or put into relation with theories that seek to account for deeper causal/modal structure. In practice, such statistically autonomous descriptions may constitute a necessary first step in the long process of elaborating a theory's probabilistic content. The structure of the data generating process may end up being fully amenable to mechanistic treatment or may end up being irreducibly probabilistic, as in the case of quantum theories (Bohm, 2004). For some, the 'irreducibly probabilistic' aspect of a theory may be a sort of temporary status that subsists up to the point where the discovery of deeper and deeper causal/modal principles peel off the layers of uncertainty and contingency. As Hacking (2006) points out, 'some authors find the question of reducibility of deep importance, partly because it is a way of preserving deterministic metaphysics'. Hacking cites Poisson's law of large numbers, Boltzmann's H-theorem and much of ergodic theory as examples. More contemporary examples are the Many-Worlds Interpretations (MWI) of quantum mechanics (Carroll, 2021; Everett, 2012), as well as David Deutsch's and Chiara Marletto's Constructor Theory (Deutsch, 2013; Marletto, 2021).

Constructive probabilistic descriptions and explanations are phenomenologically important because they force us to think about the deep nature of a process and to come up with ways of understanding the conditions that give rise to their distinctive statistical regularities. This concern goes beyond asserting long-term frequencies, counting, updating or curve-fitting; it has deeper theoretical implications.

We see, then, that it is appropriate to speak about objectively valid laws of chance, which tell us about a side of nature that is not treated completely by the causal laws alone. Indeed, the laws of chance are

just as necessary as the causal laws themselves. For example, the random character of chance fluctuations is, in a wide variety of situations, made inevitable by the extremely complex and manifold character of the external contingencies on which fluctuations depend. . . The random character of these fluctuations is quite often an inherent and indispensable part of the normal functioning of many kinds of things, and of their modes of being. Thus, it would be impossible for a modern city to continue to exist in its normal condition unless there were a tendency towards the cancellation of chance fluctuations in traffic, in the demand for various kinds of food, clothing, etc., in the times at which various individuals fall sick or die. In all kinds of fields we find a dependence on the characteristic effects of chance. (Bohm, 2004)

## 5.6 Constructive Statistical Explanations in Political Economy

There is a growing body of research in analytical political economy applying statistical equilibrium and maximum entropy methods for the analysis of economic distributions. See Yang (2018) and Scharfenaker (2020a) for thorough reviews. Despite the focus on physics related examples in the previous sections, my main motivation behind the development of the constructive explanatory schema is to make a methodological contribution to this literature, and to explore some of the philosophical implications of this line of thinking for political and economic analysis. In this section I will focus on a single case that is relevant for the account of statistical explanation provided in this paper; the problem of adducing a statistical explanation for the empirical distribution of firm profit rates, which have been shown to be well approximated by Laplace, Subbotin and QRSE distributions (Scharfenaker and Foley, 2017). In this example I will discuss how three different attempts at explanation relate to the constructive explanatory schema that I have outlined throughout the previous sections.

### 5.6.1 Explaining Profit Rates

Suppose, for the sake of argument, that we have dataset of profit rates for some well defined ensemble of firms over a certain period, as in figure 5.7. This data is simulated by sampling a laplace distribution with mean and scale parameters equal to 0 and 1 respectively.

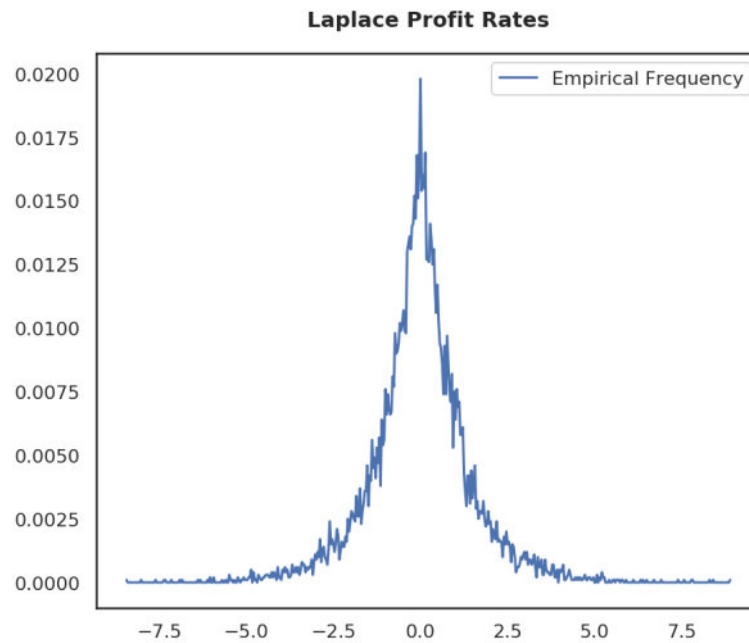


Fig. 5.7 Laplace Profit Rates

Now, suppose we observe this empirical distribution of profit rates but don't know the sampling distribution, and that we wish to explain why it is distributed in the way that it is. Thus, the histogram in figure 5.7 is the explanandum. I consider three different theories that can be proposed to explain this kind of data, which is distinctively peaked and shows sharp exponential decay around the mode. These theories appear in the recent political economy literature; Alfarano and Milaković (2008), Alfarano et al. (2012a) and Scharfenaker and Foley (2017). For the QRSE theory, I use an updated parametrization which appears in Scharfenaker and Foley (2021b). The three theories considered are based on Smith's classical theory of competition, but offer different statistical explanations.



In the next subsection I briefly restate the maximum entropy formalism in more general terms and introduce some shorthand notation, as it will help to clarify the discussion, given the theories' reliance on this approach.

### General Maximum Entropy Formalism

A predictive statistical equilibrium distribution  $\hat{f}_x$  is obtained by maximizing the entropy of a random variable  $X$  with density function  $f_x$ , subject to a set of constraints which encode empirical and theoretical information. A constraint enters the program as an expectation (over  $X$ ) of a function  $g_c(x) \in \mathcal{C}$ . The set  $\mathcal{C}$  is the set of constraints that relate the underlying theory  $T$ , or empirical information, to the predictive distribution  $\hat{f}_x$ . If there are  $m$  constraints, then  $|\mathcal{C}| = m$ . The program takes the form:

$$\begin{aligned} \max_{f_x \geq 0} \quad & H_x = - \int_{\mathcal{X}} f_x \log[f_x] dx \\ \text{st.} \quad & \\ & \int_{\mathcal{X}} f_x dx = 1 \\ & \int_{\mathcal{X}} f_x g_c(x) dx = F_c, \quad c = 1, \dots, m \end{aligned} \tag{5.11}$$

For standard references on how this program is solved via the method of Lagrange multipliers see (Borwein and Lewis, 1991; Golan, 2018; Sivia and Skilling, 2006). The solution yields a maximum entropy density  $\hat{f}_x$  with a vector  $\vec{\lambda} = [\lambda_1 \dots \lambda_m]$  of Lagrange multipliers as parameters. As pointed out by Scharfenaker (2020a), the Lagrange multipliers associated with each constraint can be understood as ‘the marginal amount of information a constraint contributes to the reduction of the entropy of the statistical equilibrium distribution’. The general form for the solution, and the partition function  $Z$ , are shown below in equations 5.12 and 5.13.

$$\hat{f}_x = \frac{1}{Z} e^{\lambda_1 g_1(x) + \dots + \lambda_m g_m(x)} \tag{5.12}$$

$$Z = \int_{\mathcal{X}} \exp \left[ \lambda_1 g_1(x) + \dots + \lambda_m g_m(x) \right] dx \tag{5.13}$$

Note that  $E[g_c(x)] = F_c$ . The constraint expectation values can be set to be both equality or inequality constraints. In the latter case, then  $F_c \leq \delta$ , for some  $\delta$ . If the support of  $X$  is discrete, the entropy and expectations are defined as sums over the state space  $\mathcal{X}$ .

### Shorthand notation

In order to avoid repeating the formulation of the program for different theories, I will state that a theory  $T$  ‘maximum entropy derives’ the predictive distribution  $\hat{f}_x$  by the following shorthand notation:  $\mathcal{C}_T \models_{\text{ME}} \hat{f}_x$ . When stated in this manner it is implicit that the entropy of  $X$  is maximized subject to the normalization condition,  $\int f_x dx = 1$ , and that  $f_x \geq 0$ .

If  $\mathcal{C} = \{\emptyset\}$ , then  $\hat{f}_x$  is a uniform distribution. if  $\mathcal{C} = \bar{x}$ , then it is an exponential distribution with a single rate parameter  $\lambda_1$ ;  $\hat{f}_x \propto e^{\lambda_1 x}$ . When  $\mathcal{C} = \{(x - \mu)^2\}$ , then  $\hat{f}_x \propto e^{-\lambda_1(x-\mu)^2}$ . In this latter case, if  $\mathcal{X}$  is the real line and  $F_c = \sigma^2$ , then  $\lambda_1 = 1/2\sigma^2$ ; which recovers the standard Gaussian density.

### Three Explanations

What are the counterfactual properties that the competitive process must possess in order for profit rate data to be Laplace distributed ? In the context of the classical theory of competition, what set of constraints yields a constructive statistical explanation for the data ? What type of set yields a statistically autonomous one ?

In previous sections, I defined a CSE as supplying either causal and/or modal content to the predictive distribution that it uses to explain the observed frequencies. A good explanation should contain a minimal set  $\mathcal{C}$  that is both hard to vary, and difficult to reduce further when considering some established knowledge about the process involved (Deutsch, 2011). In the derivation of the Boltzmann-Gibbs distribution for the energies of an isothermal gas, assuming time-reversal symmetry was too strong a requirement for obtaining a stationary distribution, and energy conservation alone was sufficient to yield the required family of distributions via the maximum entropy program. In terms of counterfactuals, I pointed out that this implies that the set of processes that are energy conserving is larger and contains the set of processes that are mechanically reversible. When considering the statistical

explanations that can be adduced to explain profit rate data, similar distinctions and issues crop up.

I first consider the two alternatives provided by Alfarano and Milaković (2008) and Alfarano et al. (2012a) to explain US firm profit rate data for the period 1980-2006. Similar to Scharfenaker and Foley (2017), Alfarano and Milaković (2008) understand the general competitive economic process as a type of congestion or negative feedback mechanism.

In Smith's classical theory of competition, the persistent entry and exit of firms into distinct sectors pursuing profitable returns works to stabilize the profit rate around a central tendency. Consequently, the two starting points for building a statistical explanation based on the theory of classical competition are i) the hypothesis of profit rate equalization, and ii) the postulation of a negative feedback or congestion mechanism that can account for the dispersion of rates around a central tendency. The first condition implies that  $\hat{f}_x$  should be a unimodal and relatively symmetric distribution. If the empirical frequencies are extremely skewed, with heavy fat left or right tails, or show more than one mode; this would be evidence against a profit rate equalizing process. The data in figure 5.7 is not skewed, so this is not a concern. But if it were, we would need to account for the 'economic forces', other than the equalizing ones, which are pushing the rates far beyond or below the mode. The second condition, which requires the postulate of a competitive feedback mechanism, is harder to translate into statistical terms. It requires a description of the competitive process which is capable of generating, through endogenous fluctuations, a laplacian law of error for the dispersion of rates around the average. Unless one is only interested in averages, this immediately rules out normality for  $\hat{f}_x$ .

