SCIENCE IN HIGH DIMENSIONS: MULTIPARAMETER MODELS AND BIG DATA

A Dissertation

Presented to the Faculty of the Graduate School

of Cornell University

in Partial Fulfillment of the Requirements for the Degree of

Doctor of Philosophy

by

Ricky Chachra

January 2014

SCIENCE IN HIGH DIMENSIONS: MULTIPARAMETER MODELS AND BIG DATA

Ricky Chachra, Ph.D.

Cornell University 2014

Complex multiparameter models such as in climate science, economics, systems biology, materials science, neural networks and machine learning have a large-dimensional space of undetermined parameters as well as a large-dimensional space of predicted data. These high-dimensional spaces of inputs and outputs pose many challenges. Recent work with a diversity of nonlinear predictive models, microscopic models in physics, and analysis of large datasets, has led to important insights. In particular, it was shown that nonlinear fits to data in a variety of multiparameter models largely rely on only a few stiff directions in parameter space. Chapter 2 explores a qualitative basis for this compression of parameter space using a model nonlinear system with two time scales. A systematic separation of scales is shown to correspond to an increasing insensitivity of parameter space directions that only affect the fast dynamics. Chapter 3 shows with the help of microscopic physics models that emergent theories in physics also rely on a sloppy compression of the parameter space where macroscopically relevant variables form the stiff directions. Lastly, in chapter 4, we will learn that the data space of historical daily stock returns of US public companies has an emergent simplex structure that makes it amenable to a low-dimensional representation. This leads to insights into the performance of various business sectors, the decomposition of firms into emergent sectors, and the evolution of firm characteristics in time.

BIOGRAPHICAL SKETCH

Ricky Chachra attended a high school that for the final two years focused heavily on physical sciences and mathematics. When he began college in 2004, the quantitative orientation led him toward computational sciences. Mostly supported by scholarships, he graduated in 2 years and 9 months, and earned a Bachelor's of Science *magna cum laude* in Applied Mathematics and Statistics from Stony Brook University. Throughout his undergraduate years, Ricky was engaged in computational biology research projects that eventually led him to graduate education in the field of Biophysics at Cornell University's Weill Graduate School of Medical Sciences. He then transferred to Cornell University's main campus in 2008 and joined the graduate field of Theoretical and Applied Mechanics. As he continued to develop his academic interests, he found himself excited by the problems in the relatively new field of data science.

Following his graduation at Cornell, Ricky will join the Business Performance Services team of IBM Corporation at its global headquarters in Armonk, NY.

To

Albert J. Libchaber

ACKNOWLEDGMENTS

An immeasurable amount of gratitude is owed to my advisor, James P. Sethna. Only a superhuman could have nurtured my intellectual development starting from my humble and unconventional beginnings as I launched into graduate school. And I was indeed lucky to have a superhuman advisor—an endless fountain of great ideas, excitement, optimism and patience for doing the most amazing work in science.

Deep gratitude is also owed to the mentors who have guided me in years prior: Albert J. Libchaber, Robert C. Rizzo and Herbert J. Bernstein. Thank you Paul H. Ginsparg, John M. Guckenheimer and Stefanos Papanikolaou for the insights and interest in my work that proved critical to the success of my projects.

Thank you to the best collaborators imaginable: Alexander A. Alemi, Benjamin B. Machta and Mark K. Transtrum, as well as to the outstanding colleagues Ashivni Shekhawat, Yan-Jiun Chen, Matthew Bierbaum, Woosong Choi, Lorien Hayden and Colin Clement.

A special thanks is due to my special committee for incisive suggestions that greatly improved this dissertation: James P. Sethna, John M. Guckenheimer and Steven H. Strogatz. And to the National Science Foundation for supporting the projects described herein.

And finally, a salute to the countless sacrifices made by my loving mother, the enduring friendship of my sister, and the unconditional affection I got from the love of my life as I plowed through the years of higher education.

TABLE OF CONTENTS

Biographical Sketch . . . iii
Dedication . . . iv
Acknowledgments . . . v
Table of Contents . . . vi
List of Figures . . . viii
List of Tables . . . xiv

CHAPTER PAGE

1 Chapter 1: Introduction 1
1.1 Phenomenological or predictive models . . . 2
1.1.1 The van der Pol Oscillator Example . . . 4
1.2 Emergent theories . . . 5
1.2.1 The Ising Model and Diffusion Equation . . . 7
1.3 Big Data or e-Science . . . 8
1.3.1 Stock Market Returns . . . 10

2 Chapter 2: Structural Susceptibility and Separation of Time Scales in the van der Pol Oscillator 12
2.1 Abstract . . . 12
2.2 Introduction . . . 13
2.3 Multiple Time Scale Dynamics . . . 14
2.4 Sloppiness in Nonlinear Fits . . . 16
2.5 Susceptibility of van der Pol system . . . 18
2.5.1 Eigenvalues and Eigenvectors . . . 21
2.6 Discussion . . . 23
2.7 Acknowledgments . . . 25

3 Chapter 3: Parameter Space Compression Yields Emergent Theories of Physics 27
3.1 Abstract . . . 27
3.2 Introduction . . . 28
3.3 The Fisher Information . . . 29
3.4 Discrete Diffusion . . . 30
3.5 Ising Model . . . 32
3.6 Discussion . . . 35
3.7 Acknowledgements . . . 37
3.8 Supplementary Information . . . 37
3.8.1 Fisher Information for Discrete Diffusive Hopping . . . 37
3.8.2 Measuring the Ising metric . . . 40
3.8.3 Scaling analysis of the Ising eigenvalue spectrum . . . 42
3.8.4 Measuring the Ising metric after coarsening . . . 44

3.8.5 Eigenvalue spectrum after coarse-graining . . . 47
3.8.6 Simulation details . . . 51

4 Chapter 4: Canonical Sectors and Evolution of Firms in the US Stock Markets 54
4.1 Abstract . . . 54
4.2 Introduction . . . 55
4.3 Canonical Sectors and Price Returns . . . 57
4.4 Constituent Firms in Canonical Sectors . . . 60
4.5 Sector Decomposition of All Firms . . . 61
4.6 Evolution of Sector Weights . . . 63
4.7 Discussion . . . 64
4.8 Acknowledgements . . . 66
4.9 Supplementary Information . . . 66
4.9.1 Dataset Particulars . . . 66
4.9.2 Returns Factorization and Sector Decomposition . . . 67
4.9.3 Calculations and Convergence . . . 69
4.9.4 Dimensionality of Space of Price Returns . . . 70
4.9.5 Low-Dimensional Projections of Price Returns . . . 72
4.9.6 Proportion of Variation Explained (PVE) . . . 73
4.9.7 Determining the Number n of Canonical Sectors . . . 74
4.9.8 Canonical Sector Indices . . . 76

APPENDIX PAGE

A Miscellaneous Results and Information 83
A.1 Fisher Information Matrix as a metric on parameter space . . . 83
A.1.1 The metric of a Gaussian model . . . 86
A.1.2 The metric of a Statistical Mechanical Model . . . 87
A.2 Derivation of FIM eigenvalues . . . 88
A.3 Derivation of α and β of the Linear Regression Equation . . . 90
A.3.1 A New Elementary Method . . . 90
A.3.2 The Standard Method . . . 91
A.4 SVD of Centered Data . . . 93
A.5 The Major Sector Classification Systems . . . 96

References 99

List of Figures

Figure Page

1.1 **Hessian eigenvalues for seventeen systems biology models taken from [1].** The distribution reaches extremely small values indicating that the cost is significantly insensitive to many directions in parameter space. These eigenvalues are squares of the singular values of the Jacobian $J_{i\alpha} = \partial f_i / \partial \theta_\alpha$. 2

1.2 **Schematic of the mapping between parameter and data space.** The sloppy directions in parameter space map to small regions in behavior space as the predictions remain largely unchanged. The stiff directions on the other hand map to long thin hyper-ribbons due to the sensitivity of these directions. . . . 3

1.3 **Hierarchies in physics.** Left: Theories in high-energy physics form a nested hierarchy. Each theory is derived from a more fundamental, unified theory, describing behavior at higher energy scales (demanding bigger particle accelerators). The unified theory explains key parameters in the derived theory: quantum chromodynamics and the electroweak theory tell you the masses of the nuclei and electron. Right: Theories in condensed-matter physics form a nested hierarchy. Each theory emerges from a more microscopic and complicated theory 'below' it, providing a simpler and more beautiful description. The emergent theory compresses the microscopic details into a few governing parameters that efficiently describe the behavior at longer distances, longer times, and lower temperatures. 6

1.4 **Eigenvalues of a digit-reconstruction neural network map.** Top: A sample of five scanned handwritten digits from the MNIST database [2]. Middle: Reconstruction of the images in the top row using a trained four-layer auto-encoder neural network. Bottom: Eigenvalues of the reconstruction map show a sloppy spectrum (every 10th eigenvalue is shown). This figure was generated with data kindly provided by Hayden *et al.* [3]. 9

2.1 (a, d) Eigenvalues of the Hessian matrix of the cost of fitting at $\mu = 1$ (left) and $\mu = 100$ (right) for a multi-parameterized van der Pol discussed in section 2.5. (b, c top row) One period of time series $x(\tau)$ (dotted line), and $y(\tau)$ (solid line), for $0 < \tau < 1$, are shown for $\mu = 1$ and $\mu = 100$ as function of time along with schematic error bars for the data-fitting of the trajectory of variable y. (b, c bottom row) The orbit in xy plane (solid line) and the critical manifold (dashed line) As $\mu \to \infty$, the orbit collapses onto the critical manifold with the trajectory spending most of its time on the slow manifold and vanishingly short on the jumps. 16

2.2 Eigenvalues of Hessian matrix are shown here as a function of μ. The range $1 \leq \mu \leq 100$ corresponds to a ratio of time scales $1 \leq \epsilon \leq 10000$. The five largest eigenvalues (solid lines) correspond to stiff directions in the parameter space: these directions perturb the slow manifold. The remainder (dashed lines) affect the transient part of the trajectory which becomes smaller with an increasing separation of time scales and hence these directions are decreasingly relevant. 20
2.3 Hessian eigenvectors are shown for $\mu = 1,\ 10,$ and 100. Each colored small square shows the magnitude of an eigenvector component (the scale bar shown on the right). Eigenvectors for each μ are sorted so that the stiffer ones appear on the left; individual components are sorted so that "slow parameters" appear on the top. Note that with increasing μ, the stiff and sloppy eigenvectors separate by parameters: The stiff eigenvectors only have projections along the slow parameters which perturb the slow manifold, whereas the sloppy directions have projections along the fast parameters which mainly perturb the jumps. 21
2.4 Top three rows: Eigenpredictions δy_k for $k = 0,\ 3,\ 6,\ 9,\ 12$ at $\mu = 1,\ 10\ \&\ 100$ are shown in solid red lines for stiff modes and dashed green for sloppy modes. These curves show the response of perturbations if the parameters are changed infinitesimally along the Hessian eigenvectors: A parameter change of norm ϵ along eigendirection n will change the trajectories by $\lambda_n \epsilon$ times the eigenpredicton. Dotted gray lines show unperturbed van der Pol solution for comparison (y scale on the right hand side). As the time scales separate, the amplitudes of the sloppiest eigenpredictions increase (roughly in proportion to μ) getting increasingly concentrated at the jumps. Bottom row shows the eigencycles for $\mu = 100$ in solid red lines and green dashed lines corresponding to the perturbations in row 3 (i.e. the new limit cycle for a perturbation of strength $\epsilon \sim 1/\lambda_n$. These curves show how the van der Pol orbit changes with perturbations along the Hessian eigenvectors. Both the stiff and the sloppy modes change the orbit at the jumps (occurring at the extrema in the dashed lines); the stiff modes also change behavior at the slow manifold, whereas the sloppy modes only affect the jumps. 26

3.1 **Normalized eigenvalues of the Fisher Information Matrix (FIM) of various models.** The diffusion and Ising models are explored here. A radioactive decay model and a neural network are taken from [4]. The systems biology model is a differential equation model of a MAP kinase cascade taken from [5] and the adjoining band marked as "Random" shows a typical eigenvalue spread from a Wishart random matrix of the same size. The 'Relaxation oscillation' model is a modified Van der Pol system taken from [6]. Eigenvalues of the genetic network describing 'Circadian rhythm' model [7] are calculated in [1]. 'Variational wave function' eigenvalues are taken from Quantum Monte Carlo simulations as Jastrow parameters are varied [8]. 'Particle accelerator' is a model of beam shape simulated using the Tool for Accelerator Optics [9]. In all models, the eigenvalues of the FIM are roughly geometrically distributed, with each successive direction significantly less important for system behavior (only the first 8 decades are shown). This means that inferring the parameter combination whose eigenvalue is smallest shown would require $\sim 10^8$ times more data than the stiffest parameter combination. Conversely, the least important parameter combination is $\sqrt{10^8}$ times less important for understanding system behavior. 31

3.2 **FIM eigenvalues of a model of stochastic motion on a 1-D lattice.** The seven parameters describe probabilities of transitioning to nearby sites (bottom inset). Observations are taken after a given number of time steps for the case where all parameters take the value $a^\mu = 1/7$. Top row shows the resulting densities plotted at times $\tau = 1, 3, 5, 7$. Bottom plot shows the eigenvalues of the FIM versus number of steps. After a single time step, the FIM is the identity, but as time progresses, the spectrum of the FIM spreads over many orders of magnitude. The first eigenvector measures deviations in the net particle creation rate R from 0, the second measures a net drift V in the density, and the third corresponds to parameter combinations that change the diffusion constant D. Further eigenvectors describe parameter combinations that do not affect these macroscopic parameters, but instead measure the skew (green), kurtosis (purple), and higher moments of the resulting density (orange and brown). 33

3.3 **FIM eigenvalues of an Ising model of ferromagnetism.** See text for definition; 13 parameters describe nearest and nearby neighbor couplings (bottom inset) and a magnetic field. Observables are spin configurations of all spins on a sub-lattice (dark sites in the insets of the top panel). The top panel shows one particular spin configuration generated by the model, suitably blurred for step > 0 to the average spin conditioned on the observed sub-lattice values. Some information about the configuration, such as the typical size of fluctuations, is preserved under this procedure, whereas other information like the nearest neighbor correlation amplitude is lost. The two largest eigenvalues, whose eigenvectors measure reduced temperature t and the applied field h do not decay substantially under coarsening. Further FIM eigenvalues shrink by a factor of $\sqrt{2}^{-2-y_i}$, where y_i is the i^{th} RG exponent (section 3.8.3). This shrinkage quantifies the information lost in each coarsening step. 36

3.4 **Eigenvalues of the FIM versus** J/J_c. The enlarged 13 parameter Ising model of size $L = 64$ is described in the text. Magnetic field h is taken to be zero. Two eigenvalues become large near the critical point, each diverging with characteristic exponents describing the divergence of the susceptibility and specific heat respectively. The other eigenvalues vary smoothly as the critical point is crossed. Furthermore they take a characteristic scale determined by the system size and are not widely distributed in log. (In the phase separated region, $\beta J > \beta J_c$ we use the connected correlation function in calculating g_{00}. This corresponds to calculating eigenvalues in 'infinitesimal field'. It allows calculation of the FIM in the phase but arbitrarily close to the phase boundary at which there is a net spontaneous magnetization. Without this the FIM would have one spuriously large eigenvalue, quantifying the large symmetry breaking affect of an arbitrarily small applied field.) 42

4.1 **Low-dimensional projection of the stock price returns data.** Stock price returns are projected onto a plane spanned by two stiff vectors from the SVD of the emergent simplex corners as described in section 4.9.5. Each colored circle corresponds to one of the 705 stocks in the dataset used in the analysis. Colors denote the sectors assigned to companies by Scottrade [11] and the scheme is shown in fig. 4.9. The grey corners of the simplex correspond to sector-defining prototype stocks, whereas all other circles are given by a suitably weighted sum of these grey corners. Projections along other singular vectors are shown in fig. 4.6. 56

4.2 **Emergent sector time series.** Annualized cumulative log price returns of the eight emergent sectors are shown. The time series capture all important features affecting different sectors: dot-com bubble (c. 2000), the energy and financial crises of 2008. Precise definitions is given in equation 4.3; other measures of sector dynamics are in fig. 4.8. 59

4.3 **Canonical sector decomposition of stocks of selected companies**. A complete set of pictures for all 705 stocks is provided on the companion website [10]. Color scheme shown on the right are used in figures throughout the paper except where noted. 62
4.4 **Evolving sector participation weights.** Results from the sector decomposition made with rolling two-year Gaussian windows are shown for selected stocks. A complete set of 705 pictures is provided on the companion website [10]. Color scheme is as in fig. 4.3 64
4.5 **Normalized distribution of singular values.** Filled blue histogram corresponds to distribution of singular values of returns from the dataset R_{ts}—one notices a clear separation of the hump-shaped bulk of singular values ascribed to random Gaussian noise, and about 20 stiff singular values (the largest singular value $\sim$952, corresponding to the *market mode* is not shown). Pink line histogram outline shows the distribution of singular values of a matrix of the same shape as R but containing purely random Gaussian entries. 71
4.6 **Low-dimensional projections of stock returns data.** Each colored circle represents a stock in our dataset is colored according to listed sectors scheme in fig. 4.9 according to sectors assigned by Scottrade [11]. The first row is repeated from fig. 4.1. Black circles represent are the archetypes found with our analysis. The $(i, j)^{th}$ figure in the grid is a plane spanned by singular vectors i and $j + 1$ (rows of $X^T R$) from the calculations described in section 4.9.5. Projections after the factorization are shown in fig. 4.7. 77
4.7 **Cross-sections along eigenplanes of the factorized returns.** Each colored circle represents a stock in our dataset is colored according to scheme in fig. 4.3 based on the primary sector association found after calculations described in this paper. Black circles represent the archetypes found with our analysis. The $(i, j)^{th}$ figure in the grid is a plane spanned by singular vectors i and $j + 1$ (rows of MN^T) from the calculations described in section 4.9.5. Projections of raw data (before the factorization) are shown in fig. 4.6. 78
4.8 **Canonical sector time series.** Top row: normalized log returns (columns of E_{tf}), middle row: cumulative log returns (same as fig. 4.2 as defined in equation 4.3, and bottom row: unweighted price index of canonical sectors (eq. 4.5). 79
4.9 **Weight distribution in canonical sectors.** Each of the eight subplots shows the constituent participation weights of all 705 companies in an canonical sector (rows of W_{fs}). Stocks are colored by listed sectors as shown at the bottom. Listed sector information was obtained from [11]. Y-axis range is from 0 to 1. 80

4.10 **Singular vectors V_{fs}^T of SVD of returns R_{ts}.** The orthonormal right singular vectors (rows of V_{fs}^T) of SVD of R_{ts} are equivalent to the eigenvectors of the stock-stock correlation matrix $\xi_{ss'} \sim R^T R$. Eight of these stiffest eigenvectors including the *market mode* are shown in rows of two at a time. Each has 705 components corresponding to stocks in an the dataset. The *market mode* with all components in the same direction describes overall fluctuations in the market; it was excluded from the analysis described in the paper. Previous work [12] has suggested that each eigenvector of the stock-stock correlation matrix describes a listed sector, however as seen above, a more correct interpretation is that each eigenvector is a mixture of listed sectors with opposite signs in components. For example, the stiffest direction (after market mode) has positive components in real estate and utility, but negative in tech. Less stiff eigenvectors (including the last one shown here), do not contain sector-relevant information. Stocks are colored by listed sectors as shown at the bottom. Listed sector information was obtained from [11]. Y-axis range is from –0.5 to 0.3. 81

4.11 **Canonical Sector Constituents (shown as columns of the C_{sf}).** C_{sf} represents a weighted combination stocks that defines of the canonical sector each of which has a time series represented by E_{tf} that is given by $E_{tf} = R_{ts}C_{sf}$. The eight subplots show the constituent participation component of stocks in each canonical sector f. Canonical sectors are labeled on the plot; their names were chosen according to the listed sectors of firms that comprise them. Noteworthy features seen above include the co-association of listed sectors: basic, capital, transport and part of cyclicals into *industrial goods*. Similarly, healthcare and non-cyclicals are coupled together in what we call *non-cyclicals*. Canonical *retail* goes primarily with listed retail and cyclicals. Stocks are colored by listed sectors as shown at the bottom. Listed sector information was obtained from [11]. Y-axis range is from 0 to 0.05. 82

List of Tables

Table Page

4.1 **Canonical sectors and major business lines of primary constituent firms.** Examples provided are firms that are strongly associated to these sectors. A full list is available on companion website [10]. 60

4.2 **Listed sectors and number of companies dataset analyzed.** Tickers for each company were obtained from [11]. 67

A.1 **The S&P/MSCI Barra Global Industry Classification Standard (GICS) [13].** 97

A.2 **The Thomson Reuters Business Classification (TRBC) [14].** 97

A.3 **The Dow Jones/FTSE Industry Classification Benchmark (ICB) [15].** . . 98

CHAPTER 1

CHAPTER 1: INTRODUCTION

There are three key ways that are employed in order to mathematically distill the essence of a phenomenon, and each one is tied to a different chapter in this dissertation:

1. phenomenological or predictive modeling,
2. application or development of a theory, and
3. statistical analysis of large sets of observations or experimental data.

Regardless of the means, the end goal is to attain an understanding that is comprehensive and in accord with other known facts. The success of science and engineering bears testimony to the efficacy of these methods, but what makes these ways of making progress possible? How is it that the complicated nature yields to often concise and comprehensible mathematical descriptions?

A central theme in this dissertation is the following: Typical nonlinear, multi-parameter models from a variety of areas of science, as well as large-scale statistical analysis algorithms are successful because the datasets they describe often effectively have surprisingly low dimensionality. The development that follows in this dissertation will support the claim that only a few directions in parameter space are sufficient to account for most of the variation seen in the general observed data, while other directions are often insignificant.

Each of the following sections of this chapter incrementally adds to a big-picture context in

which this thesis is set. Each section has a subsection with a high-level overview of a different chapter that will follow later.

1.1 PHENOMENOLOGICAL OR PREDICTIVE MODELS

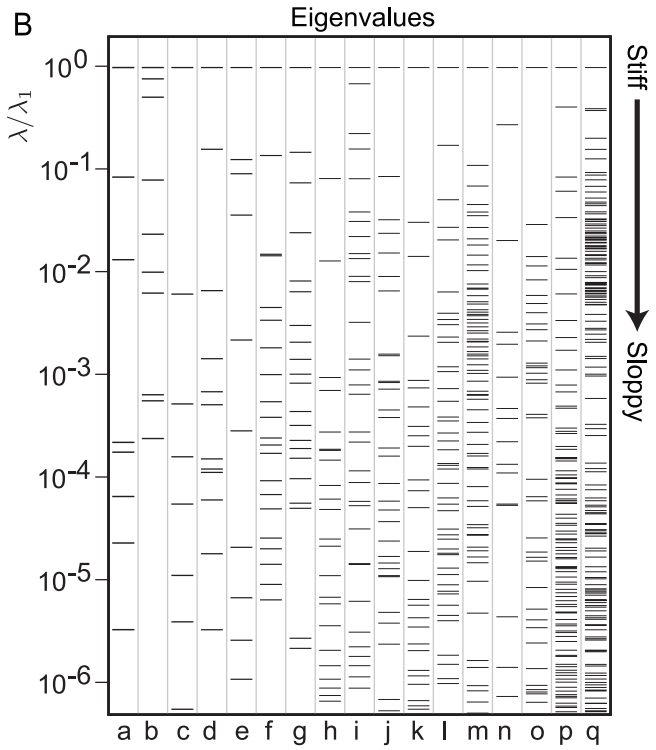

Figure 1.1: Hessian eigenvalues for seventeen systems biology models taken from [1]. The distribution reaches extremely small values indicating that the cost is significantly insensitive to many directions in parameter space. These eigenvalues are squares of the singular values of the Jacobian $J_{i\alpha} = \partial f_i / \partial \theta_\alpha$.

Predictive models in fields such as climate science, ecology, economics, neuroscience, systems biology, etc. are often complicated due to multiple nonlinear interactions among many constituent variables. For example, a typical systems biology model will include many coupled differential equations each describing an evolution law for gene concentrations $\vec{y}$ as a function

of time. Experimental data d_i is measured at times t_i with an error of measurement σ_i to which the model $\vec{y}(t) = f(\vec{\theta}, t)$ is fit with parameters $\vec{\theta}$. Seventeen models of these kinds have been previously analyzed [1] by Gutenkunst and colleagues in the Sethna group; in each case, the Hessian $\mathcal{H}_{\mu\nu} = \partial_\mu \partial_\nu C$ of the least-squares cost function $C = \frac{1}{2}\sum_i (f(\vec{\theta}, t_i) - d_i)^2/\sigma_i^2$, evaluated at the best fit has eigenvalues that are roughly geometrically distributed over many decades (fig. 1.1). The group also found the same "sloppy" character for many other models [4, 8] such as radioactive decay, variational wave function of quantum mechanics, neural networks, and even general polynomial fits to data.

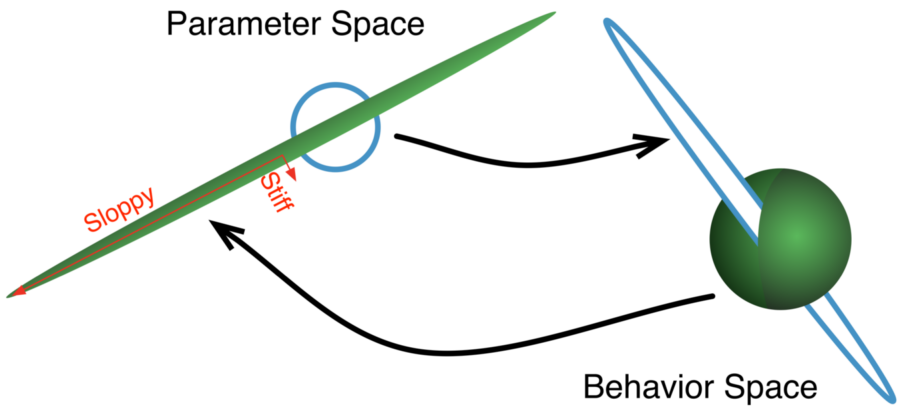

Figure 1.2: Schematic of the mapping between parameter and data space. The sloppy directions in parameter space map to small regions in behavior space as the predictions remain largely unchanged. The stiff directions on the other hand map to long thin hyper-ribbons due to the sensitivity of these directions.

In the case of predictive models, the ubiquity of sloppiness has been traced down [4] to hyper-ribbon structures (fig. 1.2) of the data (behavior) space manifold associated with these models. These hyper-ribbons have widths that are also geometrically spaced, and their curvatures and boundaries have special properties that using interpolation theorems were shown to be typical features of multi-parameter models. The observation that the manifold widths and Hessian eigenvalues were related suggested that sloppiness is not just a consequence of the chosen parameterization of the model: while the eigenvalues change

depending on model parameterization, the manifold widths are intrinsic to the model manifold. In addition, manifold boundaries correspond to physically relevant limits such as the extreme values that parameters can meaningfully have. Recent work utilizing this geometry has led to a scheme for systematically coarsening nonlinear models [16]. The success of this scheme and its possibly general applicability is fundamental evidence of the intrinsic low dimensionality of models with otherwise many parameters.

A different approach to understand the origins of sloppiness is a more qualitative one. For the variety of behaviors seen in nonlinear systems, one can systematically perturb the dynamics to target the behavior of interest, and probe the consequences on the resulting Hessian spectrum. The second-order van der Pol equation is a particularly convenient case-in-point where a single parameter alters the inherent separation of scales in the dynamics.

1.1.1 The van der Pol Oscillator Example

Chapter 2 of this dissertation is based on the analysis of the van der Pol oscillator. The van der Pol equation $x'' - \mu(1 - x^2)x' + x = 0$ has been extensively deployed in physical and biological sciences as a predictive model of relaxation oscillations in electrical circuits with vacuum tubes [17, 18], action potentials of neurons [19, 20], and plate dynamics in a seismic fault [21]. It is indeed a prototypical example of a system with two time scales in its dynamics which have a ratio long/short $= \mu$. We will examine the *structural susceptibility*[1] of van der Pol dynamics to perturbations in the dynamics as time scales are separated.

[1] The notion of susceptibility in dynamical systems is new but is inspired by its use in physics where it is used to measure the effect on free energy f of infinitesimal perturbation of fields θ_μ. Susceptibilities $g_{\mu\nu} = \partial^2_{\mu\nu} f$. For a dynamical system, fitting cost is like the free energy of physics and therefore, $\partial^2_{\mu\nu} C$ is what we refer to as structural susceptibility (Hessian of cost w.r.t. parameters).

An extension of the van der Pol equation that incorporates additional parameters selectively affecting the slow and fast parts of the dynamics is developed and utilized. As the scales separate, two classes of parameters become distinct as stiff and sloppy directions in the dynamics. The former only affect the slow manifold of the dynamics whereas the latter affect the jump in the dynamics. Sloppiness is seen to be enhanced as the scales separate, suggesting that the presence of multiple time scales in the dynamics is one of the reasons why nonlinear systems are sloppy [6].

1.2 EMERGENT THEORIES

Despite complexity on microscopic scales, physics theories have emergent behavior that is largely independent of the complicated dynamics of the constituent atoms or quantum fields. For example, our very complete phenomenological description of the subatomic world known as 'The Standard Model' is hardly informed by the more fundamental theory for which string theory is a candidate. At another level, the theory of superconductivity makes absolutely no use of The Standard Model even though it would presumably give a sufficient description of a superconductor's constituents. This kind of hierarchy is pervasive across physics; a schematic representation of this in condensed matter and high energy physics is shown in figure 1.3.

The continuum limits and renormalization group (RG) methods of physics show—at least in principle—how a micro-level physics theory can be coarsened to yield a universal theory applicable at longer length or time scales. RG is a procedure for systematically coarsening and rescaling a system to reveal the underlying, dominant interactions; with continuum limits, the microscopic complexity relevant only at short time and/or length scale is dissolved and a simpler description emerges.

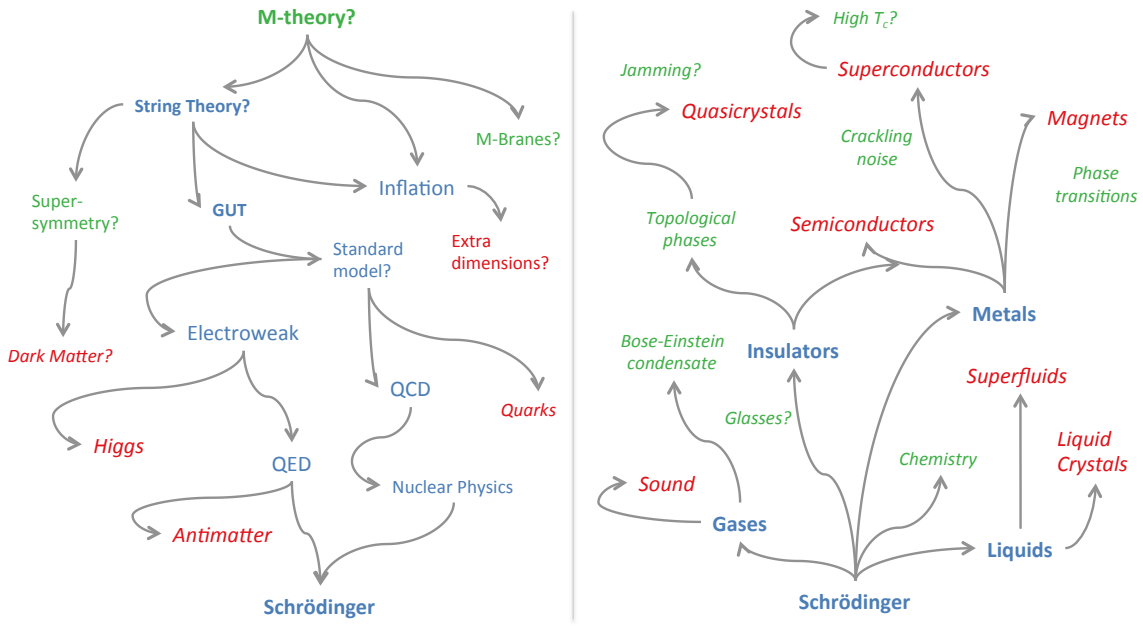

Figure 1.3: Hierarchies in physics. Left: Theories in high-energy physics form a nested hierarchy. Each theory is derived from a more fundamental, unified theory, describing behavior at higher energy scales (demanding bigger particle accelerators). The unified theory explains key parameters in the derived theory: quantum chromodynamics and the electroweak theory tell you the masses of the nuclei and electron. Right: Theories in condensed-matter physics form a nested hierarchy. Each theory emerges from a more microscopic and complicated theory 'below' it, providing a simpler and more beautiful description. The emergent theory compresses the microscopic details into a few governing parameters that efficiently describe the behavior at longer distances, longer times, and lower temperatures.