It is from this point of view, which requires a unimodal symmetric distribution with a laplacian dispersion law, that the authors in Alfarano and Milaković (2008) set up a maximum entropy program with the constraint set  $\mathcal{C}_1 = \left\{ \left| \frac{x-\mu}{\sigma} \right|^\alpha \right\}$ , and a constraint expectation value  $F_1 = 1$ . The solution to this program is a symmetric Subbotin distribution with a location parameter  $\mu$ , a scale parameter  $\sigma$ , and a shape parameter  $\alpha$ . The density is shown below in figure 5.8. For  $\alpha = 1$  and  $\alpha = 2$  the distribution reduces respectively to the Laplace and Normal densities. And for  $\alpha \rightarrow \infty$ ,  $\alpha \rightarrow 0$ , the Subbotin tends to Uniform and Delta distributions. The Laplace, Subbotin and Normal fits to the simulated data are shown in figure 5.9. I will call this Theory 1 ( $T_1$ ).

<p><b>Laplace</b></p> $\hat{f}_{x \vec{\lambda}} = \frac{1}{2\sigma} e^{-\frac{ x-\mu }{\sigma}}$ <hr/> <p><b>Normal</b></p> $\hat{f}_{x \vec{\lambda}} = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$	<p><b>Subbotin</b></p> $\hat{f}_{x \vec{\lambda}} = \frac{1}{2\sigma\Gamma(1 + 1/\alpha)} e^{-\frac{1}{\alpha} \left  \frac{x-\mu}{\sigma} \right ^\alpha}$ <hr/> <p><b>QRSE</b></p> $\hat{f}_{x \vec{\lambda}} = \frac{1}{Z} e^{H_a x} e^{-\tanh\left(\frac{x-\mu}{T}\right) \left(\frac{x-\mu}{S}\right)}$
---	--

Fig. 5.8 Densities Chart

The statistical explanations that can be built based on  $T_1$  are clearly statistically autonomous; the parameters do not supply direct information about the causal or modal principles which may be regulating the competitive processes in terms of the classical theory. But  $T_1$  has the great virtue of recovering the laplacian dispersion law with great accuracy, and is flexible enough to account for a wide range of processes, including the fully deterministic case when  $\alpha = 0$ . Since the Subbotin recovers the Laplace,  $T_1$  opens up a discussion about the stability of  $\alpha \approx 1$  across different samples and microeconomic environments. As the authors themselves note, ‘ ‘this prompts us to ask why the empirical shape parameter is close to unity, what this implies about the competitive environment that firms are facing, and whether variations in the shape parameter correspond to qualitative changes in the competitive environment” (Alfarano and Milaković, 2008). Thus, this theory opens up the range of microeconomic mechanisms which can be adduced as candidates for explanation, and which are compatible with the classical theory of competition <sup>3</sup>.

In an effort to account for the complex firm competition dynamics which generate Laplace distributed profit rates, the authors extend  $T_1$  with a second theory  $T_2$  ((Alfarano et al., 2012a)) that uses a stochastic drift-diffusion model, and which

<sup>3</sup>The firm profit rates of  $T_1$  are defined as the ratio of income flows to capital stock

### Explaining Profit Rates

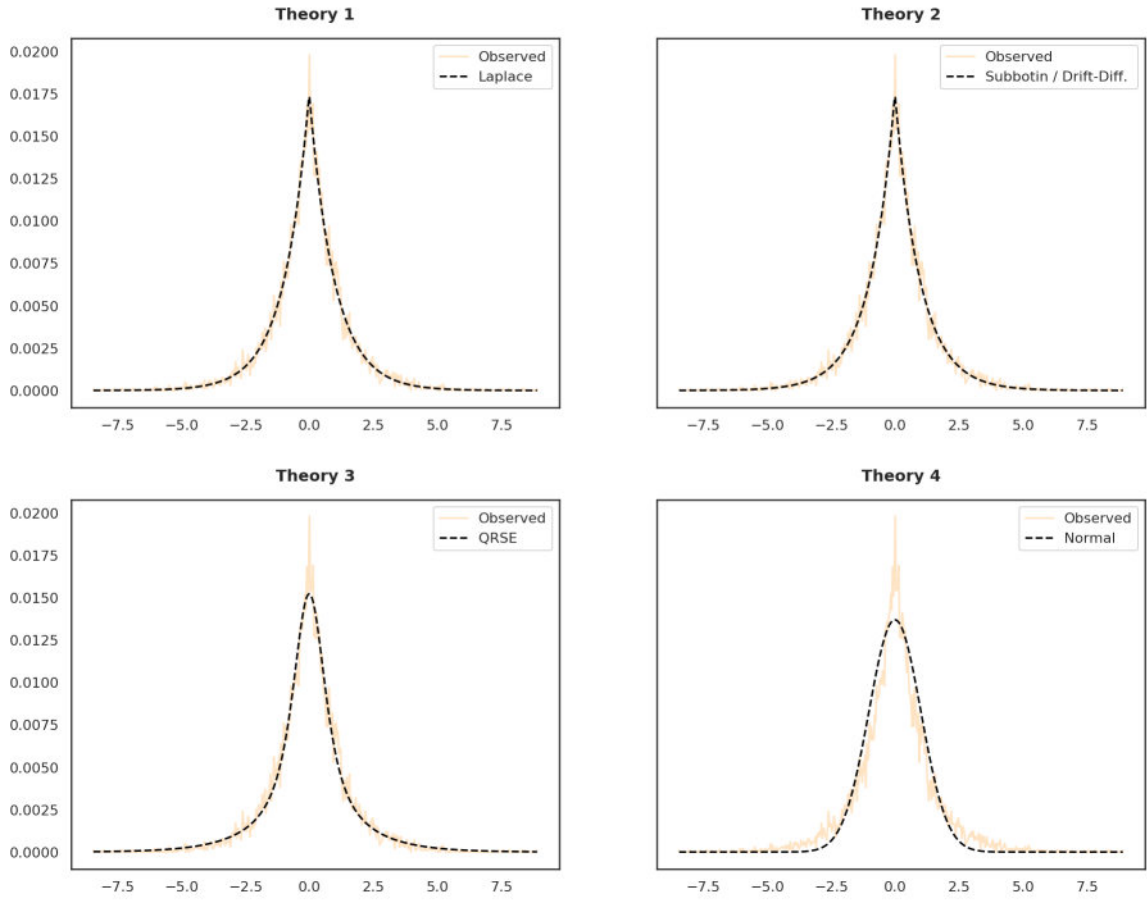


Fig. 5.9 Fits for the different theories proposed.

generates the Subbotin as the stationary distribution of the process. They propose a time-homogenous diffusion on the real line of the form:

$$dX_t = \frac{D}{2\sigma} \text{sign}(X_t - \mu) \left| \frac{X_t - \mu}{\sigma} \right|^{\alpha-1} dt + \sqrt{D} dW_t \quad (5.14)$$

It is hard to reconcile this ergodic diffusion model with the classical theory of competition, as outlined above, other than as a sampling construct for Subbotin distributed data which decomposes competitive pressures into a mean-reverting term and ‘idiosyncratic’ forces into a random dispersion term.

Good explanations are resilient to counterfactual variation, and following Occam’s razor, should be synthesized by a minimal set of defeasible (non-monotonic)

propositions. In the case of  $T_2$ , this is easy to state by simply considering the competitive process as taking place over networks with time-varying topologies and where the ordering of firm entry/exit decisions has some specified causal effect on the statistical equilibrium distribution of profit rates. As an analogy, let alone as a causal or non-causal description, the diffusion process of  $T_2$  is hard to swallow. The time-invariance and Markov properties of the  $T_2$  diffusion process may be easily swapped for others, while retaining the requirement for the statistical equilibrium distribution to be Laplace. Note that this point is not about the falsifiability of the theory, but about the minimal set of principles in a constraint set  $\mathcal{C}_T$  that can survive a loaded round of counterfactual variation, while retaining the explanandum (i.e. the shape of the data) fixed.  $T_1$  has the virtue that it opens a wide but statistically constrained space for advancing microeconomic descriptions, while  $T_2$  drastically narrows it.

The third theory I consider ( $T_3$ ), is the Quantal Response Statistical Equilibrium (QRSE) model developed in Scharfenaker and Foley (2017) and Scharfenaker and Foley (2021b). The QRSE theory advances a joint probability model  $f_{x,a}$  for the unobserved entry and exit decisions of firms into and out of competitive sectors, and the marginal profit rate  $x$ . This theory sets up a maximum entropy program that maximizes the joint entropy of the unobserved action and the profit rate variables, subject the constraint set  $\mathcal{C}_3 = \{\Delta f_{a|x}(x - \alpha)\}$ , with a constraint expectation inequality  $F_1 \leq \delta$ . In this theory, the function  $f_{a|x}$  is based on a model of bounded rational choice, and is defined as the mixed strategy of entry and exit of firms into/out of competitive sectors, conditional on the profit rate  $x$ . Formally, for a discrete set of actions  $a \in \mathcal{A}$ , and a payoff function  $\pi(a, x)$ , the boundedly rational mixed strategy  $f_{a|x}$  is obtained by maximizing its entropy subject to minimum payoff value. In shorthand notation:  $\{\pi(a, x)\} \models_{\text{ME}} f_{a|x}$  and  $F_1 \geq V_{\min}$ .

In summary, the QRSE theory involves two entropy maximization programs, one operating at the microeconomic level and the other at the macro or market-level:

- Micro Program:  $\{\pi(a, x)\} \models_{\text{ME}} f_{a|x}, F_1 \geq V_{\min}$
- Macro Program:  $\{\Delta f_{a|x}(x - \alpha)\} \models_{\text{ME}} \hat{f}_x, F_1 \leq \delta$

Since  $f_{a|x} f_x = f_{a,x}$ , the two programs completely specify a joint probability model. If the action set is binary, and  $\pi(a_1, x) = -\pi(a_2, x)$ , with  $a_1 = \text{entry}$  and

$a_2 = \text{exit}$ , this yields a familiar Gibbs or logit quantal response function with a scale parameter  $T$ :

$$f_{a_1|x} = 1 - f_{a_2|x} = \frac{1}{1 + e^{-\frac{2(x-\mu)}{T}}} \quad (5.15)$$

In this case, with two actions and symmetric payoffs,  $\Delta f_{a|x} = \tanh\left(\frac{x-\mu}{T}\right)$  and the macro constraint set becomes  $\mathcal{C}_3 = \{\tanh\left(\frac{x-\mu}{T}\right)(x-\alpha)\}$ . The QRSE predictive density  $\hat{f}_x$  and fit to the simulated Laplace data are shown above in figures 5.8 and 5.9. The QRSE parameter vector contains two scale parameters  $T$  and  $S$ , and two location parameters  $\mu$  and  $\alpha$ , which respectively operate at the micro and macro levels;  $\vec{\lambda} = [T, S, \mu, \alpha]$ .  $S$  is the inverse of the Lagrange multiplier  $\lambda_1$  that comes out of the solution to the macro ME program, which allows for expressing both scale parameters in the same units as  $x$ ,  $\mu$  and  $\alpha$ .

The QRSE translation of the classical theory of competition into stochastic terms is parsimoniously condensed in the single constraint set  $\mathcal{C}_3$ , which Scharfenaker and Foley (2017) define as a competitive feedback constraint. The expectation of this constraint,  $F_1 \leq \delta$ , expresses the idea that the expected value of a market determined rate  $\alpha$ , weighted by the difference of entry/exit probabilities, is constrained to be some positive but finite value. The constraint  $\Delta f_{a|x}(x-\alpha)$  codifies the postulate that differences in entry/exit probabilities of individual firms have a non-zero impact on the statistical dispersion of the profit rate around the average. For very low values of the behavioral scale parameter  $T$ , the QRSE density recovers Laplace shaped data, and for high values it recovers Gaussian data, as long as the impact of the overall competitive process which is captured by the market scale parameter  $S$  is non-negligible. In the simulated fit in figure 5.9,  $\hat{T} = 0.9$  and  $\hat{S} = 1.4$ . If we set  $T = 6$  and  $S = 0.3$ , for example, the QRSE density recovers a well shaped symmetric Gaussian around  $\mu = \alpha = 0$ , equivalent to  $T_4$  in the graph.

The mutual impact between the market level dispersion of the profit rate around  $\alpha$ , and the dispersion of firm level expectations around  $\mu$ , is advanced by Scharfenaker and Foley (2017) as a causal principle in this theory. The effects of the competitive feedback principle are directly traceable to estimated values of the model parameters  $T$ ,  $\mu$ , and  $S$ . The parameter  $\alpha$ , on the other hands, captures a residual statistical autonomy in the model. Differences in entry/exit probabilities work to scale and

push the market rate  $\alpha$ , but these are only one of the many potential forces that may be acting on the profit rate in a complex and high-dimensional competitive economic environment.

## 5.7 Conclusion

Having set all the details down for discussion of  $T_3$ , I conclude by considering the following questions: To what range of counterfactual variation is the principle of competitive feedback resilient? Is the principle of competitive feedback causal or non-causal ? Why is  $T_3$  a constructive statistical explanation ?

With regards to the second question,  $T_3$  is constructive on two grounds. First, the behavioral and market scale parameters directly trace the impact of the competitive feedback process on equalizing profit rates, and hence on concentrating values around the mode. The relative impact of the two scales of the economic process is not decomposable in an additive manner, due to the mutual interdependence that is established by the negative feedback mechanism. The principle of competitive feedback implies a form of probabilistic causation, but it is not a mechanical specification, and does not offer a time-dependent stochastic dynamical law that could account for the patterns of observed profit rates. And because the theory offers a joint probability model of unobserved actions and observed outcomes, it can show via simulation why and how a system constrained by the competitive feedback principle can approximate a wide range of observed distributions, from Laplace to Normal.

Now let me address the first question. The generality that is afforded by the probabilistic principle of competitive feedback, parsimoniously packed into the constraint in  $\mathcal{C}_3$ , can be applied to a very wide set of samples and reference classes. The principle is in fact so general, that some might retort that it is non-causal, in a manner similar to conservation laws in physics. If by non-causal we mean the sort of high-level principles that constraint local nomological regularities, I believe there is a valid case to be made for such an account with respect to the QRSE constructive explanation. The answer depends on what is deemed causal and non-causal in complex economic environments, which is not an easy question to address. But let me try to give an argument for why it can be deemed non-causal. In order to take it as a non-causal principle one really has to suppose the general form  $\Delta f_{a|x}(x - \alpha)$ ,



independently of any detailed microeconomic specification. Taken in its most general form, the principle may be put to test with actual data to see how well it ranges over different decentralized competitive environments. For example, does it apply to environments with different temporal orderings for entry/exit dynamics ? How does it apply to decentralized markets that exhibit different fiscal and financial liquidity conditions ? The recent QRSE literature, and my own work using it to study inter-jurisdictional competition in the local US education market, have shown that its range is quite wide. The question of its causal/modal interpretation ultimately hinges on how far a particular use case is willing to go in specifying the microeconomic potential  $\Delta f_{a|x}$ .

# References

- Aczél, J. and Kannappan, P. (1978). A mixed theory of information. iii. inset entropies of degree  $\beta$ . *Information and Control*, 39(3):315–322.
- Albin, P. (1998). *Barriers and bounds to rationality: Essays on economic complexity and dynamics in interactive systems*, volume 40. Princeton University Press.
- Albin, P. and Foley, D. (1998). *Barriers and bounds to rationality: Essays on economic complexity and dynamics in interactive systems*, volume 40. Princeton University Press.
- Albin, P. S. (1975). *The analysis of complex socioeconomic systems*. Lexington Books.
- Alfarano, S. and Milaković, M. (2008). Does classical competition explain the statistical features of firm growth? *Economics Letters*, 101(3):272–274.
- Alfarano, S., Milaković, M., Irle, A., and Kauschke, J. (2012a). A statistical equilibrium model of competitive firms. *Journal of Economic Dynamics and Control*, 36(1):136–149.
- Alfarano, S., Milaković, M., Irle, A., and Kauschke, J. (2012b). A statistical equilibrium model of competitive firms. *Journal of Economic Dynamics and Control*, 36(1):136–149. Publisher: Elsevier.
- Amari, S.-i. (1985). Differential-geometrical methods in statistics. *Lecture Notes on Statistics*, 28:1.
- Ames, E. (1983). Automation and group structures in certain economic adjustment mechanisms. *Mathematical Social Sciences*, 6(2):247–260.
- Anderson, P. W. (1988). *The economy as an evolving complex system*. CRC Press.
- Anderson, T. W. and Rubin, H. (1949). Estimation of the parameters of a single equation in a complete system of stochastic equations. *The Annals of Mathematical Statistics*, 20(1):46–63.
- Aoki, M. and Yoshikawa, H. (2011). *Reconstructing macroeconomics: a perspective from statistical physics and combinatorial stochastic processes*. Cambridge University Press.

- Ariew, A., Rice, C., and Rohwer, Y. (2015). Autonomous-statistical explanations and natural selection. *British Journal for the Philosophy of Science*, 66(3):635–658.
- Arthur, W. B. (1997). *The economy as an evolving complex system II*. CRC Press.
- Bahri, Y., Kadmon, J., Pennington, J., Schoenholz, S. S., Sohl-Dickstein, J., and Ganguli, S. (2020). Statistical mechanics of deep learning. *Annual Review of Condensed Matter Physics*, 11:501–528.
- Baker, B. D. (2021). *Educational inequality and school finance: Why money matters for America's students*. Harvard Education Press.
- Bamford, G. (1996). Popper and his commentators on the discovery of neptune: A close shave for the law of gravitation? *Studies in History and Philosophy of Science Part A*, 27(2):207–232.
- Barabási, A.-L. and Albert, R. (1999). Emergence of scaling in random networks. *science*, 286(5439):509–512.
- Barde, S. (2012). Of ants and voters. *Revue de l'OFCE*, (5):147–175.
- Barde, S. (2015). Back to the future: economic self-organisation and maximum entropy prediction. *Computational Economics*, 45(2):337–358.
- Beck, C. (2009). Generalised information and entropy measures in physics. *Contemporary Physics*, 50(4):495–510.
- Beck, N., Gleditsch, K. S., and Beardsley, K. (2006). Space Is More than Geography: Using Spatial Econometrics in the Study of Political Economy. *International Studies Quarterly*, 50(1):27–44.
- Beneš, V. E. (1965). *Mathematical theory of connecting networks and telephone traffic*. Academic press.
- Bialek, W. (2017). Perspectives on theory at the interface of physics and biology. *Reports on Progress in Physics*, 81(1):012601.
- Bialek, W., Cavagna, A., Giardina, I., Mora, T., Silvestri, E., Viale, M., and Walczak, A. M. (2012). Statistical mechanics for natural flocks of birds. *Proceedings of the National Academy of Sciences*, 109(13):4786–4791.
- Bikhchandani, S., Hirshleifer, D., Tamuz, O., and Welch, I. (2021). Information cascades and social learning. Technical report, National Bureau of Economic Research.
- Bohm, D. (2004). *Causality and chance in modern physics*. Routledge.
- Borwein, J. M. and Lewis, A. S. (1991). Duality relationships for entropy-like minimization problems. *SIAM Journal on Control and Optimization*, 29(2):325–338.

- Bottazzi, G. and Secchi, A. (2005). On the laplace distribution of firm growth rates. Technical report, LEM Working Paper Series.
- Bowles, S. (2009). *Microeconomics*. Princeton University Press.
- Brennan, G. and Buchanan, J. M. (1980). *The power to tax: Analytic foundations of a fiscal constitution*. Cambridge University Press.
- Bridle, J. S. (1990). Probabilistic interpretation of feedforward classification network outputs, with relationships to statistical pattern recognition. In *Neurocomputing*, pages 227–236. Springer.
- Bryson, M. C. (1974). Heavy-tailed distributions: properties and tests. *Technometrics*, 16(1):61–68.
- Carroll, S. (2021). Keynote: The many worlds of quantum mechanics.
- Castellano, C., Fortunato, S., and Loreto, V. (2009). Statistical physics of social dynamics. *Reviews of modern physics*, 81(2):591.
- Cavagna, A., Giardina, I., Orlandi, A., Parisi, G., Procaccini, A., Viale, M., and Zdravkovic, V. (2008). The starflag handbook on collective animal behaviour: Part i, empirical methods. *arXiv preprint arXiv:0802.1668*.
- Chakrabarti, B. K., Chakraborti, A., and Chatterjee, A. (2006). *Econophysics and sociophysics: trends and perspectives*. John Wiley & Sons.
- Chomsky, N. (1993). *Lectures on government and binding: The Pisa lectures*. Number 9. Walter de Gruyter.
- Cover, T. M. and Thomas, J. A. (2006). Elements of information theory second edition solutions to problems. *Internet Access*, pages 19–20.
- Cowan Jr, C. L., Reines, F., Harrison, F., Kruse, H., and McGuire, A. (1956). Detection of the free neutrino: a confirmation. *Science*, 124(3212):103–104.
- De Domenico, M., Brockmann, D., Camargo, C., Gershenson, C., Goldsmith, D., Jeschonnek, S., Kay, L., Nichele, S., Nicolás, J., Schmickl, T., et al. (2019). Complexity explained.
- De Finetti, B. (1974). *Theory of probability: A critical introductory treatment*, volume 6. John Wiley & Sons.
- Deutsch, D. (1985). Quantum theory, the church–turing principle and the universal quantum computer. *Proceedings of the Royal Society of London. A. Mathematical and Physical Sciences*, 400(1818):97–117.
- Deutsch, D. (2011). The beginning of infinity. *Allen Lane*.
- Deutsch, D. (2013). Constructor theory. *Synthese*, 190(18):4331–4359.