Many of the practical implications of RG and continuum limits are identical to those of the sloppiness discussed earlier. Models show weak dependence of macroscopic observables (defined at long length and time scales) on microscopic details. They thus have a smaller effective model dimensionality than their microscopic parameter space. This is shown formally with analysis using microscopic versions of two physics models, the Ising model and the diffusion equation, where the emergent behavior is already well-understood using RG and continuum limits respectively.

1.2.1 The Ising Model and Diffusion Equation

Chapter 3 of dissertation shows that the emergent theories of physics implicitly exploit the same hierarchical hyper-ribbon structure in their model manifolds that was seen earlier in sloppy models. This structure is seen to develop as microscopic observables from two model systems are coarsened in analyses for both. In the case of the Ising model of ferromagnetism, the Hamiltonian $\mathcal{H}(\sigma) = -J \sum_{\{i,j\}} \sigma_i \sigma_j - \ h \sum_i \sigma_i$ has only two-parameters, a coupling constant J and a magnetic field h, yet it describes a vast variety of systems at their self-similar critical points. Similarly, the diffusion equation governing many stochastic processes, $u_t = Du_{xx} + Vu_x + R$, has only three parameters, a diffusion constant D, an average drift V and a particle creation rate R. In both cases, the microscopic complexity is subsumed in these few parameters.

With a multi-parameter extension of the Ising model, and a hopping model of diffusion, chapter 3 provides an information theory-based approach to model condensation as seen using RG and continuum limits. In the case of Ising, J and h emerge to be the stiff parameters, whereas for diffusion, combinations corresponding to R, V and D are stiff. These physics models are not sloppy when observations are measuring microscopic level details or fluctuations, however when observations are coarse (on a macro scale) but parameters are microscopic, it is then that a low-dimensional representation is manifest in the analysis.

The similarity of sloppy hierarchies in emergent physics theories with other areas of science, clarifies the connection between the universality that arises in systems under the purview of the RG and the success of modeling more generally. Many models work not because they are microscopically correct, but because of a hyper-ribbon structure. They can be quantitatively correct even if their microscopic details are incomplete, or even wrong. Just as we can understand a superconductor without appealing to the details of its constituent

atoms, we can understand the behavior of a signal transduction cascade without knowing all of the interaction partners of its components [22].

1.3 Big Data or e-Science

Recent years have witnessed an increasing focus on the use of data and advanced analytics by businesses and researchers alike [23]. The algorithms of machine learning or artificial intelligence are allowing for statistical analysis and solutions to problems in ways that are not accessible by more traditional means employed in science. Apple Inc.'s speech recognition system (Siri) [24], Google Inc.'s self-driving car [25], IBM Corp.'s natural language processing computer (Watson) [26], and Netflix Inc.'s movie-rating prediction method [27] are among well-known services and products that use these algorithms and leverage massive scale datasets for statistical insights. As discussed next, there is compelling and increasing evidence that these algorithms either implicitly or explicitly exploit emergent low-dimensional structures in the data space.

Machine learning algorithms come under two broad buckets [28]:

- *Supervised learning* algorithms are trained with labeled input, and perform classification or regression on unlabeled items. A majority of the neural network algorithms are supervised; an example is shown in figure 1.4.

- *Unsupervised learning* algorithms such as clustering, hidden Markov models, manifold learning, matrix factorization, etc., are applied on unlabeled data to discover structure or relationships in data. A typical matrix factorization scheme $R_{ts} = E_{tf}W_{fs}$, for example, will reduce the dimensionality of the original $t \times s$ matrix as a product of two matrices of sizes $t \times f$ and $f \times s$. Depending on the application, dataset and constraints, f is

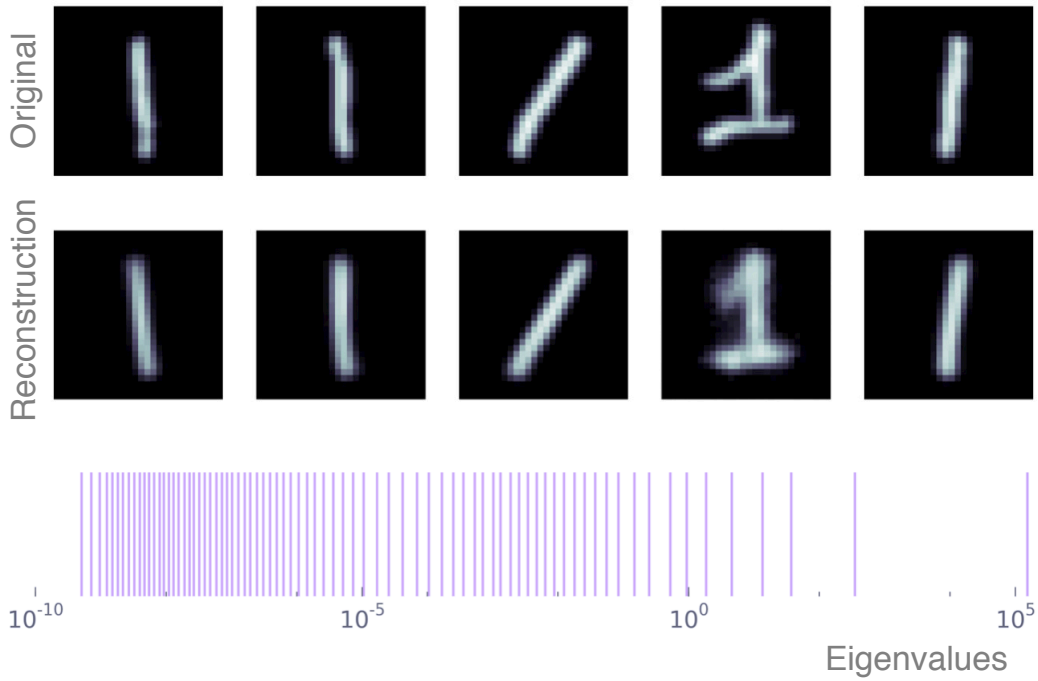

Figure 1.4: Eigenvalues of a digit-reconstruction neural network map. Top: A sample of five scanned handwritten digits from the MNIST database [2]. Middle: Reconstruction of the images in the top row using a trained four-layer auto-encoder neural network. Bottom: Eigenvalues of the reconstruction map show a sloppy spectrum (every 10th eigenvalue is shown). This figure was generated with data kindly provided by Hayden *et al.* [3].

generally user-specified so that $(t \times f + f \times s) \ll t \times s$. A movie-rating prediction can be based on this scheme when applied to a sparse matrix of known ratings R_{um} for users u and movies m. R can be factorized in dense E_{uf} and dense W_{fm}, so that $R \sim EW$. The matrix $EW = \tilde{R}$ can then be interpreted as a rating prediction because (a). it is dense, (b). R and $\tilde{R}$ agree where former is non-zero.

In general, one has very little guidance about which machine learning method or algorithm is best suited for an arbitrary problem. As we shall see with an example of stock

market returns, the choice of method or algorithm can be made in a principled manner by understanding the inherent structure in the dataset of interest.

1.3.1 Stock Market Returns

Chapter 4 presents an analysis of daily price returns of 705 US public companies' stock prices for a 20-year period spanning 1993-2013 that consists of a total of 5000 time points (one for each of the $\sim$250 business days/year). The matrix of returns R_{ts} at times t for stocks s thus has more than 3.5 million entries. A large dense matrix can be rank-reduced in a number of ways, and the common practice is to factorize using the singular value decomposition (SVD). We will see in chapter 4 that the space of stock price returns has an emergent, low-dimensional hyper-tetrahedral (simplex) structure which inspires a more meaningfully constrained factorization.

The algorithm constructs a factor matrix of f time series, E_{tf} in a purely unsupervised manner, and due to the geometry of the low-dimensional manifold considered here, each column of E (*i.e.* every basis vector) corresponds closely to a business sector (group of companies in closely related business lines). Via constraints, E is itself represented as a combination of "archetype stocks", of companies that can be considered to define a sector. These emergent sectors, which we call *canonical sectors* are comprised of a combination of (automatically) selected stocks so that $E_{tf} = R_{ts}C_{sf}$ with a relatively sparse C.

Using the property that every interior point of a convex set is a weighted-sum of the corner points, the matrix W_{fs} is constrained so that in $R_{ts} = E_{ts}W_{fs}$ each stock's return is a weighted sum of returns from the canonical sectors. For example, returns of stock of IBM, which is conventionally listed as a tech firm in most financial indices, are most accurately

described as a weighted sum of returns from the canonical tech (78%), canonical non-cyclical (21%) and canonical utility (1%) sectors.

CHAPTER 2

CHAPTER 2: STRUCTURAL SUSCEPTIBILITY AND SEPARATION OF TIME SCALES IN THE VAN DER POL OSCILLATOR

2.1 ABSTRACT

This chapter describes an extension of the van der Pol oscillator as an example of a system with multiple time scales to study the *susceptibility* of its trajectory to polynomial perturbations in the dynamics. A striking feature of many nonlinear, multi-parameter models is an apparently inherent insensitivity to large magnitude variations in certain linear combinations of parameters. This phenomenon of "sloppiness" is quantified by calculating the eigenvalues of the Hessian matrix of the least-squares cost function. These typically span many orders of magnitude. The van der Pol system is no exception: Perturbations in its dynamics show that most directions in parameter space weakly affect the limit cycle, whereas only a few directions are stiff. With this study we show that separating the time scales in the van der Pol system leads to a further separation of eigenvalues. Parameter combinations which perturb the slow manifold are stiffer and those which solely affect the jumps in the dynamics are sloppier.

[0]This chapter describes previously published work [6]: R. Chachra, M. K. Transtrum, J. P. Sethna, *Structural Susceptibility and Separation of Time Scales in the van der Pol Oscillator*. Phys. Rev. E 86 (Aug, 2012) 026712. The present author chose the model system, performed the computational analysis, analyzed the results and wrote the manuscript.

2.2 INTRODUCTION

We will analyze the sensitivity of a multiple time scales dynamical system to perturbative changes in its evolution laws. Rather than utilizing the traditional means of examining the *structural stability* for probing qualitative changes to the attractor as a response to perturbations, we study the *structural susceptibility* for quantifying the effects of the perturbations on the time series[1]. More specifically, we ask how sensitive is the dynamical system $d\mathbf{z}/dt = \mathbf{f}(\mathbf{z})$ to infinitesimal changes of the form $d\mathbf{z}/dt = \mathbf{f}(\mathbf{z}) + \mathbf{a} \cdot \mathbf{g}(\mathbf{z})$, for a family of perturbations $\mathbf{g}(\mathbf{z})$ when the parameters $\mathbf{a} \to \mathbf{0}$.

We introduce the new concept of "structural susceptibility" in dynamical systems, that is an outgrowth of our group's previous work on "sloppiness" in multiparameter systems wherein we have found that data-fitting in a number of nonlinear, multiparameter models is only sensitive to a few directions in parameter space at the best fit [1, 4, 29]. The key difference between studying sloppiness and structural susceptibilities is that in the former, the parameters are intrinsic to the system, i.e., there are no externally introduced changes in their evolution laws. Nonetheless, the methodology we have developed for studying sloppy models is also suited for studying structural susceptibilities of dynamical systems. Our approach cleanly isolates and ranks the directions in parameter space in order of relevance to observed behavior, and has previously led us to suggest improvements in experimental design [30], extract falsifiable predictions from experiments [31], and develop faster minimization algorithms [32]. Others have developed these ideas to suggest further improvements in experimental design [33] and parameter estimation [34], to quantify robustness to parameter variations [35], and to set

[1] We employ the word *structural* in the same context as its usage in dynamical systems literature on *structural stability*. The word *susceptibility* is inspired from physics wherein it is a measure of response to a perturbation (such as an applied external field) quantified by the second-derivative of the free energy w.r.t. parameters. Since cost is analogous to free energy (in that both are minimized), it is natural to call the response to perturbations in dynamics, also quantified via second derivatives, as *structural susceptibility*

confidence regions for predictions in multiscale models [36]. In this paper, we bring similar ideas together to analyze sensitivities of time series to perturbations in dynamical systems.

We demonstrate the utility of our approach with application to a dynamical system with two time scales—the van der Pol oscillator [18] which is a single parameter system and hence not amenable to sloppy model analysis. Instead, by choosing appropriate perturbations $\mathbf{g}(\mathbf{z})$, we calculate the susceptibility of its dynamics: We make perturbations on the attractor, and then systematically increase the separation of time scales in its dynamics to show how it can generally enhance the sloppiness in nonlinear systems.

2.3 Multiple Time Scale Dynamics

Multiple time scales are often found in the solutions of dynamical systems [37]. Broadly speaking, the defining criterion of these models is that the trajectory of one or more phase variables has an identifiable fast piece such as a jump or a pulse and a slow piece where the value of the variable doesn't change quickly [38]. In two dimensions, these systems are commonly studied in the contexts of slow-fast vector fields written as:

$$\begin{aligned} \epsilon\dot{x} &= X(x, y, \epsilon), \\ \dot{y} &= Y(x, y, \epsilon) \end{aligned} \tag{2.1}$$

where the parameter $\epsilon > 0$ is small and dot indicates derivative with respect to time t. For $\mathcal{O}(1)$ functions X and Y, and $X \neq 0$: $\dot{x} = \mathcal{O}(1/\epsilon)$ and $\dot{y} = \mathcal{O}(1)$, so that ϵ is the ratio of time scales in the system. On one extreme, the singular limit $\epsilon = 0$ corresponds to a differential algebraic system $X = 0,\ Y = \dot{y}$ where the solutions of $X = 0$ comprise the "critical manifold" close to which the flow in phase space is slow (the "slow manifold"). Similarly, $\epsilon = 1$ corresponds to a limit where there is no separation of time scales, with a crossover at intermediate values of ϵ.

Originally introduced in 1927, the van der Pol equation, $\ddot{x} - \mu(1 - x^2)\dot{x} + x = 0$, is a well-studied example of a second-order, nonlinear system with multiple time scales in its solution. Using the Liénard transformation $y = x - x^3/3 - \dot{x}/\mu$, and redefining time $t \to t\mu$, the equation can be written as a two dimensional system [38, 39] given by:

$$\begin{aligned} \mu^{-2}\dot{x} &= x - \tfrac{x^3}{3} - y, \\ \dot{y} &= x, \end{aligned} \tag{2.2}$$

which has the same form as (2.1) with $\epsilon = \mu^{-2}$. The global attractor of this dynamical system is a structurally stable limit cycle with two time scales[2].

The van der Pol system provides a convenient way to separate time scales by varying μ: Small values of μ in the van der Pol system correspond to a small separation of time scales. As evident from the second-order van der Pol equation, the trajectory of $x(t)$ approaches that of the harmonic oscillator as $\mu \to 0$. At large values of μ, the system shows a separation of time scales which increases with increasing μ. As shown in fig. 2.1 (b, c), with increasing μ, the trajectory of x separates into a slow part that lies $\mathcal{O}(\mu^{-2})$ close to the phase space curve given by $\dot{x} = 0$, *i.e.* the critical manifold $y = x - x^3/3$, and a fast part which connects the two branches of the slow flow. Likewise, the separation of time scales in y are associated with the increasing sharpness of the kink in its trajectory.