- Deutsch, D., Ekert, A., and Lupacchini, R. (2000). Machines, logic and quantum physics. *Bulletin of Symbolic Logic*, 6(3):265–283.
- Diaconis, P. and Freedman, D. (1980). Finite exchangeable sequences. *The Annals of Probability*, pages 745–764.
- Dias, P. M. C. and Shimony, A. (1981). A critique of jaynes’ maximum entropy principle. *Advances in Applied Mathematics*, 2(2):172–211.
- Dimitrov, V. I. (2007). On shannon-jaynes entropy and fisher information. In *AIP Conference Proceedings*, volume 954, pages 143–152. American Institute of Physics.
- Dowding, K., John, P., and Biggs, S. (1994). Tiebout: A survey of the empirical literature. *Urban studies*, 31(4-5):767–797. Publisher: Sage Publications Sage UK: London, England.
- Dragulescu, A. and Yakovenko, V. M. (2000a). Statistical mechanics of money. *The European Physical Journal B-Condensed Matter and Complex Systems*, 17(4):723–729.
- Dragulescu, A. and Yakovenko, V. M. (2000b). Statistical mechanics of money. *The European Physical Journal B-Condensed Matter and Complex Systems*, 17(4):723–729. Publisher: Springer.
- Eaves, B. C. (1971). Computing kakutani fixed points. *SIAM Journal on Applied Mathematics*, 21(2):236–244.
- Edel, M. and Sclar, E. (1974). Taxes, spending, and property values: Supply adjustment in a Tiebout-Oates model. *Journal of political Economy*, 82(5):941–954. Publisher: The University of Chicago Press.
- Ehrenfest, P. (1917). In what way does it become manifest in the fundamental laws of physics that space has three dimensions. In *Proc. Amsterdam Acad*, volume 20, pages 200–200.
- Eigen, M. (2013). *From strange simplicity to complex familiarity: a treatise on matter, information, life and thought*. OUP Oxford.
- Einstein, A. (1907). 252 the relativity principle. *Jahrbuch der Radioaktivität und Elektronik*, 4:411–462.
- Epple, D. and Nechyba, T. (2004a). Fiscal decentralization. In *Handbook of regional and urban economics*, volume 4, pages 2423–2480. Elsevier.
- Epple, D. and Nechyba, T. (2004b). Fiscal decentralization. In *Handbook of regional and urban economics*, volume 4, pages 2423–2480. Elsevier.
- Epple, D., Peress, M., and Sieg, H. (2010). Identification and Semiparametric Estimation of Equilibrium Models of Local Jurisdictions. *American Economic Journal: Microeconomics*, 2(4):195–220.

- Epple, D., Romer, T., and Sieg, H. (1999). The Tiebout Hypothesis and Majority Rule: An Empirical Analysis. Technical Report w6977, National Bureau of Economic Research, Cambridge, MA.
- Epple, D., Romer, T., and Sieg, H. (2001). Interjurisdictional Sorting and Majority Rule: An Empirical Analysis. *Econometrica*, 69(6):1437–1465.
- Epple, D., Zelenitz, A., and Visscher, M. (1978). A Search for Testable Implications of the Tiebout Hypothesis. *Journal of Political Economy*, 86(3):405–425.
- Everett, H. (2012). *The Everett interpretation of quantum mechanics: Collected works 1955-1980 with commentary*. Princeton University Press.
- Farjoun, E. and Machover, M. (1983). *Laws of chaos*. Verso.
- Feynman, R. (1965). The feynman lectures on physics; vol. i. *American Journal of Physics*, 33(9):750–752.
- Feynman, R. (2017). *The character of physical law*. MIT press.
- Fischer, H. (2010). *A history of the central limit theorem: from classical to modern probability theory*. Springer Science & Business Media.
- Floridi, L. (2005). Semantic conceptions of information.
- Foley, D. K. (2017). Crisis and theoretical methods: equilibrium and disequilibrium once again. Technical report.
- Foley, D. K. (2020). Information theory and behavior. *The European Physical Journal Special Topics*, 229(9):1591–1602.
- Friedman, K. and Shimony, A. (1971). Jaynes’s maximum entropy prescription and probability theory. *Journal of Statistical Physics*, 3(4):381–384.
- Friston, K. (2009). The free-energy principle: a rough guide to the brain? *Trends in cognitive sciences*, 13(7):293–301.
- Gallegati, L., Di Guilmi, Mauro, S., and Corrado (2008). Modeling maximum entropy and mean-field interaction in macroeconomics. *Economics Discussion Paper*, (2008-36).
- Gelman, A. and Imbens, G. (2013). Why ask why? forward causal inference and reverse causal questions. Technical report, National Bureau of Economic Research.
- Gelman, A. and Shalizi, C. R. (2013). Philosophy and the practice of bayesian statistics. *British Journal of Mathematical and Statistical Psychology*, 66(1):8–38.
- Gentner, D. and Gentner, D. R. (2014). Flowing waters or teeming crowds: Mental models of electricity. In *Mental models*, pages 107–138. Psychology Press.

- Gentner, D. and Smith, L. (2012). Analogical reasoning. *Encyclopedia of human behavior*, 2:130–136.
- Geraci, M. and Borja, M. C. (2018). Notebook: The laplace distribution.
- Giffin, A. and Caticha, A. (2007). Updating probabilities with data and moments. In *AIP conference proceedings*, volume 954, pages 74–84. American Institute of Physics.
- Golan, A. (2018). *Foundations of info-metrics: Modeling, inference, and imperfect information*. Oxford University Press.
- Goodman, C. B. (2019). Local government fragmentation: What do we know? *State and Local Government Review*, 51(2):134–144.
- Gottwald, S. and Braun, D. A. (2020). The two kinds of free energy and the bayesian revolution. *PLoS computational biology*, 16(12):e1008420.
- Gull, S. F. and Skilling, J. (1984). Maximum entropy method in image processing. In *IEE Proceedings F-Communications, Radar and Signal Processing*, volume 131, pages 646–659. IET.
- Hacking, I. (2006). *The emergence of probability: A philosophical study of early ideas about probability, induction and statistical inference*. Cambridge University Press.
- Hayek, F. A. et al. (1948). The meaning of competition.
- Hempel, C. G. et al. (1965). *Aspects of scientific explanation*, volume 3. Free Press New York.
- Hertert, L., Busch, C., and Odden, A. (1994). School financing inequities among the states: The problem from a national perspective. *Journal of Education Finance*, pages 231–255. Publisher: JSTOR.
- Hirsch, M. W., Marsden, J. E., and Shub, M. (2012). *From topology to computation: proceedings of the Smalefest*. Springer Science & Business Media.
- Hogg, D. W. and Foreman-Mackey, D. (2018). Data analysis recipes: Using markov chain monte carlo. *The Astrophysical Journal Supplement Series*, 236(1):11.
- Hood, W. C., Koopmans, T. C., et al. (1953). *Studies in econometric method*.
- Hoover, K. D. (2005). The methodology of econometrics. *Available at SSRN* 728683.
- Hoover, K. D. (2008). Causality in economics and econometrics. *The new Palgrave dictionary of economics*, 2.
- Imbens, G. W. and Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of economic literature*, 47(1):5–86.

- Ising, E. (1925). Beitrag zur theorie des ferromagnetismus. *Zeitschrift für Physik*, 31(1):253–258.
- Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical review*, 106(4):620. Publisher: APS.
- Jaynes, E. T. (1968). Prior probabilities. *IEEE Transactions on systems science and cybernetics*, 4(3):227–241.
- Jaynes, E. T. (1983). Concentration of distributions at entropy maxima. *ET Jaynes: Papers on probability, statistics, and statistical physics*, pages 315–336. Publisher: Reidel Dordrecht.
- Jaynes, E. T. (1986). Monkeys, kangaroos and n. *Maximum-Entropy and Bayesian Methods in Applied Statistics*, page 26.
- Jaynes, E. T. (2003). *Probability theory: The logic of science*. Cambridge university press.
- Jeans, J. (1921). *The dynamical theory of gases*. University Press.
- Jha, N. K. (2020). *Political Economy of Public Education Finance: Equity, Political Institutions, and Inter-school District Competition*. Lexington Books.
- Jimenez, B. S. and Hendrick, R. (2010). Is government consolidation the answer? *State and Local Government Review*, 42(3):258–270. Publisher: Sage Publications Sage CA: Los Angeles, CA.
- Kalenscher, T. and Van Wingerden, M. (2011). Why we should use animals to study economic decision making—a perspective. *Frontiers in Neuroscience*, 5:82.
- Keynes, J. M. (1909). A treatise on probability. *Diamond*, 3(2):12.
- Keynes, J. M. (1921). *A treatise on probability*. Macmillan and Company, limited.
- Khandker, S. R., Koolwal, G. B., and Samad, H. A. (2009). *Handbook on impact evaluation: quantitative methods and practices*. World Bank Publications.
- Khemlani, S. and Johnson-Laird, P. N. (2013). The processes of inference. *Argument & Computation*, 4(1):4–20.
- Kirman, A. (1993). Ants, rationality, and recruitment. *The Quarterly Journal of Economics*, 108(1):137–156.
- Kirman, A. (2010). *Complex economics: individual and collective rationality*. Routledge.
- Kitcher, P. (1989). Explanatory unification and the causal structure of the world.
- Klein, J. (2015). The cold war hot house for modeling strategies at the carnegie institute of technology. *Institute for New Economic Thinking Working Paper Series*, (19).



- Kolmogorov, A. N. (1983). On logical foundations of probability theory. In *Probability theory and mathematical statistics*, pages 1–5. Springer.
- Koopmans, T., Rubin, H., and Leipnik, R. (1950). Measuring the equation systems of dynamic economics, chapter 2 in statistical inference in dynamic economic models, tc koopmans (editor), cowles commission for research in economics, monograph no. 10.
- Kotz, S. and Kozubowski, T. (2001). *The Laplace distribution and generalizations: a revisit with applications to communications, economics, engineering, and finance*. Number 183. Springer Science & Business Media.
- Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86.
- Kuminoff, N., Smith, V. K., and Timmins, C. (2010). The New Economics of Equilibrium Sorting and its Transformational Role for Policy Evaluation. Technical Report w16349, National Bureau of Economic Research, Cambridge, MA.
- Kumpula, J. M., Onnela, J.-P., Saramäki, J., Kaski, K., and Kertész, J. (2007). Emergence of communities in weighted networks. *Physical review letters*, 99(22):228701.
- Kydland, F. E. and Prescott, E. C. (1982). Time to build and aggregate fluctuations. *Econometrica: Journal of the Econometric Society*, pages 1345–1370.
- Landau, L. D. and Lifshitz, E. M. (2013). *Course of theoretical physics*. Elsevier.
- Landauer, R. (1996). The physical nature of information. *Physics letters A*, 217(4-5):188–193.
- Lange, M. (2013). Really statistical explanations and genetic drift. *Philosophy of Science*, 80(2):169–188.
- Lange, M. (2016). *Because Without Cause: Non-Casual Explanations In Science and Mathematics*. Oxford University Press.
- Lange, M. (2020). What makes a scientific explanation distinctively mathematical? *The British Journal for the Philosophy of Science*.
- Lee, E. D., Broedersz, C. P., and Bialek, W. (2015). Statistical mechanics of the us supreme court. *Journal of Statistical Physics*, 160(2):275–301.
- Lee, E. D. and Daniels, B. C. (2018). Convenient interface to inverse ising (coniii): A python 3 package for solving ising-type maximum entropy models. *arXiv preprint arXiv:1801.08216*.
- Lifschitz, E. and Pitajewski, L. (1983). Physical kinetics. In *Textbook of theoretical physics*. 10.