We view the van der Pol system as one member of a multiparameter family of models sharing a periodic cycle. The fact that with an increasing separation of time scales the trajectory spends an increasing amount time on the slow manifold and a decreasing amount of time on the jumps has important implications for distinguishing one model from another in the limit of large separation of time scales. With increasing scale separation, one expects that the

[2]Incidentally, the set of equations (2.2) can also be considered a special case of the FitzHugh-Nagumo model (*See* R. FitzHugh. Impulses and Physiological States in theoretical models of nerve propagation *Biophys J.*, 1(6):445, 1961) introduced three decades later as simplification of the Hodgkin-Huxley equations of neuronal spikes in the squid giant axons, and is sometimes referred to as the Bonhoeffer-van der Pol model

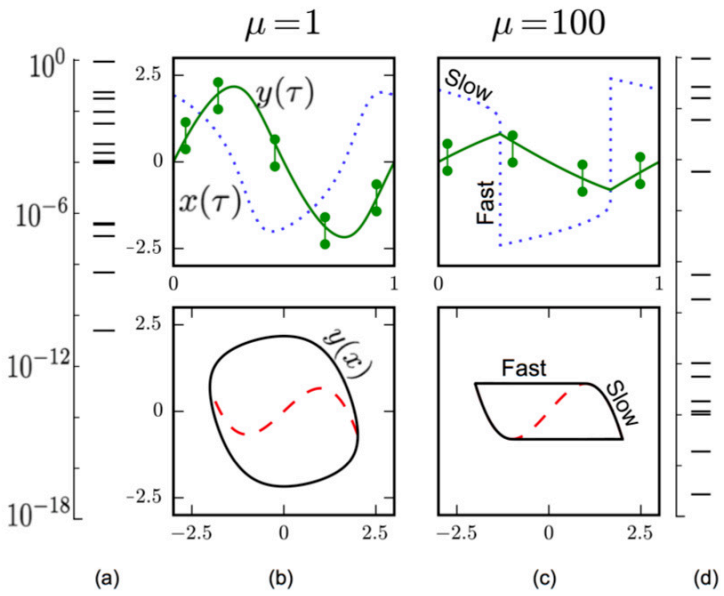

Figure 2.1: (a, d) Eigenvalues of the Hessian matrix of the cost of fitting at $\mu = 1$ (left) and $\mu = 100$ (right) for a multi-parameterized van der Pol discussed in section 2.5. (b, c top row) One period of time series $x(\tau)$ (dotted line), and $y(\tau)$ (solid line), for $0 < \tau < 1$, are shown for $\mu = 1$ and $\mu = 100$ as function of time along with schematic error bars for the data-fitting of the trajectory of variable y. (b, c bottom row) The orbit in xy plane (solid line) and the critical manifold (dashed line) As $\mu \to \infty$, the orbit collapses onto the critical manifold with the trajectory spending most of its time on the slow manifold and vanishingly short on the jumps.

cost of fitting (the sum of squared-residuals) will be decreasingly sensitive to changes in the jumps of the trajectory as they get progressively shorter in duration.

2.4 Sloppiness in Nonlinear Fits

In this section, we discuss the concepts of sloppiness and structural susceptibility in more detail with examples as a prelude to the calculations. For time series $z(t, \mathbf{a})$, a least-squares fit to data d_i minimizes a cost $C = \frac{1}{2}\sum_i (z(t_i, \mathbf{a}) - d_i)^2/\sigma_i^2$ in the space of system parameters

which are collectively denoted as **a**. The discovery of sloppiness is essentially that the eigenvalues of the Hessian of the cost with respect to parameters, $\mathcal{H}_{\alpha\beta} = \partial^2 C/\partial a_\alpha \partial a_\beta$, at the best fit span many orders of magnitude. The larger and smaller eigenvalues correspond to stiffer and sloppier directions respectively. For concreteness, consider a time series of a multi parameter model, such as the one denoted by $y(\tau)$ in fig. 2.1(b, top row). The error bars schematically show the least-squares fit of $y(\tau)$ and the sidebar (fig. 2.1(a)) shows the eigenvalues of the corresponding Hessian matrix. Note the broad range of eigenvalues ($\sim 10^{11}$, corresponding to a factor of almost a million in parameter range)— a feature that turns out to be typical in nonlinear fits.

Another vivid example of sloppiness in nonlinear models is provided by the well-established formalism behind the characterization of the sensitivities of initial conditions using Lyapunov exponents [40]. Consider $d\mathbf{z}/dt = \mathbf{f}(\mathbf{z})$ as a model whose parameters are the initial conditions $a_\alpha = z_\alpha(0)$ and whose predictions are the final positions $z_i(t)$ at time t. At the best fit, $\mathcal{H}_{\alpha\beta} = (J^T J)_{\alpha\beta}$ where $J_{i\alpha} = \partial z_i(t)/\partial z_\alpha(0)$ is the Jacobian of the sensitivities to perturbations in the initial conditions. The Lyapunov exponents, which are defined to be the eigenvalues ℓ_n of $\mathbf{L} = \lim_{t\to\infty} 1/(2t)\log(J^T J)$, utilize the same Hessian we would use in calculating the sloppy model eigenvalues $\lambda_n = exp(2t\ell_n)$. The typical roughly equal spacing of Lyapunov exponents naturally explains both the typical exponentially broad range of sloppy model exponents and the associated roughly equal spacing of $\log(\lambda_n)$ for a model with initial conditions as parameters.

Instead of the sensitivities with respect to the initial conditions or other intrinsic parameters, we focus here on the sensitivity of the dynamics to changes in the dynamical evolution laws. Therefore, for the remainder of this paper we will be interested in a cost function that measures the square of the distance between two time series for the system $d\mathbf{z}/dt = \mathbf{f(z)} + \mathbf{a}\cdot\mathbf{g}(z)$—

one with perturbation $\mathbf{a}$, $\mathbf{z}(t, \mathbf{a})$, and the other one with no perturbation, $\mathbf{z}(t, \mathbf{a} = \mathbf{0})$:

$$C = \frac{1}{2} \int_0^T ||\mathbf{z}(t, \mathbf{a}) - \mathbf{z}(t, \mathbf{0})||^2 \, \mathrm{d}t \tag{2.3}$$

with the perturbing terms $g_i(z)$ giving a power series in the components of $\mathbf{z}$. Further in the manuscript, we will use this form of the cost to compute the susceptibility of the van der Pol system and show how sloppiness is enhanced by increasing separation of time scales in the van der Pol equations. This is in essence captured by fig. 2.1(a & d) where we show that an increase in the van der Pol parameter μ from 1 to 100 produces roughly a million-fold increase in the spread of eigenvalues.

2.5 Susceptibility of van der Pol system

We perturb the van der Pol system in (2.2) by adding a series of additional terms. There is a long tradition in dynamical systems of studying equations of motion of polynomial form [40, 41]; indeed, the theory of normal forms [41, 42] suggests that a broad range of dynamical systems near bifurcations can be generically mapped into a polynomial form by a nonlinear but smooth change of variables. Adding extra polynomial terms can be used to 'unfold' the qualitative behavior near bifurcations [42]. Here we focus on quantitative changes far from bifurcations. In choosing our perturbations, we must cut off the polynomials at some order. There are two ways in which we specialize our general susceptibility analysis to the two time scale, periodic limit cycle of the van der Pol system. First, we choose the family of perturbations of order $3N$ as follows:

$$\begin{aligned} \mu^{-2}\dot{x} &= x - \tfrac{x^3}{3} - y + \textstyle\sum_{m+n\leq N} a_{m,n}(x - \tfrac{x^3}{3} - y)^m x^n \\ \dot{y} &= x. \end{aligned}$$

This choice has two noteworthy features— (a) We have grouped the polynomial perturbations so that, for $m \neq 0$ they vanish on the critical manifold, $y = x - x^3/3$. That is, the parameters

$a_{m,n}$ with $m \neq 0$ do not significantly affect the dynamics on the slow manifold; we call these the "fast parameters" and correspondingly the $a_{0,n}$ are "slow parameters". The parameter $a_{1,0}$ duplicates the effect of μ and thus we omit it. Surely, the eigenvalue spectrum of the general polynomial expansion, $a_{m,n}x^m y^n$, behaves qualitatively similarly to the one presented here but our parametrization greatly simplifies the analysis of the eigenvector perturbations. (b) We only perturb the $\dot{x}$ equation. Our choice corresponds to a general expansion of a second-order equation, with the acceleration $\ddot{y} = \dot{x}$ written as a polynomial in the position y and velocity $\dot{y} = x$. Perturbing both equations produces similar behavior.

Second, we modified the cost to focus on the limit cycle of the van der Pol system in two ways— (a) by rescaling all trajectories in our analysis so that they have the same unit period, and (b) by changing the initial condition[3] so that the perturbed orbit and the unperturbed orbit both start at $\mathbf{y}_0 = (x_0, y_0)$ with $y_0 = 0$. When we correct the period T by δT, initial conditions $\mathbf{y}_0$ by $\delta\mathbf{y}_0$, and do an overall rescaling of the time variable $t \to \tau T$, the cost functional for the time series of $y(\tau)$ at each μ takes the following form:

$$C(\mu) = \frac{1}{2}\int_0^1 [y(\tau, \mathbf{a} + \delta\mathbf{a}, \mathbf{y}_0 + \delta\mathbf{y}_0, T + \delta T) - y(\tau, \mathbf{a}, \mathbf{y}_0, T)]^2 \, \mathrm{d}\tau \tag{2.4}$$

In principle, changes in both time series, $x(\tau)$ and $y(\tau)$ could be incorporated in the cost function, but we get qualitatively similar results by keeping either or both variables. Choosing to measure changes only in $y(\tau)$ corresponds again to studying the second-order equation for $\ddot{y}$ as an expansion in y and $\dot{y}$.

The susceptibilities are still given by the Hessian matrix at the best fit ($\mathbf{a} = \mathbf{0}$):

$$\mathcal{H}(\mu)_{\alpha\beta} = \frac{\partial^2 C(\mu)}{\partial a_\alpha \partial a_\beta} \tag{2.5}$$

[3] Perturbations distort the dynamics so that the attractor and its period change. We addressed these issues by setting the periods to unity, and by moving the initial conditions to the new attractor to remove any transients. Alternatively, if we fit data over many periods without making the said changes, the parameter combinations determining the period and phase would become stiff modes in our dynamics.

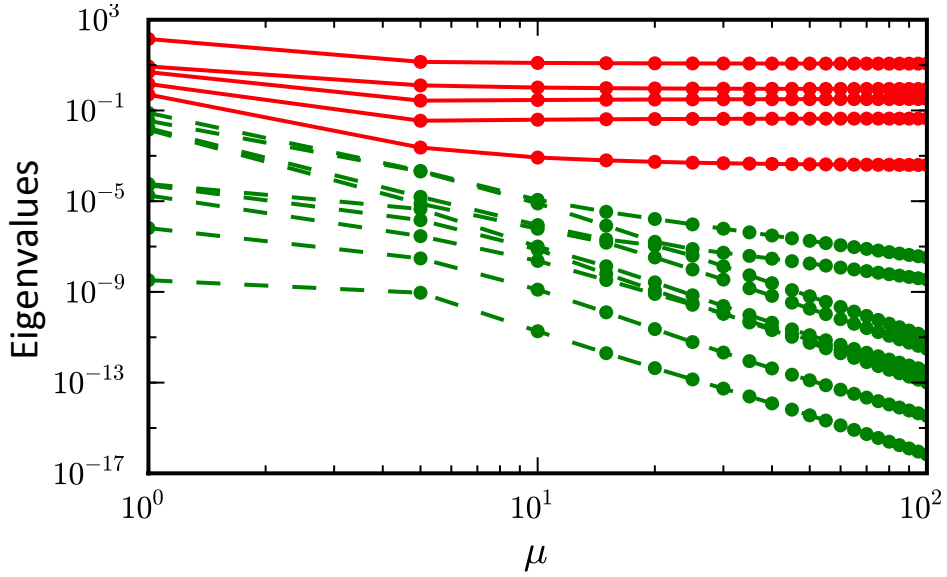

Figure 2.2: Eigenvalues of Hessian matrix are shown here as a function of μ. The range $1 \leq \mu \leq 100$ corresponds to a ratio of time scales $1 \leq \epsilon \leq 10000$. The five largest eigenvalues (solid lines) correspond to stiff directions in the parameter space: these directions perturb the slow manifold. The remainder (dashed lines) affect the transient part of the trajectory which becomes smaller with an increasing separation of time scales and hence these directions are decreasingly relevant.

which can be written out more completely as:

$$\mathcal{H}(\mu)_{\alpha\beta} = \int_0^1 \left(\frac{\partial y}{\partial a_\alpha} + \frac{\partial y}{\partial \mathbf{y}_0}\frac{\partial \mathbf{y}_0}{\partial a_\alpha} + \frac{\partial y}{\partial T}\frac{\partial T}{\partial a_\alpha} \right) \times \left(\frac{\partial y}{\partial a_\beta} + \frac{\partial y}{\partial \mathbf{y}_0}\frac{\partial \mathbf{y}_0}{\partial a_\beta} + \frac{\partial y}{\partial T}\frac{\partial T}{\partial a_\beta} \right) \mathrm{d}\tau$$

Here, each of the two terms in the integral is to be interpreted as a Jacobian matrix, a mapping from the finite dimensional parameter space to the infinite dimensional data space:

$$J_{\tau\alpha} = \frac{\partial y(\tau)}{\partial a_\alpha} + \frac{\partial y(\tau)}{\partial \mathbf{y}_0}\frac{\partial \mathbf{y}_0}{\partial a_\alpha} + \frac{\partial y(\tau)}{\partial T}\frac{\partial T}{\partial a_\alpha} \tag{2.6}$$

The sensitivity trajectories in the Jacobian, $\partial y/\partial a_\alpha$, $\partial y/\partial \mathbf{y}_0$, and $\partial y/\partial T$, were computed using the open source SloppyCell package [43, 44]. The expressions for the time invariant quantities, $\partial \mathbf{y}_0/\partial a_\alpha$ and $\partial T/\partial a_\alpha$, were found by enforcing periodicity of the perturbed time series denoted by $\mathbf{y}(\tau) \equiv (x(\tau), y(\tau))$ as follows:

$$\mathbf{y}(\tau = 0, \mathbf{a} + \delta\mathbf{a}, \mathbf{y}_0 + \delta\mathbf{y}_0, T + \delta T) = \mathbf{y}(\tau = 1, \mathbf{a} + \delta\mathbf{a}, \mathbf{y}_0 + \delta\mathbf{y}_0, T + \delta T),$$

Taylor expansion of both sides of the previous equation leads to a vector equation:

$$\delta \mathbf{y}_0 = \left.\frac{\partial \mathbf{y}}{\partial T}\right|_{\tau=1} \delta T + \left.\frac{\partial \mathbf{y}}{\partial \mathbf{a}}\right|_{\tau=1} \delta \mathbf{a} + \left.\frac{\partial \mathbf{y}}{\partial \mathbf{y}_0}\right|_{\tau=1} \delta \mathbf{y}_0,$$

from which both constants can be computed following the convention that the component denoting the change in initial condition of $y(\tau)$ in $\delta \mathbf{y}_0$ is set to zero. Now with the Jacobian calculated, the Hessian at best fit is simply $\mathcal{H} = J^T J$.

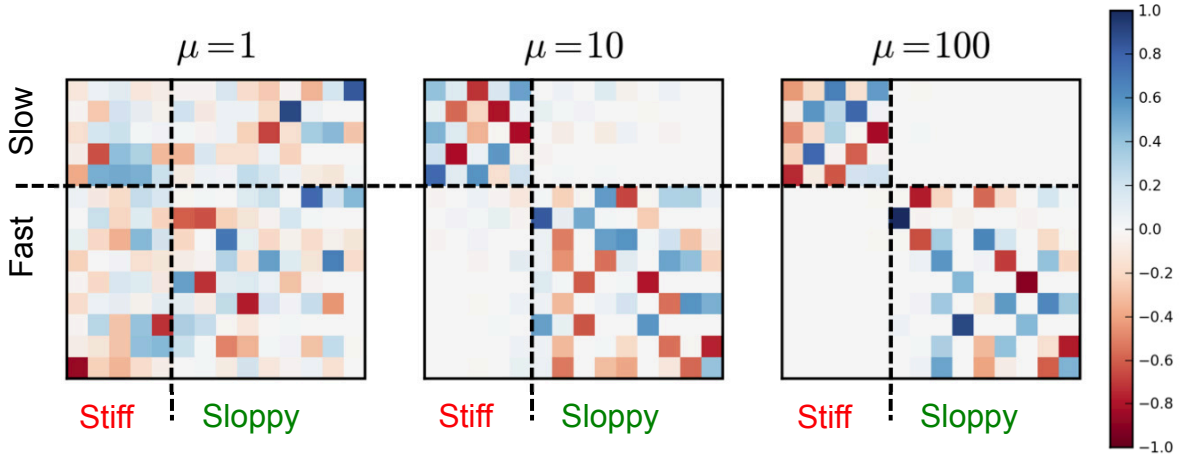

Figure 2.3: Hessian eigenvectors are shown for $\mu = 1,\ 10$, and 100. Each colored small square shows the magnitude of an eigenvector component (the scale bar shown on the right). Eigenvectors for each μ are sorted so that the stiffer ones appear on the left; individual components are sorted so that "slow parameters" appear on the top. Note that with increasing μ, the stiff and sloppy eigenvectors separate by parameters: The stiff eigenvectors only have projections along the slow parameters which perturb the slow manifold, whereas the sloppy directions have projections along the fast parameters which mainly perturb the jumps.