- Lovász, L. (2010). Discrete and continuous: two sides of the same? In *Visions in Mathematics*, pages 359–382. Springer.
- Lucas, R. E. and Sargent, T. (1981). After keynesian macroeconomics. *Rational expectations and econometric practice*, 1:295–319.
- Marletto, C. (2021). The science of can and can't. *New Scientist*, 250(3330):34–37.
- Melo, J. (2021). A statistical equilibrium model of public school expenditures. \*.
- Mora, T. and Bialek, W. (2011). Are biological systems poised at criticality? *Journal of Statistical Physics*, 144(2):268–302.
- Mora, T., Walczak, A. M., Bialek, W., and Callan, C. G. (2010). Maximum entropy models for antibody diversity. *Proceedings of the National Academy of Sciences*, 107(12):5405–5410.
- Morgan, M. (1990). *The history of econometric ideas*. Cambridge University Press.
- Mundt, P. and Oh, I. (2019). Asymmetric competition, risk, and return distribution. *Economics Letters*, 179:29–32. Publisher: Elsevier.
- Munoz, M. A. (2018). Colloquium: Criticality and dynamical scaling in living systems. *Reviews of Modern Physics*, 90(3):031001.
- Musgrave, R. A. (1939). The voluntary exchange theory of public economy. *The quarterly journal of economics*, 53(2):213–237. Publisher: MIT Press.
- Nechyba, T. (2003). School finance, spatial income segregation, and the nature of communities. *Journal of Urban Economics*, 54(1):61–88.
- Nechyba, T. J. (2020). Tiebout sorting and competition. In *The Economics of Education*, pages 471–478. Elsevier.
- Nersessian, N. J. (1992). How do scientists think? capturing the dynamics of conceptual change in science. *Cognitive models of science*, 15:3–44.
- Nersessian, N. J. (2010). *Creating scientific concepts*. MIT press.
- Nersessian, N. J., Kurz-Milcke, E., Newstetter, W. C., and Davies, J. (2003). Research laboratories as evolving distributed cognitive systems. In *Proceedings of the annual meeting of the Cognitive Science Society*, volume 25.
- Oates, W. E. (1969). The Effects of Property Taxes and Local Public Spending on Property Values: An Empirical Study of Tax Capitalization and the Tiebout Hypothesis. *Journal of Political Economy*, 77(6):957–971.
- Parker, W. D. and Prechter, R. R. (2005). Herding: An interdisciplinary integrative review from a socionomic perspective. *Available at SSRN 2009898*.
- Pearl, J. (2009). *Causality*. Cambridge university press.

- Peterson, J., Dixit, P. D., and Dill, K. A. (2013). A maximum entropy framework for nonexponential distributions. *Proceedings of the National Academy of Sciences*, 110(51):20380–20385.
- Porta Mana, P. et al. (2017). Maximum-entropy from the probability calculus: exchangeability, sufficiency. Technical report, Center for Open Science.
- Rajapakse, I. and Smale, S. (2017). Emergence of function from coordinated cells in a tissue. *Proceedings of the National Academy of Sciences*, 114(7):1462–1467.
- Rényi, A. (1961). 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.
- Rosenkrantz, R. D. (2012). *ET Jaynes: Papers on probability, statistics and statistical physics*, volume 158. Springer Science & Business Media.
- Rotemberg, J. J. and Woodford, M. (1997). An optimization-based econometric framework for the evaluation of monetary policy. *NBER macroeconomics annual*, 12:297–346.
- Rubinstein, A. (1986). Finite automata play the repeated prisoner’s dilemma. *Journal of economic theory*, 39(1):83–96.
- Salmon, W. C. (1971). *Statistical explanation and statistical relevance*, volume 69. University of Pittsburgh Pre.
- Salmon, W. C. (1989). Four decades of scientific explanation.[part 1] introduction.
- Salvatier, J., Wiecki, T., and Fonnesbeck, C. (2015). Probabilistic Programming in Python using PyMC. *arXiv:1507.08050 [stat]*. arXiv: 1507.08050.
- Samuelson, P. A. (1954). The pure theory of public expenditure. *The review of economics and statistics*, 36(4):387–389. Publisher: JSTOR.
- Scarf, H. E. (1967). On the computation of equilibrium prices.
- Scharfenaker, E. (2020a). Statistical equilibrium methods in analytical political economy. *Journal of Economic Surveys*.
- Scharfenaker, E. (2020b). Statistical equilibrium methods in analytical political economy. *Journal of Economic Surveys*.
- Scharfenaker, E. and Foley, D. (2017). Quantal Response Statistical Equilibrium in Economic Interactions: Theory and Estimation. *Entropy*, 19(9):444.
- Scharfenaker, E. and Foley, D. K. (2021a). Unfulfilled expectations and labor market interactions: A statistical equilibrium theory of unemployment. Technical report, University of Utah, Department of Economics.

- Scharfenaker, E. and Foley, D. K. (2021b). Unfulfilled expectations and labor market interactions: A statistical equilibrium theory of unemployment. Technical report, University of Utah, Department of Economics.
- Scharfenaker, E. and Semieniuk, G. (2015). A Mixture Model for Filtering Firms' Profit Rates. In Frühwirth-Schnatter, S., Bitto, A., Kastner, G., and Posekany, A., editors, *Bayesian Statistics from Methods to Models and Applications*, volume 126, pages 153–164. Springer International Publishing, Cham. Series Title: Springer Proceedings in Mathematics & Statistics.
- Scharfenaker, E. and Semieniuk, G. (2017). A statistical equilibrium approach to the distribution of profit rates. *Metroeconomica*, 68(3):465–499. Publisher: Wiley Online Library.
- Schelling, T. C. (2006). *Micromotives and macrobehavior*. WW Norton & Company.
- Schinckus, C. (2016). 1996–2016: Two decades of econophysics: Between methodological diversification and conceptual coherence. *The European Physical Journal Special Topics*, 225(17):3299–3311.
- Schinckus, C. (2021). The santa fe institute and econophysics: A possible genealogy? *Foundations of Science*, pages 1–21.
- Schneidman, E., Bialek, W., and Berry, M. J. (2003). Synergy, redundancy, and independence in population codes. *Journal of Neuroscience*, 23(37):11539–11553.
- Sequoiah-Grayson, S. (2007). The metaphilosophy of information. *Minds and Machines*, 17(3):331–344.
- Shaikh, A. (2016). *Capitalism: Competition, conflict, crises*. Oxford University Press.
- Shore, J. and Johnson, R. (1980). Axiomatic derivation of the principle of maximum entropy and the principle of minimum cross-entropy. *IEEE Transactions on information theory*, 26(1):26–37.
- Simon, H. A. (1957). Models of man; social and rational.
- Simon, H. A. (1977). Causal ordering and identifiability. In *Models of Discovery*, pages 53–80. Springer.
- Sims, C. A. (2003). Implications of rational inattention. *Journal of monetary Economics*, 50(3):665–690. Publisher: Elsevier.
- Sivia, D. and Skilling, J. (2006). *Data analysis: a Bayesian tutorial*. OUP Oxford.
- Skilling, J. (1988). The axioms of maximum entropy. In *Maximum-entropy and Bayesian methods in science and engineering*, pages 173–187. Springer.
- Sklar, L. (1995). *Physics and chance: Philosophical issues in the foundations of statistical mechanics*. Cambridge University Press.

- Skyrms, B. (1987). Updating, supposing, and maxent. *Theory and Decision*, 22(3):225–246.
- Smale, S. (1976). Dynamics in general equilibrium theory. *The American Economic Review*, 66(2):288–294.
- Smets, F. and Wouters, R. (2007). Shocks and frictions in us business cycles: A bayesian dsge approach. *American economic review*, 97(3):586–606.
- Smith, A. (1937). *The wealth of nations [1776]*, volume 11937. na.
- Sokolov, I. M. and Klafter, J. (2005). From diffusion to anomalous diffusion: a century after einstein’s brownian motion. *Chaos: An Interdisciplinary Journal of Nonlinear Science*, 15(2):026103.
- Soofi, E. S. and Retzer, J. J. (2002). Information indices: unification and applications. *Journal of Econometrics*, 107(1-2):17–40. Publisher: Elsevier.
- Sornette, D. (2007). Probability distributions in complex systems. *arXiv preprint arXiv:0707.2194*.
- Stachel, J. J. (2005). Einstein’s miraculous year: five papers that changed the face of physics.
- Stalnaker, R. C. (1968). A theory of conditionals. In *Ifs*, pages 41–55. Springer.
- Stauffer, D. (2008). Social applications of two-dimensional ising models. *American Journal of Physics*, 76(4):470–473.
- Stigler, S. M. (1986). *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press.
- Strogatz, S. H. (2018). *Nonlinear dynamics and chaos with student solutions manual: With applications to physics, biology, chemistry, and engineering*. CRC press.
- Tiebout, C. M. (1956). A Pure Theory of Local Expenditures. *Journal of Political Economy*, 64(5):416–424.
- Traub, J. F. (2003). Information-based complexity. In *Encyclopedia of Computer Science*, pages 850–854.
- Tsallis, C. (1988). Possible generalization of boltzmann-gibbs statistics. *Journal of statistical physics*, 52(1):479–487.
- Tsallis, C. and Brigatti, E. (2004). Nonextensive statistical mechanics: A brief introduction. *Continuum Mechanics and Thermodynamics*, 16(3):223–235.
- Turing, A. M. (1952). The chemical basis of morphogenesis. *Bulletin of mathematical biology*, 52(1):153–197.

- Uffink, J. (1995). Can the maximum entropy principle be explained as a consistency requirement? *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 26(3):223–261.
- van Fraassen, B. C. (2011). On stance and rationality. *Synthese*, 178(1):155–169.
- Van Fraassen, B. C. et al. (1980). *The scientific image*. Oxford University Press.
- Vehtari, A., Gelman, A., Simpson, D., Carpenter, B., and Bürkner, P.-C. (2019). Rank-normalization, folding, and localization: An improved <sup>f</sup>or assessing convergence of mcmc. *arXiv preprint arXiv : 1903.08008*.
- Vinković, D. and Kirman, A. (2006). A physical analogue of the schelling model. *Proceedings of the National Academy of Sciences*, 103(51):19261–19265.
- Waldrop, M. M. (1993). *Complexity: The emerging science at the edge of order and chaos*. Simon and Schuster.
- Wold, H. (1954). Causality and econometrics. *Econometrica: Journal of the Econometric Society*, pages 162–177.
- Wolfram, S. (1983). Statistical mechanics of cellular automata. *Reviews of modern physics*, 55(3):601.
- Woodward, J. (1989). Data and phenomena. *Synthese*, pages 393–472.
- Wu, F.-Y. (1982). The potts model. *Reviews of modern physics*, 54(1):235.
- Yang, J. (2018). Information theoretic approaches in economics. *Journal of Economic Surveys*, 32(3):940–960.

# Appendix A

## Résumé de thèse

L'histoire de l'utilisation des idées de la physique statistique dans les sciences sociales est une histoire de retour au pays. Les rapatriements peuvent être complexes à gérer.

Au 19<sup>ème</sup> siècle, les physiciens intéressés par l'étude des propriétés de la matière ont été confrontés au formidable défi de réfléchir à la façon de prédire le comportement de grandes collections d'atomes se déplaçant à travers l'espace en rebondissant au hasard les uns sur les autres dans un récipient fermé. Les collections étaient tellement grandes, de l'ordre de  $10^{23}$ , qu'il semblait désespéré d'essayer de les suivre individuellement en utilisant les lois du mouvement de Newton.