2.5.1 Eigenvalues and Eigenvectors

We computed the Hessian matrix given by the previous equation at multiple values of μ for $N = 4$ where there are 14 parameters. The spread of eigenvalues (fig. 2.2) increases as a function of μ confirming that sloppiness increases with an increasing separation of time scales.

Not surprisingly[5] the eigenvalues for $\mu = 1$ already span 11 orders of magnitude, while for $\mu = 100$, we observe that the stiffest eigenvalue is 18 orders of magnitude larger than the smallest one— the spread increases by 10^7 when μ increases to 100.

Taken together with the eigenvectors shown in fig. 2.3, some interesting facts come to light: Fig. 2.2 shows that with increasing μ, the eigenvalues separate into two clusters of closely related decay exponents. The largest N eigenvalues approach constants. The other eigenvalues decay with power laws: two modes with exponents between -2 and -3 and the remaining with exponents between -5 and -6. Similarly, fig. 2.3 shows that the eigenvectors also separate into two groups with increasing μ: The stiffest directions are linear combinations of the slow parameters whereas the sloppy directions are comprised of other parameters as expected.

We can understand the effect of perturbations in parameter combinations given by the eigenvectors (called *eigenparameters*) $\hat{\mathbf{e}}_k$ more clearly by observing their behavior in the data space. The Jacobian transformation of (2.6) projects the eigenvectors to the data space: $\delta y_k = J \cdot \hat{\mathbf{e}}_k / \sqrt{\lambda_k}$ where λ_k corresponds to the k^{th} largest eigenvalue. Defined this way, these data space vectors, called *eigenpredictions* [4], δy_k, are also orthonormal. Alternatively, the eigenpredictions are the left singular vectors in the singular value decomposition of the Jacobian (i.e. they are the columns of the unitary matrix U in $J = U\Sigma V^T$ [45]). As shown in fig. 2.4 for $\mu = 1,\ 10\ \&\ 100$ (top three rows), we learn from the eigenpredictions that the stiff modes affect behavior both along the slow manifold and at the jumps. Moreover with increasing μ, as the eigenvalues associated with the stiff directions approach constants (fig. 2.2), so do the stiff eigenpredictions (fig. 2.4 rows 2, 3 columns (a) and (b)). The sloppy modes on the other hand, affect the dynamics on the jumps only. The maximum amplitudes of the (normalized) sloppiest eigenpredictions appears to increase roughly proportional to μ

[5] We understand this as sloppiness as arising due to the generalized interpolation argument [4].

(corresponding to the jump duration of $\sim \mu^{-2}$). In the limit, these become δ-functions and derivatives concentrated at the jumps. Fig. 2.4 (bottom row) also shows the limit cycles (*eigencycles*) with eigenparameter perturbations as phase space trajectories $(x, y + \eta\ \delta y_k)$ for small η.

2.6 DISCUSSION

In this paper, we have introduced a formalism we call "structural susceptibility" for analyzing the quantitative dependence of dynamical systems to perturbations of the equations of motion. It is a generalization of the Lyapunov exponents governing the dependence on initial conditions. It is in the spirit of 'unfolding' methods of bifurcation theory. And finally, it exposes the ubiquitous presence of broad range of sloppy eigendirections in parameter space— largely unimportant to the dynamics. We used this method to study the role of time scale separation in enhancing the sloppiness of the susceptibility spectrum in the particular case of the van der Pol oscillator.

By extending the framework of our sloppy model analysis to systems where changes in evolution laws are to be studied, our method offers a simple way to calculate the effects of broad classes of perturbations. By studying the structural susceptibility of a dynamical system with two time scales, the analysis presented here showed that sloppiness of nonlinear systems is enhanced by separation of time scales in the dynamics. With increasing separation of time scales in the van der Pol oscillator, the trajectory spends an increasing amount of time on the slow manifold and a vanishingly small amount of time in the transition region. The cost of perturbations is integrated over time and therefore we are unsurprised that the perturbations that change the slow manifold will accrue the most cost and therefore manifest as stiff modes of the Hessian matrix. The remaining directions are sloppy as they only affect

the behavior at the jumps or the fast pieces. These perturbations manifest as δ-functions and their derivatives in the limit of $\mu \to \infty$, significantly affecting the phase-space trajectory, but over only the fast times asymptotically ignored in the least-squares cost. It remains a challenge to connect separation of time scales to parameter sensitivity in more complicated systems, but the analogy of the van der Pol system's behavior with other nonlinear physical systems of interest is clear.

Many important dynamical systems have multiple time scales in their solutions: examples include models in neuroscience (such as Hodgkin-Huxley model), systems biology or chemical reaction systems (such as protein network models), and in engineering (such as models of combustion, lasers, locomotion, etc.). Our analysis suggests that any system with multiple time scales should become sloppier as the scales separate for the same reasons as we found in the van der Pol: Some parameter combinations will only affect the fast dynamics, and lead to sloppy modes. Perturbations which affect the slow dynamics will presumably accrue more cost and be stiff.

More broadly, the sloppiness exposed by our structural susceptibility analysis has clear implications for attempting to reconstruct the equations of motion from experimental data [46] because parameter identification along any sloppy eigendirection will be relatively poorly determined by the dynamics. This discovery has already influenced work on experimental design optimization: estimating parameters is challenging [33, 47], but extracting predictions without constraining parameters is straightforward [31]. We further anticipate that the concept of structural susceptibility will be useful for studying systems with chaos, bifurcations and phase transitions; quantifying the unfoldings of these systems may also be useful for gaining a deeper understanding of the phenomena they model.

2.7 ACKNOWLEDGMENTS

The research described in this chapter was done in collaboration with M. Transtrum, and was directed by J. Sethna. Helpful discussions with S. Papanikolaou yielded important insights about perturbations to slow manifold, and dynamical systems analogues of the thermodynamic susceptibilities. The calculations and analysis also gained from the valuable input of J. Guckenheimer who also made helpful suggestions in editing the published manuscript for inclusion into this thesis. The project was financially supported in part via NSF grant DMR 1005479.

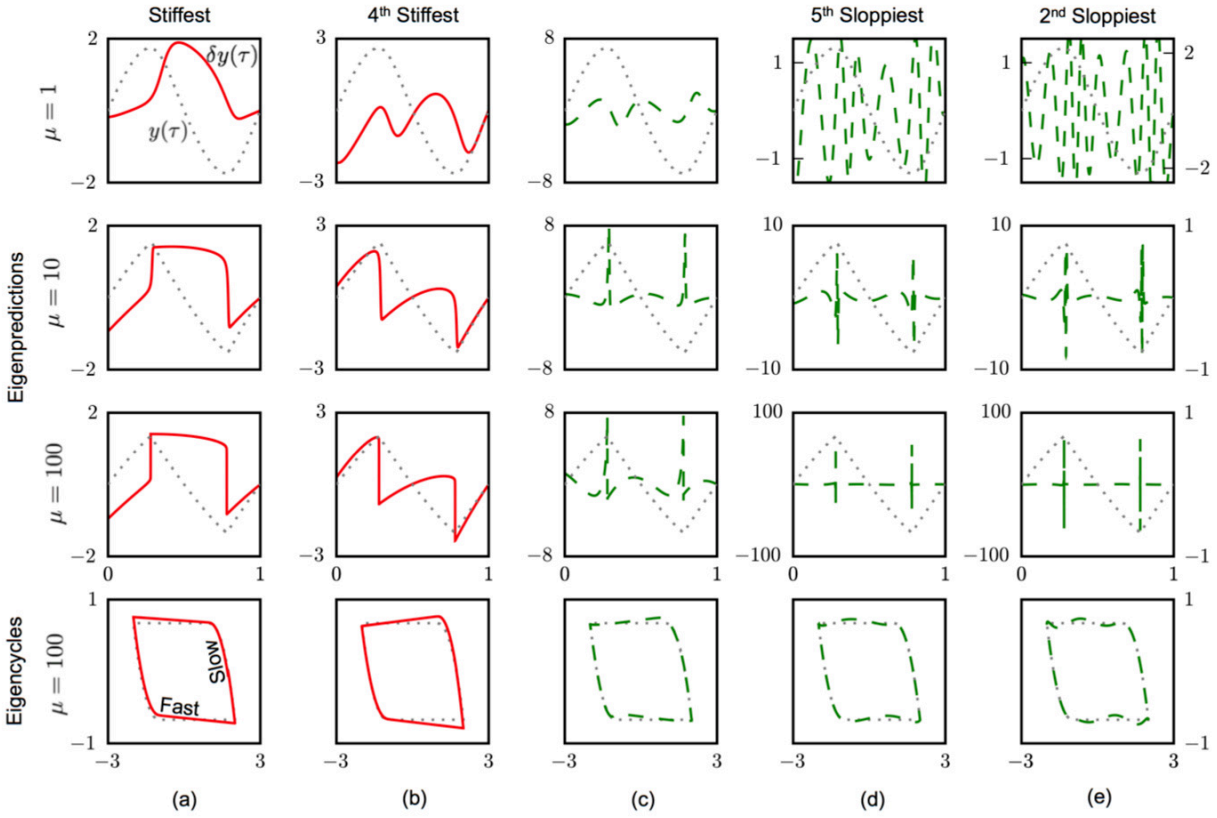

Figure 2.4: Top three rows: Eigenpredictions δy_k for $k = 0,\ 3,\ 6,\ 9,\ 12$ at $\mu = 1,\ 10\ \&\ 100$ are shown in solid red lines for stiff modes and dashed green for sloppy modes. These curves show the response of perturbations if the parameters are changed infinitesimally along the Hessian eigenvectors: A parameter change of norm ϵ along eigendirection n will change the trajectories by $\lambda_n \epsilon$ times the eigenpredicton. Dotted gray lines show unperturbed van der Pol solution for comparison (y scale on the right hand side). As the time scales separate, the amplitudes of the sloppiest eigenpredictions increase (roughly in proportion to μ) getting increasingly concentrated at the jumps. Bottom row shows the eigencycles for $\mu = 100$ in solid red lines and green dashed lines corresponding to the perturbations in row 3 (i.e. the new limit cycle for a perturbation of strength $\epsilon \sim 1/\lambda_n$. These curves show how the van der Pol orbit changes with perturbations along the Hessian eigenvectors. Both the stiff and the sloppy modes change the orbit at the jumps (occurring at the extrema in the dashed lines); the stiff modes also change behavior at the slow manifold, whereas the sloppy modes only affect the jumps.

CHAPTER 3

CHAPTER 3: PARAMETER SPACE COMPRESSION YIELDS EMERGENT THEORIES OF PHYSICS

3.1 ABSTRACT

The microscopically complicated real world exhibits behavior that is comprehensible, often yielding to simple yet quantitatively accurate descriptions. Predictions are possible despite large uncertainties in microscopic parameters, in physics and in multiparameter models in other areas of science. This chapter shows a connection between the two through an examination of parameter sensitivities in a discrete diffusion model and a generalized Ising model of ferromagnetism. In both cases, the emergence of an effective theory for long-scale observables is linked to a compression of the parameter space quantified by the eigenvalues of the Fisher Information Matrix. Strikingly similar compressions appear ubiquitously in models taken from diverse areas of science, suggesting that the parameter space structure underlying effective continuum and universal theories in physics also permits predictive modeling more generally.

[0]The work described in this chapter was published [22] with the citation: B. B. Machta, R. Chachra, M. K. Transtrum, J. P. Sethna. *Parameter Space Compression Underlies Emergent Theories and Predictive Models* Science, 342, 604-607 (2013). A preliminary version of this chapter also appears in Machta's dissertation [48]. The present author significantly contributed to the text and figures as seen in this chapter sections 3.1–3.6, in addition to assisting Machta with computational implementation and analyses relevant to the Ising model.

3.2 INTRODUCTION

The success of physics [49] and the comprehensibility of nature is owed to the hierarchical character of scientific theories [50]. These theories of our physical world, ranging in scales from the sub-atomic to the astronomical, model natural phenomena as if physics at macroscopic length scales were almost independent of the underlying, shorter length scale details. For example, understanding string theory or some other fundamental high energy theory is not necessary for quantitatively modeling the behavior of superconductors that operate in a lower energy regime. The fact that many lower level theories in physics can be systematically coarsened (renormalized) into macroscopic effective models, establishes and quantifies their hierarchical character. Moreover, a similar hierarchy of theories is also at play in multiparameter models in other areas of science [46, 51–56]. Disparate as they may seem, the key finding discussed in this chapter is that the hierarchy of theories in physics relies on the same parameter space compression that is ubiquitous in general multiparameter models. This suggests that even where model reduction cannot be systemically generated, a smaller effective theory could still capture most of the observable behavior.

Recent studies of nonlinear, multiparameter models drawn from disparate areas in science have shown that predictions from these models largely depend only on a few 'stiff' combinations of parameters [1, 8, 57]. This recurring characteristic (termed 'sloppiness') appears to be an inherent property of these models and may be a manifestation of an underlying universality. Indeed, many of the practical and philosophical implications of sloppiness are identical to those of the renormalization group (RG) and continuum limit methods of statistical physics: models show weak dependance of macroscopic observables on microscopic details. They thus have a smaller effective model dimensionality than their microscopic parameter space. To clarify their connection to sloppiness, we develop and

apply an information theory based analysis to models where the continuum limit and the renormalization group already give a quantitative explanation for the emergence of effective models—a discrete model of diffusion and an Ising model of the ferromagnetic phase transition. In both cases, our results show that at long time and length scales there is a similar compression of the microscopic parameter space, where sensitive, or 'stiff' directions correspond to the relevant macroscopic parameters (such as the diffusion constant in the hopping model).

3.3 The Fisher Information

The sensitivity of model predictions to changes in parameters is quantified by the Fisher Information Matrix (FIM)[1]. The FIM forms a metric that converts parameter space distance into a unique measure of distinguishability between a model with parameters θ^μ (for $1 \leq \mu \leq N$) and a nearby model with parameters $\theta^\mu + \delta\theta^\mu$ [58–60]). This divergence is given by $ds^2 = g_{\mu\nu}\delta\theta^\mu\delta\theta^\nu$ where $g_{\mu\nu}$ is the FIM defined by

$$g_{\mu\nu} = -\sum_{\vec{x}} P_\theta(\vec{x}) \frac{\partial^2 \log P_\theta(\vec{x})}{\partial\theta^\mu \partial\theta^\nu}. \tag{3.1}$$

Here, $P_\theta(\vec{x})$ is the probability that a (stochastic) model with parameters θ^μ would produce observables $\vec{x}$. In the context of nonlinear least squares, g is the Hessian of χ^2, the sum of squares of residuals of the data fit (derivation is shown in section A.1.1). Distance in this metric space is a fundamental measure of distinguishability in stochastic systems. Sorted by decreasing eigenvalues, eigenvectors of g describe progressively less important linear combinations of parameters that govern system behavior. Previously, it was shown that in nonlinear least squares models, the eigenvalues form a roughly geometrical sequence,

[1] A simple derivation is shown in section A.1

reaching extremely small values in many models (fig. 3.1). Thus, the eigenvalues of the FIM quantify parameter space compression: few 'stiff' eigenvectors in each model point along directions where observables are sensitive to changes in parameters, while progressively sloppier directions make little difference for observables. These sloppy parameters cannot be inferred from data, and conversely, their exact values do not need to be known to quantitatively understand system behavior [8]. To see how this relates to continuum models in physics, we now turn to a 'microscopic' model of stochastic motion from which the diffusion equation emerges.