James Clerk Maxwell, inspiré par le nombre croissant de régularités statistiques qui étaient trouvées dans les données sociales, notamment par des personnalités comme Laplace, John Graunt, et Edmund Halley, proposa de relever le défi en recensant les populations d'atomes. Partant de l'hypothèse que les propriétés des atomes étaient gaussiennes, il a rapidement réalisé, avec Ludwig Boltzmann, que les distributions exponentielles étaient mieux adaptées à l'étude des propriétés de la matière d'un point de vue "social-statistique".

Cette constatation a conduit au développement de la mécanique statistique, un domaine qui a connu un regain d'intérêt pour les sciences sociales récemment. Ce regain d'intérêt pour l'application des idées de la physique statistique dans les sciences sociales est le résultat de trois évolutions principales. La première est l'augmentation de la puissance de calcul et de la quantité de données disponibles. Le second est la prise de conscience que les institutions sociales et les groupes d'individus s'auto-

organisent de manière interactive en structures complexes qui complexes qui ne semblent pas être facilement saisissables par des perspectives de théorie du contrôle ou mécanistes. La troisième est l'essor des collaborations interdisciplinaires entre économistes, physiciens, sociologues, biologistes et théoriciens de l'information, entre autres, sous l'égide de la science des systèmes complexes.

Pendant les trente dernières années, les économistes politiques d'orientation classiques ont, de manière indépendante, exploré différentes façons de théoriser les marchés et les institutions économiques à travers le prisme de la mécanique statistique. L'article de Duncan Foley de 1994 intitulé *A statistical equilibrium theory of markets* a été la première tentative en économie de théoriser les échanges décentralisés et les institutions économiques avec cette approche. Dans un article de 2017, Ellis Scharfenaker et Duncan Foley proposent une distribution de probabilité exponentielle pour étudier les environnements économiques décentralisés. Il s'agit de la distribution QRSE qui est utilisée et étudiée en détail dans cette thèse. La distribution QRSE est autant une théorie de l'interaction économique décentralisée qu'un modèle de probabilité descriptif, similaire à la distribution de Gauss et de Laplace.

Cette thèse mêle deux projets distincts mais interdépendants. Le premier est l'application du cadre de l'équilibre statistique à l'étude des données des biens publics locaux. Il constitue une première étape d'un projet à plus long terme visant à construire une image quantitative d'un processus complexe et évolutif, celui de la fragmentation des gouvernements locaux aux États-Unis. Le second est une exploration philosophique de l'inférence modale (basée sur des principes) dans le cadre des modèles d'équilibre statistiques en économie. Cette analyse se situe à l'intersection des théories sémantiques de l'information, de la mécanique statistique et de l'économie politique.

### **Complexité et économie**

La science des systèmes complexes étudie comment de grandes collections de composants interagir localement à différentes échelles s'auto-organiser en modèles et comportements globaux. Au cœur de la science de la complexité se trouve l'idée que les tous sont irréductibles à leurs parties, et que l'analyse et la prédiction de la dynamique complexe des parties entières dans les systèmes sociaux, biologiques et physiques nécessitent de nouvelles inter- et méthodes transdisciplinaires.



Le premier grand thème qui définit la vision de la complexité est l'étude des **interactions**. Les composants des systèmes complexes peuvent être vus comme interagissent de multiples façons les uns avec les autres et avec leurs environnement. La science de la complexité résiste à l'idée qu'une telle les interactions doivent être étudiées à travers des cadres réductionnistes isolés. Ce n'est pas seulement une question de élucider les relations en réseau et hiérarchiques qui existent entre les composants d'un système, mais aussi de comprendre comment les modèles d'interdépendance forte et faible conduisent à des statistiques distinctes régularités et signatures.

Un deuxième thème majeur est l'*émergence*. Les modèles globaux non triviaux qui émergent des interactions locales en réseau imposent des obstacles à la pensée linéaire et statistique traditionnelle. Molécules d'air et de vapeur se combinent pour former des tornades, des milliards de neurones dans le cerveau interagissent pour produire de la cognition, et des réseaux bancaires mondiaux mais fragiles les institutions et les bilans font basculer par inadvertance les politiques dans effondrement financier et crise politique. L'étude de l'émergence dans les systèmes complexes peuvent être abordés par des méthodes dynamiques, statistiques et homologues informatiques.

Le problème de l'analyse du comportement émergent et dépendant du temps des des systèmes complexes à plus de deux degrés de liberté conduit à troisième grand thème, qui est celui de l'analyse des **systèmes dynamiques** et *chaos*. L'interaction de composantes linéaires prévisibles indépendantes conduit à des ensembles présentant des trajectoires non linéaires et chaotiques dans un l'espace des phases du système. Il existe deux principaux types de systèmes dynamiques : *équations différentielles* et *cartes itérées* (ou différence équations). Systèmes non linéaires à plus de deux degrés de liberté sont souvent impossibles à résoudre analytiquement, du fait que les principe de superposition échoue majoritairement dans ces contextes. C'est une autre façon de dire que dans les systèmes non linéaires le tout est précisément pas égal à la somme de ses parties. Pour les systèmes présentant comportement chaotique, petites erreurs dans la mesure de la les conditions s'amplifient avec le temps et les systèmes deviennent efficacement imprévisible. Comme les systèmes dynamiques stables, les systèmes chaotiques ont tendance à attracteurs, mais ceux-ci se révèlent souvent complexes et difficiles espaces topologiques (Albin and Foley, 1998; Strogatz, 2018). Des exemples courants de les systèmes non linéaires et chaotiques sont le double pendule, le temps et les marchés financiers.

La science de la complexité pointe les limites de l'analyse de tels systèmes dans termes d'approches de loi dynamique de forme réduite. Les barrières que de tels les systèmes posés aux dynamiques ont engendré des connaissances méthodologiques profondes réflexions, et ont conduit naturellement au développement de alternatives géométriques/topologiques, informatiques et statistiques.

Les trois grands thèmes qui ont émergé de l'interdisciplinarité de SFI l'exploration des questions économiques ont été présentées dans le volume de 1988 *L'économie en tant que système complexe en évolution* (Anderson, 1988). Notes de David Pines que SFI a initialement établi trois groupes de travail sur les sujets de 'Cycles', 'Webs' et 'Motifs' (voir Schinckus, 2021). "Cycles" axés sur comportement déterministe non linéaire et chaos, "Webs" axés sur l'interaction et l'émergence, en particulier vues à travers le prisme de science des réseaux et modélisation à base d'agents. Et 'Patterns' axé sur étudier les propriétés statistiques et les invariants d'économies complexes systèmes. Schinckus (2021) note que l'on trouve un intérêt décroissant pour le chaos *déterministe* tout au long de la l'évolution du programme de recherche en économie du SFI.

Propulsé par la vision que les processus axés sur le marché sont complexes, systèmes de haute dimension, dynamiques et stochastiques, l'éconophysique a commencé à se développer au début des années 1990 comme une puissante alternative à la conduite travail empirique. L'afflux massif de "quants" et de physiciens dans l'actif des sociétés de gestion ont également accompagné ce développement, une demande croissante de renouvellement recherche interdisciplinaire à l'intersection de la physique, de la biologie, théorie des jeux et économie. L'Institut Santa Fe continue de jouer un rôle clé dans le soutien de la recherche liée à l'éconophysique.

Parallèlement aux développements de la littérature éconophysique mis en évidence dans la section précédente, qui étaient pour la plupart menées par des physiciens, un nombre croissant de recherches empiriques a émergé ces trente dernières années au carrefour des hétérodoxes économie politique, économie non walrasienne, théorie de l'information et mécanique statistique. Cette littérature se distingue en ce qu'elle est pas une tentative tabula rasa d'appliquer les idées de la physique statistique à l'étude des processus économiques. Cette ligne de travail tente de récupérer et soumettre à un examen empirique, en utilisant les cadres de mécanique statistique

de la théorie de l'information, hypothèses bien établies à partir du stock de classique<sup>1</sup> et traditions non walrasiennes (Scharfenaker, 2020a).

Une critique que cette littérature propose à la fois aux approches basées sur la simulation et aux approches statistiques, issues de l'éconophysique, est l'application lâche des concepts de symétrie et d'ergodicité. Par exemple, comme lorsque Yakovenko traite l'argent et les flux financiers comme une forme d'énergie conservée afin de dériver les lois Pareto-Exponentielles pour les distributions de revenus américaines. (Dragulescu and Yakovenko, 2000b).

Un catalyseur majeur pour le développement de cette ligne de raisonnement est venu hors des travaux de Farjoun and Machover (1983). Dans Farjoun and Machover (1983), les auteurs critiquent la traitement déterministe des agrégats économiques dans la politique classique économique et proposer un cadre probabiliste capable de capter présence de fluctuations aléatoires autour d'une tendance centrale taux de profit. Leurs recherches ont donné une nouvelle vie (statistique) au Smithian et au Marxian hypothèses d'égalisation de long terme du taux de profit, et aux idée classique selon laquelle les taux observés gravitent autour d'une "quantité naturelle". Leur théorie considérations les ont amenés à conjecturer une densité gamma pour le profit taux de distribution stationnaire, mais cela s'est avéré peu convaincant puisque la distribution gamma est contrainte au domaine positif et les taux de profit négatifs sont loin d'être rares dans le capitalisme moderne. économies (Scharfenaker and Semieniuk, 2017).

### **Complexité et probabilité**

Les systèmes sociaux, biologiques et économiques complexes sont souvent bien décrits par des distributions de probabilité exponentielles et à queue lourde. Quelques exemples sont les distributions des taux de profit, les populations des villes et les fluctuations des marchés financiers (voir Peterson et al., 2013; Sornette, 2007). L'une des raisons de distinguer les distributions est qu'elles peuvent être liées à différents mécanismes, et l'examen de leur forme entière donne des indices sur les processus en jeu. Souvent, si les données ne manquent pas, trouver la forme d'une distribution peut être aussi simple que de regarder un histogramme ou d'ajuster une fonction mathématique appropriée. L'ajustement d'un modèle de probabilité disponible peut être suivi en réfléchissant à la raison pour laquelle cette fonction particulière est un

---

<sup>1</sup>Économie politique classique

bon ajustement. Autrement dit, quelles sont certaines des raisons qui pourraient amener un système à présenter cette forme particulière.

L'étude des distributions remonte à Abraham De Moivre et Laplace. À l'époque de De Moivre et de Laplace, un problème central de la science était celui de trouver des "courbes d'erreur" universelles. Laplace a montré comment les fluctuations des statistiques sociales pouvaient être décrites par la distribution normale, initialement proposée par De Moivre dans sa **Doctrine des chances** (1718) comme méthode d'approximation des coefficients binomiaux. Le problème de trouver une loi d'erreurs était couramment rencontré en astronomie et en géodésie, où le but était de décrire les variations de mesure à partir d'une vraie grandeur physique, par ex. la distance à la lune (Geraci and Borja, 2018). L'étude de la distribution éponyme de Laplace (1774), qui est une courbe d'erreur non normale, a récemment fait l'objet d'un regain d'intérêt. Les distributions de Laplace et normale appartiennent toutes deux à la famille exponentielle. Ce qui les distingue est la spécification de leurs fonctions d'erreur ; le Laplace est une exponentielle de la *magnitude* numérique de l'erreur, tandis que la Normale est une exponentielle du *carré* de l'erreur (voir Kotz and Kozubowski, 2001).