3.4 Discrete Diffusion

The diffusion equation is the canonical example of a continuum limit. It governs behavior whenever small particles undergo stochastic motion. Given translation invariance in space and time, the complex microscopic collisions are subsumed into a dynamical equation for the particle density, $\rho(r, \tau)$, with only three coefficients: $\partial_\tau \rho(r, \tau) = D\nabla^2\rho - \vec{v} \cdot \nabla\rho + R\rho$. Here D is the diffusion constant, $\vec{v}$ is the drift, and R is the particle creation rate. Microscopic parameters describing the particles and their environment enter into this continuum description only through their effects on these three coefficients. To see this, consider a microscopic model of stochastic motion on a discrete one-dimensional lattice of sites, with $2N + 1$ parameters θ^μ, for $-N \leq \mu \leq N$ which describe the probability that in a discrete time step a particle will hop from site j to site $j + \mu$ (fig. 3.2 inset). At the initial time, all particles are at the origin, $\rho_0(j) = \delta_{j,0}$. The observables, $\vec{x} \equiv \rho_\tau(j)$, are the densities of particles at some later time τ. After a single time step the distribution of particles is given by $\rho_1(j) = \theta^j$. This distribution depends independently on all of its parameters, thus the FIM is the identity, $g_{\mu\nu} = \delta_{\mu\nu}$ (the calculation shown in section 3.8.1). After a single time step, there is no

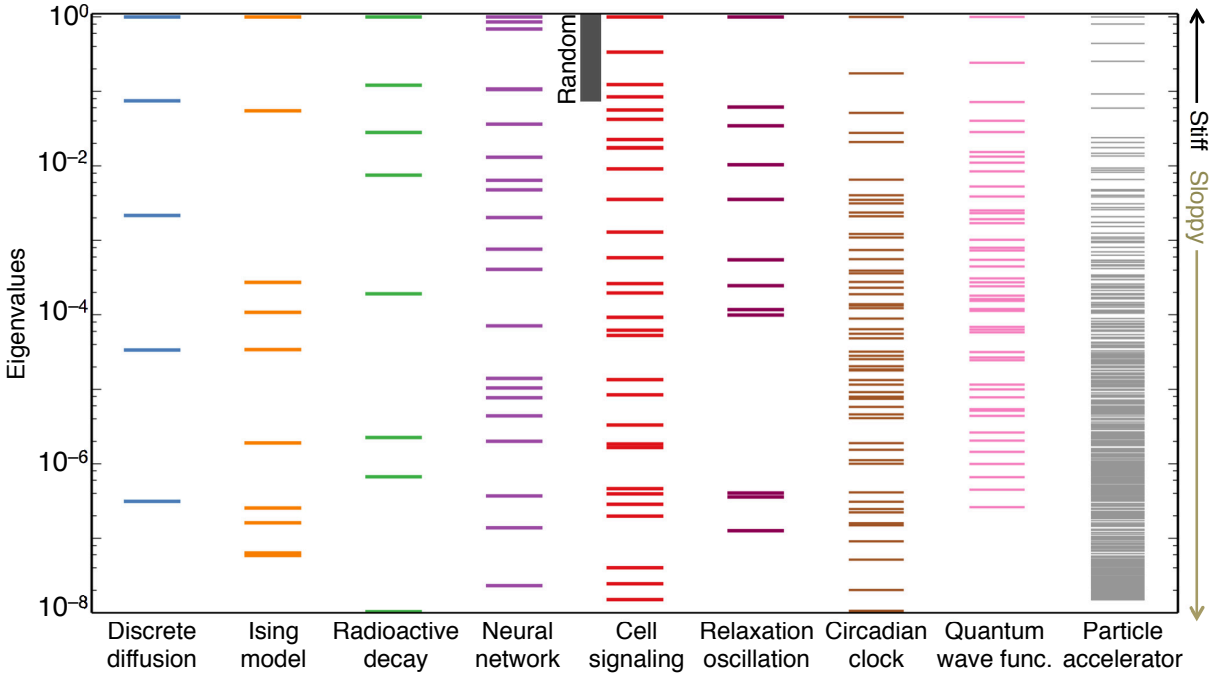

Figure 3.1: Normalized eigenvalues of the Fisher Information Matrix (FIM) of various models. The diffusion and Ising models are explored here. A radioactive decay model and a neural network are taken from [4]. The systems biology model is a differential equation model of a MAP kinase cascade taken from [5] and the adjoining band marked as "Random" shows a typical eigenvalue spread from a Wishart random matrix of the same size. The 'Relaxation oscillation' model is a modified Van der Pol system taken from [6]. Eigenvalues of the genetic network describing 'Circadian rhythm' model [7] are calculated in [1]. 'Variational wave function' eigenvalues are taken from Quantum Monte Carlo simulations as Jastrow parameters are varied [8]. 'Particle accelerator' is a model of beam shape simulated using the Tool for Accelerator Optics [9]. In all models, the eigenvalues of the FIM are roughly geometrically distributed, with each successive direction significantly less important for system behavior (only the first 8 decades are shown). This means that inferring the parameter combination whose eigenvalue is smallest shown would require $\sim 10^8$ times more data than the stiffest parameter combination. Conversely, the least important parameter combination is $\sqrt{10^8}$ times less important for understanding system behavior.

parameter space compression—each parameter is measured independently. When particles take several time steps before their positions are observed, some parameter combinations affect observable behavior more sensitively than others. At late times, the parameter

combination that controls the particle creation rate, R, becomes the most sensitive as the mean particle number changes exponentially with time. Further, combinations corresponding to the drift $\vec{v}$, and diffusion constant D, then emerge as the next most sensitive directions. Finally, eigenvectors describing the skew, kurtosis and higher moments of the final distribution become progressively less important, each with a higher negative power of time τ (fig. 3.2, section 3.8.1). This gives an information theoretic explanation for the wide applicability of the diffusion equation. Any system with stochastic motion and conservation of particle number will be dominated by the drift, $\vec{v}$, if it is present (for example, particles falling through honey under gravity), and by diffusion if the drift is constrained to be zero. Since the diffusion constant cannot be removed for stochastic systems, there is never a need for higher terms to enter into a minimalist continuum description. These results quantify a widely held intuition: one cannot infer microscopic parameters, such as the bond angle of a water molecule, from a diffusion measurement, and conversely it is unnecessary to have such knowledge to quantitatively understand the long length and time behavior of diffusing particles in water.

3.5 Ising Model

Continuum models like the diffusion equation arise when fluctuations are only large on the micro scale. Their success can be said to rely on the largeness and slowness of observables when compared with the natural scale of fluctuations. However, RG methods clarify that system behavior can be universal even when fluctuations are large on all scales, as occurs near critical points and for quantum field theories. The Ising model is the simplest model which exhibits these nontrivial fluctuations. Near its critical point, the Ising model predicts fractal domains whose statistics are universal. That is, it not only describes magnetic fluctuations in ferromagnets, but also the density fluctuations near a liquid-gas transition and

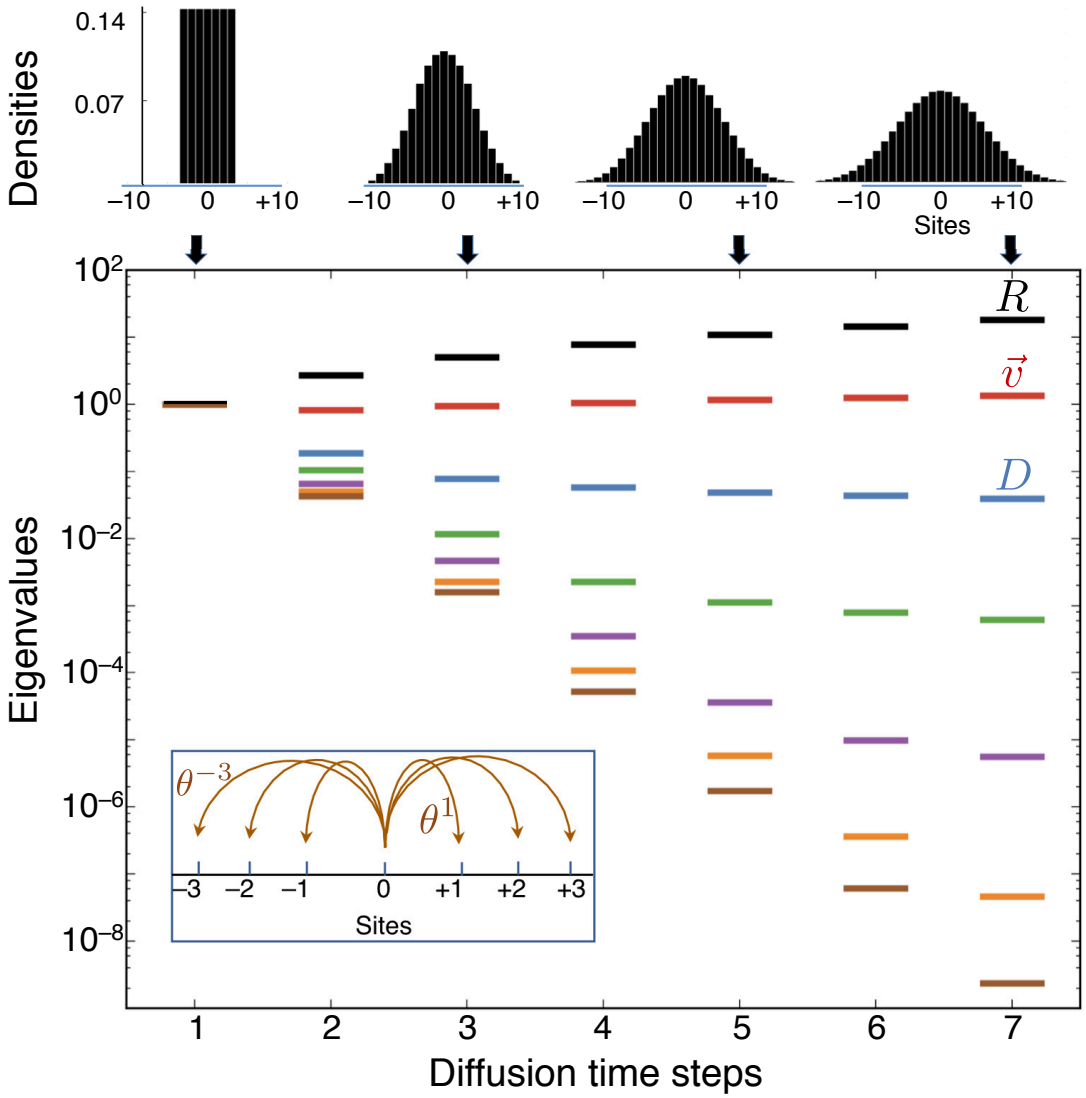

Figure 3.2: FIM eigenvalues of a model of stochastic motion on a 1-D lattice. The seven parameters describe probabilities of transitioning to nearby sites (bottom inset). Observations are taken after a given number of time steps for the case where all parameters take the value $a^{\mu} = 1/7$. Top row shows the resulting densities plotted at times $\tau = 1, 3, 5, 7$. Bottom plot shows the eigenvalues of the FIM versus number of steps. After a single time step, the FIM is the identity, but as time progresses, the spectrum of the FIM spreads over many orders of magnitude. The first eigenvector measures deviations in the net particle creation rate R from 0, the second measures a net drift V in the density, and the third corresponds to parameter combinations that change the diffusion constant D. Further eigenvectors describe parameter combinations that do not affect these macroscopic parameters, but instead measure the skew (green), kurtosis (purple), and higher moments of the resulting density (orange and brown).

composition fluctuations near a liquid-liquid miscibility transition [62, 63]. Consider a two dimensional square lattice Ising model where at every site a 'spin' takes a value of $s_{i,j} = \pm 1$. Observables are spin configurations ($\vec{x} = \{s_{i,j}\}$) or subsets of spin configurations ($\vec{x}^n$, as defined below). The Ising model assigns to each spin configuration a probability given by its Boltzmann weight, $P_\theta(\vec{x}) = e^{-\mathcal{H}_\theta(\vec{x})}/Z$ and the model is parametrized through its Hamiltonian $\mathcal{H}_\theta(\vec{x}) = \theta^\mu \Phi_\mu(\vec{x})$. Parameters $\theta^{\alpha\beta}$ describe the coupling between spins and their neighbors at coordinates (α, β) away, so that $\Phi_{\alpha\beta}(\vec{x}) = \sum_{i,j} s_{i,j} s_{i+\alpha,j+\beta}$, while θ^0 is the external field multiplying $\Phi_0(\vec{x}) = \sum_{i,j} s_{i,j}$ (see inset of fig. 3.3). We examine the vicinity of the nearest neighbor Ising model in zero field, where $\theta^{01} = \theta^{10} = \beta J$ and $\theta = 0$ otherwise.

At the microscopic level, all spins are observable and the Ising FIM (derived in section 3.8.2) is a sum of 2 and 4-spin correlation functions that can be readily calculated using Monte-Carlo techniques [64]. Near the critical point, it has two 'relevant' eigenvectors with eigenvalues that diverge like the specific heat and magnetic susceptibility [65, 66]. These two large eigenvalues have no analog in the diffusion equation, and reflect the presence of fluctuations at scales much larger than the microscopic lattice constant. The remaining eigenvalues all take a characteristic scale given by the system size. The non-sloppy clustering of the remaining eigenvalues is reminiscent of the spectrum seen in the diffusion equation when viewed at its microscopic (time) scale. When observables are microscopic spin configurations, the nearest neighbor Ising model is a poor description of a binary liquid, and even of a ferromagnet.

To coarsen the Ising model (section 3.8.4), the observables are restricted to a subset of lattice sites chosen via checkerboard decimation procedure (fig. 3.3 top row insets). The FIM of equation A.5 is now measured using as our observables only those sites in a sub-lattice decimated by a factor 2^n, $\vec{x}^n = \{s_{i,j}\}_{\{i,j\} \in n}$. For example, after 1 level of decimation, this corresponds to the black sites on the checkerboard, while after 2 steps, only sites $\{i, j\}$ with

even i and j remain. Importantly, the distribution is still drawn from the ensemble defined by the original Hamiltonian defined on the full lattice. The calculation is implemented using compatible Monte-Carlo [67].

The results from Monte-Carlo[2] are presented for a 64×64 lattice at its critical point in fig. 3.3. The irrelevant and marginal eigenvalues of the metric continue to behave much as the eigenvalues of the metric in the diffusion equation, becoming progressively less important under coarsening with characteristic eigenvalues. However, the large eigenvalues, dominated by singular corrections, do not become smaller under coarsening; they are measured by their collective effects on the large scale behavior, which is primarily informed by large distance correlations. Later, we use RG analysis to explain the scaling of the FIM eigenvalues with the coarse-graining level. The analysis clarifies that 'relevant' directions in the RG are exactly those whose FIM eigenvalues do not contract on coarsening. They control the large-wavelength fluctuations of the model and dominate the behavior provided that the correlation length of fluctuations is larger than the observation scale.

3.6 Discussion

We have seen that neither the hopping model nor the Ising model are sloppy at their microscopic scales. It is only upon coarsening the observables, either by allowing several time steps to pass, or by only observing a subset of lattice sites, that a typical sloppy spectrum of parameter combinations emerges. Correspondingly, multiparameter models such as in systems biology and other areas of science are sloppy only when fit to experiments that probe collective behavior— if experiments are designed to measure one parameter at a time, no model compression can be expected [30, 33].

[2] Simulation details are discussed in 3.8.6.