L'une des raisons d'utiliser une distribution de Laplace sur une distribution normale est s'il existe une hétérogénéité dans la mesure des erreurs. Geraci and Borja (2018) donne l'exemple d'astronomes à différents endroits essayant de mesurer la distance à un objet céleste. Chaque astronome essaie de mesurer la même quantité, mais inévitablement, leurs mesures auront des variances différentes. Donc, si vous regardez les ensembles de données de chaque astronome, ils seront probablement normaux. Mais si vous êtes chargé de collecter les mesures de tous les astronomes dans un seul histogramme, vous obtiendrez probablement une courbe de type Laplace. La variabilité observateur-instrument est donc une raison possible d'utiliser la distribution de Laplace. Cette distribution apparaît couramment en hydrologie, en finance, en reconnaissance de la parole et des formes, et plus récemment, dans l'économie politique statistique et les littératures d'organisation industrielle (voir Bottazzi and Secchi, 2005).

Il y a des cas où il peut être intéressant, non seulement de trouver la forme appropriée d'une distribution, mais de savoir quels principes générateurs pourraient conduire à sa signature statistique distinctive. Barabási and Albert (1999), par exemple, ont montré comment des modèles de réseaux dynamiques peuvent conduire

à des distributions de loi de puissance sans échelle. Scharfenaker and Foley (2017) montrent comment un principe de rétroaction du marché conduit à une courbe de type Laplace-Normal pour la distribution des taux de profit.

Les dérivations basées sur des principes sont mieux comprises comme des problèmes *inverse*. Ils sont inverses car le but est de déduire l'ensemble des mécanismes ou règles sous-jacents possibles qui génèrent les données. « Principe » signifie simplement que nous utilisons une description suffisamment générale pour englober un large éventail de spécifications fonctionnelles ou dynamiques possibles. La mécanique statistique de la théorie de l'information fournit un cadre puissant pour résoudre les problèmes inverses, en particulier ceux dans lesquels les données sont distribuées de manière exponentielle. La méthode de maximisation de l'entropie contrainte est un principe variationnel qui donne une solution informationnellement efficace au problème inverse.

Dans la science de la complexité, nous sommes confrontés à des systèmes qui ont de nombreux degrés de liberté. Le comportement d'un système  $S$  peut être caractérisé comme dépendant d'un ensemble de variables  $\{x_j\}$ , pour  $j = 1, 2, \dots, N$ . Lorsque l'on vise une description probabiliste de  $S$ , dans laquelle on ne décrit ni ne calculent les trajectoires individuelles de chacun  $x_j$ , la première question naturelle à se poser est la suivante : qu'est-ce que la distribution de probabilité conjointe  $P(\{x_j\})$  de tous ces variables ? Dans la vue la plus simple, nous demandons ce que  $P$  est à un moment donné point dans le temps. Autrement dit, quelle est la probabilité conjointe des variables à partir de laquelle nous pouvons échantillonner un instantané en coupe du système ?

Du point de vue de la physique statistique, les  $x_j$ s sont les variables d'état microscopiques pour les composants du système, tels que les atomes et molécules. Un objectif central de la physique statistique est de construire un pont qui nous emmène d'une compréhension de ce que microscopique composants font à une description au niveau macro de l'émergence comportement collectif d'un grand nombre de pièces en interaction.

Les ponts micro-macro de la science de la complexité sont avant tout statistique et non déductive. Les matériaux de construction nécessaires à la l'assemblage de ces ponts complexes sont : i) des connaissances pratiques, ou hypothèses, sur les principes régissant l'interaction au niveau micro, et ii) probabilité et données.

Il y a une simplicité et une universalité émergentes qui sortent de la problème de description de systèmes complexes en termes de distribution conjointe  $P(\{x_j\})$ . De nouvelles variables et équations émergent qui ont pour propriété remarquable d'être applicable à une large gamme de systèmes à différentes échelles, physiques, sociales et biologiques. domaines. Dans les systèmes physiques classiques, comme les gaz parfaits, l'analyse statistique d'un grand nombre de molécules se déplaçant au hasard et indépendamment au sein d'un volume conduit à l'émergence de lois et constantes déterministes, par ex. la loi des gaz parfaits (voir Feynman, 1965). la pression sur une section d'un volume de gaz, en raison de l'indépendance et aléatoire collectif de ses particules, s'avère être la moyenne force exercée par toutes les particules, divisée par la surface concernée. ;  $P = \frac{\langle F \rangle}{A}$ . Ainsi, dans de tels systèmes, les effets sont "additif" et sans surprise. Ce qui est intéressant, cependant, c'est notion qu'une constante physique émerge de la statistique agrégation de collisions aléatoires de particules sur un pan de mur (Bialek, 2017).

Pour les systèmes déterministes classiques, qu'est-ce qui fonctionne dans garantissant l'émergence de constantes, ou à déterminer que les moyennes peuvent effectivement être considérées comme des nombres exacts, est l'élément central théorème limite (CLM) (voir Fischer, 2010). En un mot, ce que dit la CLM, c'est que la valeur moyenne d'un grand nombre d'observations *indépendantes* suivra a Distribution normale (en forme de cloche). Mais plus crucialement pour physiciens, il indique que comme le nombre d'observations tend à l'infini ( $N \rightarrow \infty$ ), la variance de la distribution va rétrécir et avoir tendance à disparaître. Étant donné que les systèmes considérés par physique classique contiennent souvent un très grand nombre de particules, l'ordre du numéro d'Avogadro ( $10^{23}$ ), les variations autour de la moyenne peut être effectivement ignorée. C'est ainsi que le déterminisme et constantes émergent dans un cadre probabiliste, et c'est ce qui permet aux physiciens de se concentrer sur leurs équations et évite d'avoir à modifier constamment les principales hypothèses distributionnelles (les lois physiques sont stables).

Dans le monde social et biologique, où le hasard est contraint par des principes distincts d'interaction et d'organisation (ex. feedback), il faut manipuler le CLM avec précaution. De plus, dans de tels contextes ce qui finit par être d'un intérêt fondamental est le type de variation que nous percevoir autour de la moyenne. De la même manière que ce qui distingue un chute de pierre d'un oiseau sont les capacités d'adaptation de ce dernier, ce qui distingue les mesures d'organismes vivants des mesures de solides est la statistique non normale du premier variance. Quelle que

soit la limite, la variance nulle ou infinie ne sont pas des recettes efficaces pour maintien de l'organisation biologique et sociale.

### **Décentralisation fiscale et l'hypothèse de Tiebout**

Un problème central dans l'analyse économique de l'offre de biens publics est le manque d'incitations des électeurs à révéler leur véritable demande. Tiebout (1956) a proposé d'étudier le problème de biens publics locaux à travers un modèle de quasi-marché dans lequel les électeurs-consommateurs expriment leurs préférences pour les biens publics locaux en entrer et sortir des juridictions locales. L'hypothèse de Tiebout stipule que les juridictions locales auront tendance à trier en blocs homogènes avec respect de la demande de biens publics locaux et des niveaux de taxation, lorsque ceux-ci sont considérés comme une forme de prix dans le modèle. L'idée centrale derrière cette hypothèse est qu'un équilibre de tri de Tiebout, s'il existe, éliminera les inefficacités associées à la diversité de la demande ; les ménages ne seront pas obligés de payer des impôts plus élevés qu'ils ne le feraient préfèrent autrement, ni ne sont en mesure de faire du free-ride sur les voisins contributions relativement plus élevées des ménages au service fiscal local.

L'article de Tiebout de 1956 a été, et continue d'être, un important catalyseur pour une recherche renouvelée dans l'analyse de l'administration décentralisée la finance. La principale contribution de Tiebout a été de défier la norme croyance de l'époque qu'il n'y avait pas de solution fondée sur le marché à la problème de la fourniture de biens publics locaux. Il l'a fait en plaçant la localisation géographique et la mobilité au cœur de l'analyse, et par en utilisant ce dernier comme proxy pour la révélation des choix et des préférences.

Le consommateur ne peut en aucun cas éviter de révéler son préférences dans une économie spatiale. La mobilité spatiale offre contrepartie de produits locaux au voyage d'achat du marché privé. Tiebout (1956)

Le mécanisme hypothétique de concurrence de Tiebout, vu à travers les lentille des modèles d'équilibre néoclassiques, peut être comprise comme ayant trois traits essentiels (Nechyba, 2020). Le premier est que lorsque les communautés locales sont considérées comme analogues à des concurrents entreprises, la décentralisation permettra la fourniture optimale des services publics services en présence de demandes hétérogènes des ménages. le deuxièmement, la notion que la concurrence réduira les

incitations aux gouvernements à se comporter comme des ‘Léviathans’ (Jha, 2020). La dernière notion repose sur la conviction que l’approvisionnement décentralisé et la fourniture de biens publics locaux contrera la tendance des gouvernements à prélever arbitrairement des impôts plus élevés sur leurs résidents (Brennan and Buchanan, 1980; Jimenez and Hendrick, 2010). Une troisième caractéristique implique que dans ‘équilibre’, le mécanisme de Tiebout conduira les ménages à trier (à dans une certaine mesure) sur la base de la capacité de paiement et du revenu du ménage. Cette dernière caractéristique est, bien sûr, loin d’être sans équivoque souhaitable. La caractérisation du tri de Tiebout comme un tri optimal résultat, possédant un mérite intrinsèque principalement en raison de sa capacité pour générer des gains d’efficacité productifs, s’avère en contradiction avec notions juridiques de base concernant les droits des citoyens à l’éducation (Jha, 2020). Un équilibre dans lequel les dépenses des écoles publiques et la qualité sont fortement corrélées aux caractéristiques des ménages présente des défis moraux et juridiques non négligeables. La portée de ces défis a été dûment mis en évidence par la poursuite des batailles juridiques et débats politiques sur les inégalités de financement dans l’éducation publique américaine système depuis 50 ans (Hertert et al., 1994; ?). Le fait que l’optimum résultat dans une formulation hautement idéalisée de l’hypothèse de Tiebout s’avère fondamentalement en contradiction avec ce qui peut être souhaitable à au niveau de la politique ou des ménages (ou est à tout le moins hautement contestable), n’exclut pas que le tri et la le rationnement des ressources gouvernementales sont en fait façonnés par des les forces. Elle pose cependant de sérieux défis à la modélisation et spécification des primitives microéconomiques qui animent processus concurrentiel.

L’un des problèmes fondamentaux qui découle de l’utilisation de modèles d’équilibre général est qu’ils nous obligent à considérer distributions économiques comme résultant principalement et mécaniquement de interaction d’agents optimisateurs (ménages et gouvernements) dont les préférences sont pleinement satisfaites. Il s’agit d’une stratégie de modélisation qui exclut a priori la possibilité que les attentes des agents restent insatisfaits à l’équilibre.