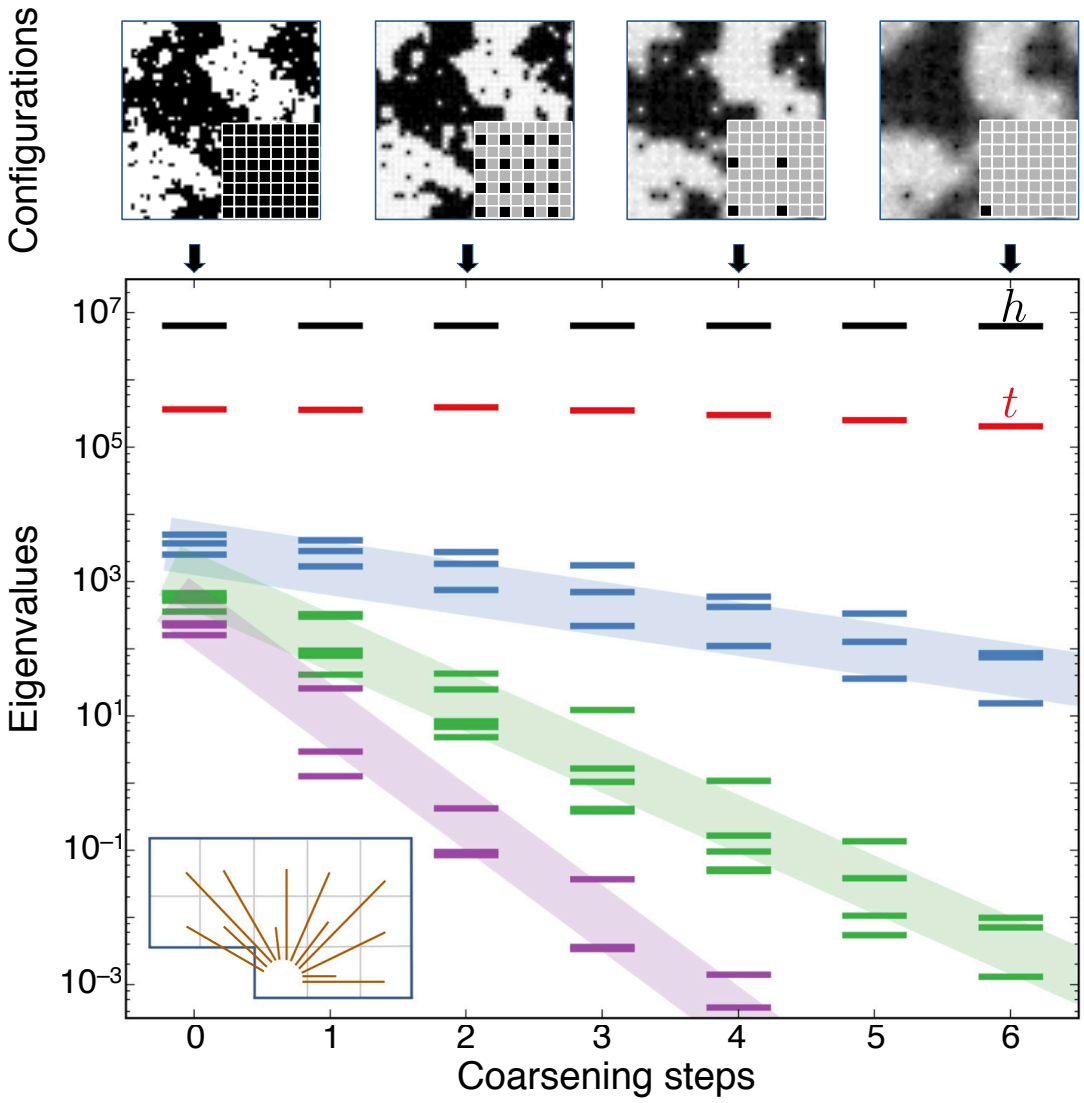

Figure 3.3: FIM eigenvalues of an Ising model of ferromagnetism. See text for definition; 13 parameters describe nearest and nearby neighbor couplings (bottom inset) and a magnetic field. Observables are spin configurations of all spins on a sub-lattice (dark sites in the insets of the top panel). The top panel shows one particular spin configuration generated by the model, suitably blurred for step > 0 to the average spin conditioned on the observed sub-lattice values. Some information about the configuration, such as the typical size of fluctuations, is preserved under this procedure, whereas other information like the nearest neighbor correlation amplitude is lost. The two largest eigenvalues, whose eigenvectors measure reduced temperature t and the applied field h do not decay substantially under coarsening. Further FIM eigenvalues shrink by a factor of $\sqrt{2}^{-2-y_i}$, where y_i is the i^{th} RG exponent (section 3.8.3). This shrinkage quantifies the information lost in each coarsening step.

In the models examined here, there is a clear distinction between the short time or length scale of the microscopic theory, and the long time or length scale of observables. As we showed, sloppiness in these systems can be precisely traced to the ratio of these two scales—an important small variable. On the other hand, in many other areas of science such a distinction of scales cannot be made. However, the striking similarity in the parameter sensitivities to those in physics lends perspective to the surprising power of mathematical modeling despite parameter uncertainty.

3.7 Acknowledgements

The research described in this chapter was done in collaboration with B. Machta and directed by J. Sethna. Helpful comments and discussions with M. Transtrum, S. Kuehn and S. Papanikolaou are hereby also gratefully acknowledged. The project was financially supported in part by NSF grants DMR 1005479 and DMR 1312160, and a Lewis-Sigler Fellowhip (B. Machta).

3.8 Supplementary Information

3.8.1 Fisher Information for Discrete Diffusive Hopping

To calculate the density of particles at position j and time τ, $\rho_\tau(j)$, it is useful to introduce the Fourier transform of the hopping rates, as well as the Fourier transform of the particle density

at time τ

$$\begin{aligned} \tilde{\theta}^k &= \sum_{\mu=-N}^{N} e^{-ik\mu}\theta^\mu, \\ \tilde{\rho}_\tau^k &= \sum_{j=-\infty}^{\infty} e^{-ikj}\rho_\tau(j), \\ \rho_\tau(j) &= \tfrac{1}{2\pi}\int_{-\pi}^{\pi} dk e^{ikj}\tilde{\rho}_\tau^k. \end{aligned} \tag{3.2}$$

In a time step the density distribution is convolved by the hopping rates. In Fourier space, this is written as[3]

$$\tilde{\rho}_\tau^k = \tilde{\theta}^k \tilde{\rho}_{\tau-1}^k. \tag{3.3}$$

Initially, all particles are at the origin $\rho_0(j) = \delta_{j,0}$, hence $\tilde{\rho}_0^k \equiv 1$ and

$$\begin{aligned} \tilde{\rho}_\tau^k &= (\tilde{\theta}^k)^\tau, \\ \rho_\tau(j) &= \tfrac{1}{2\pi}\int_{-\pi}^{\pi} dk e^{ikj}(\tilde{\theta}^k)^\tau. \end{aligned} \tag{3.4}$$

The Jacobian and metric at time τ can now be written

$$\begin{aligned} J_{j\mu}^\tau &= \partial_\mu \rho_\tau(j) = \tfrac{\tau}{2\pi}\int_{-\pi}^{\pi} dk e^{ik(j-\mu)}(\tilde{\theta}^k)^{\tau-1}, \\ g_{\mu\nu}^\tau &= \tfrac{\tau^2}{2\pi}\int_{-\pi}^{\pi} dk e^{ik(\mu-\nu)}(\tilde{\theta}^k)^{\tau-1}(\tilde{\theta}^{-k})^{\tau-1}. \end{aligned} \tag{3.5}$$

Note that the metric now depends on θ. The preceding formulae were used to calculate the sloppy spectrum of the fig. 3.2. After many steps, the three stiffest eigendirections of $g_{\mu\nu}$ become the three terms in the diffusion equation as discussed next.

The late time behavior of $g_{\mu\nu}^\tau$ is dominated by small k values appearing in the integrand of equation 3.5. For small values of k

$$\begin{aligned} \tilde{\theta^k} &= (1+R)(1 - ikV - \tfrac{k^2}{2}(D+V^2)) + \mathcal{O}(k^3) \\ &= (1+R)\exp(-ikV - D\tfrac{k^2}{2}) + \mathcal{O}(k^3), \\ R &= \textstyle\sum_\mu \theta^\mu - 1 \\ V &= \tfrac{1}{1+R}\textstyle\sum_\mu \mu\theta^\mu, \\ D &= \tfrac{1}{1+R}\textstyle\sum_\mu \mu^2\theta^\mu - V^2. \end{aligned} \tag{3.6}$$

[3]This is due to the convolution theorem. For example, see [61]

In the preceding, note that the first two equations are identical up to second order in k, R is the particle creation rate, V is the drift, and D is the diffusion constant. For the case where the drift $V = 0$ and particle creation rate $R = 0$, at late times

$$\begin{aligned} g^{\tau}_{\mu\nu} &\approx \tfrac{\tau^2}{2\pi} \int_{-\infty}^{\infty} dk e^{ik(\mu-\nu)} e^{-D\tau k^2} \\ &\sim \tfrac{\tau^2}{(D\tau)^{1/2}} e^{-(\mu-\nu)^2/4D\tau}. \end{aligned} \tag{3.7}$$

Expanding this in powers of the small parameter $(\mu - \nu)^2/D\tau$ gives

$$\begin{aligned} g^{\tau}_{\mu\nu} &\sim \tau^2((D\tau)^{-1/2} - (D\tau)^{-3/2}(\mu-\nu)^2/4 + \cdots) \\ &= \tau^2 \sum_{n=0}^{\infty} \tfrac{(-1)^n(\mu-\nu)^{2n}}{n!(4D\tau)^{n+1/2}}. \end{aligned} \tag{3.8}$$

Each term in the series contributes a single new non-zero eigenvalue which scales like

$$\lambda_n \sim \tau^2 \left(\frac{D\tau}{N^2}\right)^{-n-1/2}, n \geq 0. \tag{3.9}$$

The corresponding eigenvectors are best understood by considering their projection onto the observables and are proportional to the left singular vectors of J as $v_{L,n} = (1/\lambda_n) J_{i\mu} v^{\mu}_n$. These are exactly the Hermite polynomials multiplied by a Gaussian with width $2\sigma = \sqrt{D\tau}$. Thus at late times, when the Gaussian goes to a constant in the range $-N$ to N, the stiffest eigendirection is proportional to the non-conservation of particle number $R = \sum_{\mu} \theta^{\mu} - 1$, the second measures drift $V = \frac{1}{1+R} \sum_{\mu} \mu\theta^{\mu}$, and next is the diffusion constant, D. The next terms are less familiar; those past $n = 2$ never appear in a continuum description, because they are always harder to observe than the diffusion constant by a factor of the ratio of the observation scale ($\sqrt{D\tau}$) to the microscopic scale (N) raised to a negative integer power. It is not possible for the diffusion constant, as defined here, to be zero while any higher cumulants are non-zero, explaining why though drift and the diffusion constant both appear in continuum limits, the physical parameter that describes the third cumulant does not. The next eigendirection measures the skew of the resulting density distribution, while the next one measures the distribution's kurtosis, and so on. It is worth noting that careful observation of a

particular θ^μ, somewhat analogous to knowing the bond-angle of a water molecule, would give very little insight on the relevant observables. The exact eigenvalues, measured at steps $\tau = 1$–7 are plotted in fig. 3.2 for an $N = 3$ (seven parameter) model where $\theta^\mu = 1/7$ for all μ.

3.8.2 Measuring the Ising metric

The 2d square lattice Ising model discussed here has lattice sites $1 < i, j < L$, and degrees of freedom $s_{i,j}$ taking the values of ± 1. The probability of observing a particular configuration on the whole lattice (denoted by $\{s_{i,j}\}$) is defined by a Hamiltonian $H\{s_{i,j}\}$ that assigns each configuration of spins an energy (equation A.10). The usual nearest neighbor Ising model has two parameters: a coupling strength J, and a magnetic field h defined through the equation

$$H(\{s_{i,j}\}) = J\sum_{i,j}(s_{ij}s_{ij+1} + s_{ij}s_{i+1j}) + h\sum_{i,j} s_{ij}. \tag{3.10}$$

The Ising model discussed here generalizes this to a larger dimensional space of possible models by including in its Hamiltonian the magnetic field θ^0, the usual nearest neighbor coupling term, and 12 other nearby couplings parameterized by $\theta^{\alpha\beta}$. Vertical and horizontal couplings are also allowed to be different. In the form of equation A.10

$$\begin{aligned} H(x) = &\sum_{\alpha,\beta} \theta^{\alpha\beta}\Phi_{\alpha\beta}\left(\{s_{i,j}\}\right) + \theta^h\Phi_h\left(\{s_{i,j}\}\right), \\ \Phi_{\alpha\beta}\left(\{s_{i,j}\}\right) = &\sum_{i,j} s_{ij}s_{i+\alpha j+\beta}, \\ \Phi_h\left(\{s_{i,j}\}\right) = &\sum_{i,j} s_{ij}. \end{aligned} \tag{3.11}$$

As discussed next, the FIM of this model is calculated along the line through parameter space that describes the usual Ising model ($\theta^{01} = \theta^{10} = J$ and $\theta^{\alpha\beta} = 0$ otherwise) with no magnetic field ($\theta^h = 0$).

From equation A.11, the metric for the generalized Ising model, evaluated at the nearest-neighbor standard zero-field point, can be written in terms of expectation values of

observables as follows (except where necessary, the indices $\alpha\beta$ and h are condensed into a single μ)

$$g_{\mu\nu} = \partial_\mu \partial_\nu \log z = \langle \Phi_\mu \Phi_\nu \rangle - \langle \Phi_\mu \rangle \langle \Phi_\nu \rangle . \tag{3.12}$$

Furthermore, given a configuration $x = \{s_{i,j}\}$, $\Phi_\mu(x)$ is just a particular two point correlation function (or the total sum of spins for Φ_h) [4]. The Wolff algorithm [68] was employed to generate an ensemble of configurations $x_p = \{s_{i,j}\}_p$, for $1 < p < M$, for systems with $L = 64$ to estimate the distribution defined in equation 3.12. (Results were checked against exact enumeration of all possible states on lattices up to $L = 4$.) Thus, for an ensemble of M lattice configurations x_i

$$g_{\mu\nu} = \frac{1}{M^2 - M} \sum_{p,q=1, p\neq q}^{M} \Phi_\mu(x_p)\Phi_\nu(x_p) - \Phi_\mu(x_q)\Phi_\nu(x_p). \tag{3.13}$$

The results are plotted in fig. 3.4. Away from the critical point in the high temperature phase (small βJ), the results seem somewhat analogous to those found for the diffusion equation viewed at its microscopic scale. All of the parameter eigendirections that control two spin couplings ($\theta^{\alpha\beta}$) are roughly of similar distinguishability. However, as the critical point is approached, the system becomes extremely sensitive both to θ^h and to a certain combination of the $\theta^{\alpha\beta}$ parameters. This divergence has been previously shown for the continuum Ising universality class [66] and for the nearest neighbor Ising model [69]. As discussed in the next section, these two metric eigenvalues diverge with the scaling of the susceptibility ($\chi \sim \xi^{7/4}$, whose eigenvector is θ^h) and specific heat ($C \sim \log(\xi)$, whose eigenvector is a combination of $\theta^{\alpha\beta}$ proportional to the gradient of the critical temperature, $\partial T_c / \partial \theta^{\alpha\beta}$). From an information theoretic point of view, these two parameter combinations seem to become particularly easy to measure near the critical point because the system's behavior becomes extremely sensitive

[4] $\Phi_h(\{s_{i,j}\}) = \sum_{i,j} s_{i,j}$ is efficiently calculated for a given configuration $\{s_{i,j}\}$. $\Phi_{\alpha\beta}(\{s_{i,j}\})$ is less trivial: one defines the translated lattice $s'_{i,j}(\alpha,\beta) = s_{i+\alpha,j+\beta}$, in terms of which we write $\Phi_{\alpha\beta}(\{s_{i,j}\}) = \sum_{i,j} s_{i,j} s'_{i,j}(\alpha,\beta)$.

to changes in field and temperature. The behavior of these two eigenvalues seems to have no parallel in the diffusion equation viewed at its microscopic scale.

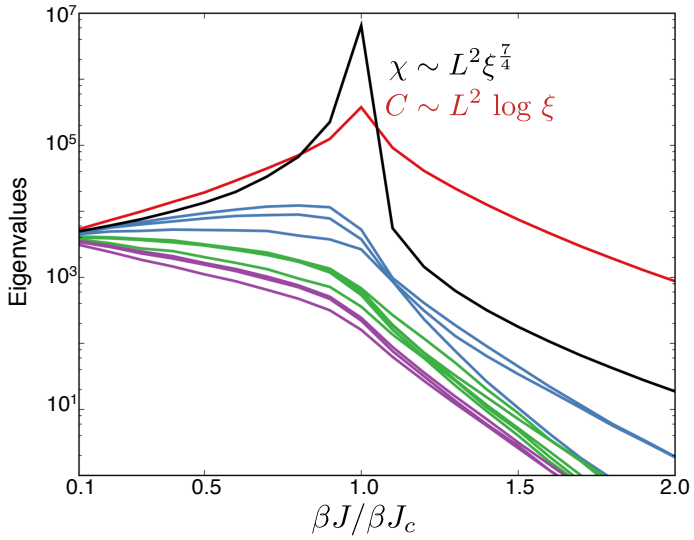

Figure 3.4: Eigenvalues of the FIM versus J/J_c. The enlarged 13 parameter Ising model of size $L = 64$ is described in the text. Magnetic field h is taken to be zero. Two eigenvalues become large near the critical point, each diverging with characteristic exponents describing the divergence of the susceptibility and specific heat respectively. The other eigenvalues vary smoothly as the critical point is crossed. Furthermore they take a characteristic scale determined by the system size and are not widely distributed in log. (In the phase separated region, $\beta J > \beta J_c$ we use the connected correlation function in calculating g_{00}. This corresponds to calculating eigenvalues in 'infinitesimal field'. It allows calculation of the FIM in the phase but arbitrarily close to the phase boundary at which there is a net spontaneous magnetization. Without this the FIM would have one spuriously large eigenvalue, quantifying the large symmetry breaking affect of an arbitrarily small applied field.)