Dans le cadre de l’analyse économique des déterminants de hétérogénéité des niveaux de dépenses et de la demande des écoles publiques, où une grande partie des débats théoriques et politiques tournent autour de la la reconnaissance que les marchés de l’éducation sont structurés par des politiques et des processus de production, l’exigence d’une optimisation complète comportement cohérent avec les équilibres observés est difficile à soutenir. De plus, en l’absence de caractérisations



plausibles pour les environnements microéconomiques et politiques, il est difficile de voir comment des idées utiles peuvent être extraites de l'étude des forces d'équilibre et les résultats.

Cette préoccupation a orienté la littérature sur Tiebout et le financement de l'éducation vers une voie de construction de modèles de mathématiques et de complexité de calcul, où des éléments tels que le vote hétérogène les préférences et les intrants non financiers sont incorporés afin de fournir des descriptions plus riches et plus empiriquement pertinentes, plausible d'un point de vue microéconomique (Kuminoff et al., 2010; Nechyba, 2003).

Il y a eu un changement récent dans la littérature de la construction générale modèles d'équilibre à la construction de modèles d'équilibre computationnels qui à cheval sur un large spectre couvrant à la fois purement théorique et formulations empiriquement motivées. Comme Nechyba (2020) notes, tous ces modèles commencent par spécifier explicitement le sous-jacent structure mathématique de l'environnement économique modélisé. Ce c'est-à-dire qu'ils fournissent une spécification entièrement structurelle pour le ménage préférences, fonctions de production scolaire, distributions pour les ménages caractéristiques du modèle (telles que le revenu), ainsi que des descriptions du processus politique (modèles de vote), du budget l'environnement et les marchés du logement et des écoles privées. Par études de simulation, l'étude des résultats d'équilibre dans ces on s'attend alors à ce que les modèles fournissent des informations politiques significatives et fournir un bac à sable pour expérimenter la politique hors échantillon interventions.

Le problème est que la pertinence de ces simulations études dépend de la plausibilité empirique des structure microéconomique utilisée pour représenter la structure sous-jacente mécanique du processus de génération de données, et sur la confiance que nous pouvons avoir dans les paramètres du modèle pour saisir de manière adéquate de manière empirique processus pertinents. Mais si l'on considère le fait que la tâche de déterminer la plausibilité empirique d'un modèle donné spécification pour les environnements sociaux complexes avec de grands degrés de la liberté peut être mal posée et sous-déterminée (Scharfenaker and Foley, 2017), alors il est difficile de voir comment la voie de l'augmentation de la complexité des modèles dans modèles d'équilibre général (ou computationnels) entièrement micro-fondés est susceptibles de produire des résultats non ambigus et non biaisés sur le plan normatif. Il y a un très large continuum de modèles et de solutions cohérents avec tout

ensemble de données circonstancielle et de preuves (Golan, 2018). Et une erreur de spécification peut apparaître soit au niveau des formes fonctionnelles (production et préférence fonctions), critères ou fonctions de décision, la spécification des modèles de vote, ainsi que des priors pour les entrées stochastiques dans le modèle (par exemple, les caractéristiques du ménage).

Cette thèse adopte une approche alternative qui utilise le maximum méthodes d'entropie et un cadre d'équilibre statistique pour modéliser et étudier l'effet de la concurrence dans la formation des distributions des dépenses publiques d'éducation pour la période 2000-2016 États-Unis. Les avantages de cette entropie maximale/statistique cadre d'équilibre sont nombreux, mais un cadre central que nous considérons voilà qu'il nous permet d'étudier la dynamique concurrentielle des USA marché de l'éducation publique (un système social complexe avec de grands diplômes de liberté) sans avoir à s'engager *a priori* dans une lourde échafaudage mathématique de la structure microéconomique sous-jacente environnement. Elle permet plutôt d'étudier une façon plausible de la structure probabiliste des dépenses des districts scolaires peut être vu émerger d'une paire d'actions comportementales et institutionnelles parcimonieuses. contraintes que nous imposons à l'environnement microéconomique sous-jacent.

L'un des aspects intéressants de la formulation originale de Tiebout en 1956 est qu'il reste sans engagement à l'égard de tout formulation du modèle d'équilibre, même s'il met en évidence un ensemble de faits et caractéristiques que le processus concurrentiel hypothétique devrait rencontrer. Mais comme l'histoire des tests empiriques du Tiebout hypothèse a montré (Edel and Sclar, 1974; Nechyba, 2020; Oates, 1969), il n'est pas vraiment possible de tester toutes les hypothèses du plus grand Tiebout hypothèse à la fois sans se heurter à des contradictions. Pour exemple, en testant l'hypothèse de mobilité résidentielle parallèlement capitalisation des variables budgétaires dans les prix des logements peut aller à l'encontre L'hypothèse d'efficacité plus large de Tiebout (puisque la présence de capitalisation est la preuve de l'existence d'une demande excédentaire de logement dans la juridiction où les taxes et les niveaux de service locaux sont en majuscule) (Epple and Nechyba, 2004a)>. De même, comme nous pointé plus haut, l'existence du tri de Tiebout est d'être mieux compris comme un résultat macroéconomique potentiellement inattendu (à moins une partie des ménages). Vu sous cette lumière, le perspective de pouvoir concilier les politiques sous-jacentes contradictions du marché de l'éducation avec l'hypothèse d'une maximiser les ménages dans un modèle d'équilibre général

semble farfelu. Cela dit, nous croyons qu'il est nécessaire et qu'il y a amplement de place pour se concentrer sur certains aspects de l'hypothèse de Tiebout, et qu'il est possible d'étudier le support empirique de l'affirmation générale selon laquelle les dépenses en biens publics locaux sont fortement façonnées (et au moins expliquée en partie) par des forces concurrentielles et un rationnel limité arbitrage qui a lieu au niveau des ménages en termes de consommation d'éducation.

Cette thèse applique le cadre théorique de la Réponse Quantique Modèle d'équilibre statistique (QRSE) développé dans Scharfenaker and Foley (2017). Comme mentionné ci-dessus, l'approche adoptée par le document n'est pas totalement agnostique par rapport à la structure microéconomique, car elle utilise une entropie modèle contraint de mobilité résidentielle et de choix juridictionnel la caractérisation de base du comportement des ménages. Cette ligne de base modèle fait l'hypothèse comportementale que les ménages essaient de maximiser le taux de rentabilité des dépenses fiscales (considérées comme les prix des services d'éducation), sous la contrainte d'une capacité limitée à traiter les signaux du marché et les signaux politiques. Dans le contexte de faible revenu et ménages des centres-villes, nous avons avancé l'idée que ce la capacité peut également être interprétée comme une forme de restriction économique agence. Les grandes lignes de ce modèle comportemental sont très similaires à celui trouvé dans le programme d'inattention rationnelle des Sims (Sims, 2003). Grâce à l'insertion d'une contrainte de la théorie de l'information sur le programme de maximisation de l'utilité de ménages, cette spécification de base fournit une description probabiliste du comportement des ménages.

Retours scolaires dans les districts scolaires à travers les États-Unis pour la La période 2000-2016 présente des pics distincts, positivement biaisés distributions avec des queues droites de largeur variable. La mise en forme de leur distribution d'équilibre statistique est le résultat d'une évolution processus de concurrence inter-juridictionnelle, ménage résidentiel tri sur la base d'un large ensemble de caractéristiques (telles que revenu) et les régimes politiques changeants aux niveaux local, étatique et fédéral niveaux.

En utilisant un cadre d'équilibre statistique, dans cet article, nous avons cherché à examiner le rôle joué par la concurrence entre les juridictions et choix des ménages pour façonner la distribution observée revient pour un cas d'ensemble complet qui couvre tous les districts scolaires américains dans la période 2000-2016. Il s'agit d'un échantillon considérablement plus important que le celles trouvées dans d'autres

traitements empiriques de la littérature, qui se concentrent généralement sur des États ou des régions uniques. Un aspect important de nos résultats empiriques est qu'ils corroborent la nécessité de dissocier les notions normatives sur l'efficacité du marché des affirmations sur la présence du tri Tiebout et de la concurrence. Nous avons proposé un modèle parcimonieux qui capture de manière significative la différence entre la concurrence et le tri forces via deux ensembles de paramètres d'échelle et de localisation.

Notre analyse empirique corrobore également les résultats précédents du QRSE la littérature concernant l'utilisation des contraintes de rétroaction comme caractérisations significatives de la concurrence sur un marché décentralisé réglages. L'histogramme et l'ajustement du modèle affichés dans cet article sont clairement évocateur du rôle joué par la concurrence décentralisée dans l'accentuation des pics modaux, et par un marché élevé (loin d'être concurrentiel) taux en créant un biais positif.

Au cœur de l'hypothèse de Tiebout se trouve l'idée que la mobilité peut être utilisée comme proxy de la demande de biens publics. Dans la formulation abstraite décrite par Tiebout (1956), parfaite tri résidentiel des ménages sur la base des préférences pour les les paquets de services fiscaux sont supposés être une condition d'équilibre.

La littérature empirique sur le financement de l'éducation a du mal à donner un sens des preuves solides montrant les incompatibilités entre le tri ménager et la parfaite concurrence inter-juridictionnelle, ainsi que les lignes définies par Tiebout. Généralement, la plupart des épreuves prennent l'un des deux approches pour résoudre ce problème. Une première approche porte sur mesurer l'hétérogénéité entre et au sein des juridictions, et sur associant cette analyse de variance à une analyse de capitalisation sur le marché du logement (Dowding et al., 1994), ou avec autre mesure qui peut signaler la forte présence de Tiebout des incitations. Cette approche est constamment confrontée à la difficulté de distinguer le statistique pur du type de tri qui est supposé être motivé par une concurrence de type Tiebout.

Une deuxième approche, en économétrie, commence par se demander si les les modèles de dépenses en biens publics locaux satisfont aux conditions prises par un modèle de Tiebout d'équilibre général (Epple et al., 1999). La condition forte de satisfaire le nécessaire et les conditions d'équilibre déterministes ont placé les chercheurs le long de la chemin de la construction de modèles d'équilibre calculables

de croissance complexité mathématique (Nechyba, 2020). L'objectif central de ce type d'approche est de prédire la fragmentation juridictionnelle et stratification, sans sacrifier le choix et l'allocation néoclassiques Efficacité. Le problème avec les modèles statiques génériques est qu'ils Caractéristiques des conditions trop strictes pour obtenir un équilibre distribution de la variable de résultat, et qui sont facilement échangeable contre d'autres. Cette échangeabilité des mathématiques les conditions d'équilibre donnent les modèles très idiosyncratiques, et spécifique au-delà de ce qui est nécessaire pour fournir une statistique significative description du tri et de la compétition de Tiebout. Pour les récents et examen approfondi des développements récents dans la littérature Tiebout voir (Epple and Nechyba, 2004b; Jha, 2020; Nechyba, 2020). Pour des approches économétriques pour étudier et tester l'hypothèse de Tiebout, voir Dowding et al. (1994) et (Epple et al., 2010, 2001, 1978).

Dans l'approche d'équilibre statistique proposée ici et dans Melo (2021), qui est une extension des biens publics locaux mise en place du modèle développé dans Scharfenaker and Foley (2017) et dans Scharfenaker and Foley (2021a), une approche parcimonieuse modèle du processus de quasi-marché de Tiebout est utilisé pour distinguer entre les forces concurrentielles et de tri en jeu dans le public américain marché de l'éducation.