3.8.3 Scaling analysis of the Ising eigenvalue spectrum

Monte Carlo results were also analyzed with renormalization group (RG) techniques focusing on the critical region, close to the RG fixed point θ_0. After an RG transformation that reduces lengths by a factor of b, the remaining degrees of freedom are described by an effective theory

with parameters θ' related to the original ones by the relationship $\theta'^{\mu} - \theta_0^{\mu} = T_{\nu}^{\mu}(\theta^{\nu} - \theta_0^{\nu})$ where[5] T has left eigenvectors and eigenvalues given by $\mathbf{e}_{\alpha,\mu}^{L}$ and b^{y_α}. It is convenient to switch to the so-called scaling variables, $u_\alpha = \sum_\mu \mathbf{e}_{\alpha,\mu}^{L}\theta^{\mu}$, which have the property that under a renormalization group transformation

$$u'_\alpha = b^{y_\alpha} u_\alpha. \tag{3.14}$$

It is also convenient to separate the free energy into a singular part and an analytic part so that

$$\begin{aligned} F(\theta) = &\quad Af^s(u_\alpha(\theta)) + Af^a(u_\alpha(\theta)), \\ f^s = &\quad u_1^{d/2y_1}\mathcal{U}(r_0, ..., r_\alpha), \\ r_\alpha = &\quad u_\alpha / u_1^{y_\alpha/y_1}. \end{aligned} \tag{3.15}$$

Here functions f are free energy densities, A is the system size and f^a and $\mathcal{U}$ are both analytic functions of their arguments. Notice that by construction the variables r do not change under an RG transformation: the rescaling of component variables u_α and u_1 cancel. The FIM can be similarly divided into two pieces

$$\begin{aligned} g_{\mu\nu} &= g^s_{\mu\nu} + g^a_{\mu\nu} = -A\partial_\mu\partial_\nu f^s - A\partial_\mu\partial_\nu f^a, \\ \\ g^s_{\mu\nu} &= A\textstyle\sum_{\alpha,\beta}(\frac{\partial r_\alpha}{\partial\theta^\mu}\frac{\partial r_\beta}{\partial\theta^\nu})(\frac{\partial}{\partial r^\alpha}\frac{\partial}{\partial r^\beta}\mathcal{U}) \\ &= A\textstyle\sum_{\alpha,\beta}(\frac{\partial u_\alpha}{\partial\theta^\mu}\frac{\partial u_\beta}{\partial\theta^\nu})u_1^{-(y_\alpha+y_\beta-d)/y_1}(\frac{\partial}{\partial r^\alpha}\frac{\partial}{\partial r^\beta}\mathcal{U}) \\ &= A\sum_{\alpha,\beta}(\frac{\partial u_\alpha}{\partial\theta^\mu}\frac{\partial u_\beta}{\partial\theta^\nu})(\frac{\partial}{\partial r^\alpha}\frac{\partial}{\partial r^\beta}\mathcal{U})\xi^{y_\alpha+y_\beta-d}, \\ \\ g^a_{\mu\nu} &= A\textstyle\sum_{\alpha,\beta}(\frac{\partial u_\alpha}{\partial\theta^\mu}\frac{\partial u_\beta}{\partial\theta^\nu})\frac{\partial}{\partial u_\alpha}\frac{\partial}{\partial u_\beta}f^a \end{aligned} \tag{3.16}$$

where ξ is the correlation length, which diverges like u_1^{-1/y_1}. By using the dimensionless r variables for analysis of $g^s_{\mu\nu}$, the singular behavior is isolated and expressed in powers of ξ. Now f^a is by assumption an analytic function at the critical point, and the coordinate

[5] Note that in this part of the analysis, the parameter vector θ is infinite dimensional, representing not only long-range couplings but also multi-spin interactions and fields.

changes $u_\alpha(\vec{\theta})$ are analytic there, so $g^a_{\mu\nu}$ will have eigenvalues that are all of the same order of magnitude, given by the area A:

$$\lambda^a_i \sim A. \tag{3.17}$$

The singular behavior of $g^s_{\mu\nu}$ as the correlation length $\xi \to \infty$ at the critical point controls its eigenvalues. As shown in Appendix A, its eigenvalues scale as

$$\lambda^s_i \sim A\xi^{2y_i - d}. \tag{3.18}$$

Hence the singular piece will dominate wherever $2y_i - d \geq 0$. In the 2d Ising model, this is true for the magnetic field as it becomes the largest eigenvector $e_0 = \theta^h$ (with $y_h = 15/8$) along with $e_1 = \partial_\mu u_1$ whose RG exponent is $y_1 = 1$ (in the latter case $2y_i - d = 0$ so there is a logarithmic divergence, as with the Ising model's specific heat). The remaining eigenvectors of $g_{\mu\nu}$ are dominated by analytic contributions. These analytic contributions, just as in the diffusion equation viewed at its fundamental scale, cause the corresponding eigenvalues to cluster together at a characteristic scale and not exhibit sloppiness (though not necessarily to be exactly the identity). This analysis agrees with the Monte Carlo results plotted in fig. 3.4.

3.8.4 Measuring the Ising metric after coarsening

The FIM after n steps of coarsening is $g_{\mu\nu} = -\langle \partial_\mu \partial_\nu \log\left(P(x^n)\right)\rangle$ where $x^n = \{s_{i,j}\}_{\text{for } \{i,j\} \text{ in level } n}$. The levels are defined as follows: If n is even then $\{i,j\}$ is in level n iff $i/2^{n/2}$ and $j/2^{n/2}$ are both integers. If n is odd than $\{i,j\}$ is in level n if and only if $\{i,j\}$ is in level $n-1$ and $(i+j)/2^{n/2+1}$ is an integer[6]. The mapping to level n from level 0 (giving the configuration of

[6]The first level is thus a checkerboard, the second has only even sites, the third has a checkerboard of even sites, etc.

retained subset of spins) is denoted[7] as $x^n = C^n(x)$. It will be useful to write $P(x^n)$ in terms of a restricted partition function

$$\begin{aligned} P(x^n) = &\quad \tilde{Z}(x^n)/Z, \\ \tilde{Z}(x^n) = &\quad \sum_x \exp(-H(x))\delta(C^n(x) = x^n). \end{aligned} \tag{3.19}$$

Here $\tilde{Z}(x^n)$ is the coarse-grained partition function conditioned on the sub-lattice at level n taking the value x^n, summing over the remaining degrees of freedom. The expectation value of an operator defined at level 0 over configurations which coarsen to the same configuration x^n will be denoted as

$$\{Q\}_{x^n} = \frac{\sum_x Q(x)\delta(C^n(x) = x^n)\exp(-H(x))}{\tilde{Z}(x^n)}. \tag{3.20}$$

$\tilde{Z}(x^n)$ can be treated like a partition function in the usual way. In particular, it is possible to take parameter derivatives of the log of $\tilde{Z}(x^n)$ yielding familiar equations for cumulants

$$\begin{aligned} -\partial_\mu \log(\tilde{Z}(x^n)) &= \{\Phi^\mu\}_{x^n} \\ \partial_\mu\partial_\nu \log(\tilde{Z}(x^n)) &= \{\Phi^\mu\Phi^\nu\}_{x^n} - \{\Phi^\mu\}_{x^n}\{\Phi^\nu\}_{x^n}. \end{aligned} \tag{3.21}$$

The calculation will also use nested brackets wherein an outer triangular bracket refers to an expectation value over microscopic configurations and inner curly brackets denote an expectation value in the set of configurations that coarsen to the same x^n. Importantly, a single curly bracket nested in a triangular bracket does not affect expectation values, as every micro state x appears the same number of times in total. However, the presence of two curly brackets in the same one does. For example:

$$\begin{aligned} \langle\{\Phi^\mu\Phi^\nu\}_{x^n}\rangle &= \langle\Phi^\mu\Phi^\nu\rangle \\ \langle\{\Phi^\mu\}_{x^n}\{\Phi^\nu\}_{x^n}\rangle &\neq \langle\Phi^\mu\Phi^\nu\rangle \end{aligned} \tag{3.22}$$

[7] The mapping $C^n(x)$ here simply discards all of the spins that do not remain at level N, leaving an $L/2^{n/2} \times L/2^{n/2}$ square lattice for even N and a rotated 'diamond' lattice for odd N. However, this formalism would also apply to other schemes, such as the commonly used block-spin procedure.

With these the FIM can be written as

$$\begin{aligned} g^n_{\mu\nu} &= -\partial_\mu\partial_\nu \big\langle \log\left(P(x^n)\right)\big\rangle \\ &= \partial_\mu\partial_\nu \log(Z) - \big\langle \partial_\mu\partial_\nu \log(\tilde{Z}(C^n(x)))\big\rangle \\ &= g_{\mu\nu} - \big\langle \big\{\Phi_\mu\Phi_\nu\big\}_{C^n(x)}\big\rangle + \big\langle \big\{\Phi_\mu\big\}_{C^n(x)}\big\{\Phi_\nu\big\}_{C^n(x)}\big\rangle \\ &= \big\langle \big\{\Phi_\mu\big\}_{C^n(x)}\big\{\Phi_\nu\big\}_{C^n(x)}\big\rangle - \big\langle \big\{\Phi_\mu\big\}_{C^n(x)}\big\rangle\big\langle \big\{\Phi_\nu\big\}_{C^n(x)}\big\rangle. \end{aligned} \tag{3.23}$$

Going from the first to the second line uses equation 3.19, going from the second to the third uses equation 3.21 and going from the third to the fourth uses equation 3.22. The quantity $\big\langle \big\{\Phi_\mu\big\}_{C^n(x)}\big\{\Phi_\nu\big\}_{C^n(x)}\big\rangle$ can be measured by taking each member of an ensemble, x_q, and generating a sub-ensemble of $x'_{q,r}$ according to the distribution defined by

$$P(x'_{q,r}|x_q) = \frac{\sum_x \exp(-H(x))\delta(C^n(x'_{q,r}) = C^n(x_q))}{\tilde{Z}(C^n(x_q)))}. \tag{3.24}$$

Techniques for generating this ensemble, using a form of 'Compatible Monte Carlo' [67] are discussed in section 3.8.6. From an ensemble of M configurations, with x_q taken from the ensemble of full lattice configurations, and $x_{q,r}$ from the ensemble given by $P(x'_{q,r}|x_q)$ for each x_q, the metric becomes

$$g^n_{\mu\nu} = \tfrac{1}{(M)(M'^2-M')} \sum_{q,r,s=1 r\neq s}^{q=M\ r,s=M'} \Big(\Phi_\mu(x'_{q,r})\Phi_\nu(x'_{q,s}) - \tfrac{1}{M-1}\sum_{p=1\ p\neq q}^{M} \Phi_\mu(x'_{q,r})\Phi_\nu(x'_{p,s})\Big). \tag{3.25}$$

The results of this Monte Carlo are presented for a 64×64 system at its critical point in fig. 3.3. The analytic corrections to scaling are reduced under coarse-graining, revealing a sloppy spectrum of marginal and irrelevant metric eigenvalues. These irrelevant and marginal eigenvalues continue to behave much as the eigenvalues of the metric in the diffusion equation, becoming progressively less important under coarsening with characteristic eigenvalues. The large eigenvalues are dominated by singular corrections and do not become smaller under coarsening, presumably because they are measured by their collective effects on the large scale behavior measured from large distance correlations.

3.8.5 EIGENVALUE SPECTRUM AFTER COARSE-GRAINING

The scaling of the FIM's eigenvalues after coarsening can be estimated by using an RG-like procedure that uses the following steps: (a) discarding the information in certain degrees of freedom, (b) constructing an effective Hamiltonian for the remaining degrees of freedom in a new parameter basis, (c) repeating the analysis for the metric's eigenvalues in the parameter coordinates of this new effective Hamiltonian, and (d) transforming back into the original coordinates. It is helpful to contrast this approach to a usual RG calculation for a lattice Ising model. In a usual RG calculation, information about certain degrees of freedom is discarded as in (a) and, just as in (b), an effective theory is built that describes the behavior of the remaining degrees of freedom. The approach described below departs from this usual picture in that the goal is not to find this effective theory, but instead to calculate parameter sensitivities of the original microscopic theory. To this end, steps (c) and (d) are added; the effective theory is used only as an intermediate in calculating parameter sensitivities of the original model.

After coarse-graining n times, each observation yields only the spins $\{i, j\}$ remaining at level n, $x^n = \{s_{i,j}\}\Big|_{\{i,j\} \text{ in level n}}$. The probability of a given configuration of these spins x^n can be written in terms of a renormalized model as is typical in RG

$$P(x^n) = \frac{\exp\left(-H^n(x^n)\right)}{Z(A^n, u^n)}, \tag{3.26}$$

where H^n is an effective Hamiltonian describing just those spins that are observable after n coarse-graining steps. H^n has new parameters that can be expressed in terms of the scaling variables defined in equation 3.14 with $u_\alpha^n = b^{y_\alpha n} u_\alpha$. In addition, the area A of the system, in lattice spacings, is reduced to[8] $A^n = b^{-dn} A$, $\partial u_\alpha^n / \partial \theta^\mu = b^{y_\alpha} \partial u_\alpha / \partial \theta^\mu$.

After rescaling, the entropy of the model is smaller by an amount ΔS^n from the original

[8] here, $b = \sqrt{2}$, $d = 2$

model's entropy. It is customary in RG analysis to subtract this constant from the Hamiltonian, so as to preserve the free energy of the system after rescaling:

$$F^n = F^{n,s} + F^{n,a} + \Delta S^n = F^s + F^a = F \tag{3.27}$$

The new model's Hamiltonian is still linear in new parameters, allowing us to use the algebra of equation A.11 if we remove the constant ΔS from the new Hamiltonian. This would, of course, be an identical model, since the addition of a constant to the free energy does not change any observables. Now expressing the metric for the new observables in terms of the original parameters yields

$$g^n_{\mu\nu}(\theta) = \partial_\mu \partial_\nu (F^{n,s} + F^{n,a}) = \partial_\mu \partial_\nu (F^s + F^a - \Delta S). \tag{3.28}$$

Analyzing the singular and analytic contributions to the FIM separately

$$\begin{aligned} g^{s,n}_{\mu\nu} &= \partial_\mu \partial_\nu F^{n,s} = \partial_\mu \partial_\nu F^s = g^s_{\mu\nu}, \\ g^{a,n}_{\mu\nu} &= \partial_\mu \partial_\nu F^{n,a} = b^{-dn} A \partial_\mu \partial_\nu f^{n,a} \\ &= b^{-dn} A \frac{\partial u^n_\alpha}{\partial \theta^\mu} \frac{\partial u^n_\beta}{\partial \theta^\nu} \left(\frac{\partial}{\partial u^n_\alpha} \frac{\partial}{\partial u^n_\beta} f^{n,a}\right) \\ &= A \sum_{\alpha,\beta} b^{(y_\alpha + y_\beta - d)n} \left(\frac{\partial u_\alpha}{\partial \theta^\mu} \frac{\partial u_\beta}{\partial \theta^\nu}\right) \left(\frac{\partial}{\partial u_\alpha} \frac{\partial}{\partial u_\beta} f^a\right) \end{aligned} \tag{3.29}$$

The singular piece of the metric is maintained *exactly* because the singular part of the free energy is preserved after an RG step. The implication is that the singular part of the free energy contains long wave-length information. On the other hand, the analytic piece is smaller by $\partial_\mu \partial_\nu \Delta S^n$. The matrix $\left(\frac{\partial u_\alpha}{\partial \theta^\mu} \frac{\partial u_\beta}{\partial \theta^\nu}\right)\left(\frac{\partial}{\partial u_\alpha} \frac{\partial}{\partial u_\beta} f^a\right)$ should be smoothly varying, with n, depending only the u^n which vary only small amount with n near the RG fixed point. Importantly, all of its eigenvalues should continue to take a characteristic value. Thus, after rescaling n times (see equations 3.17, 3.18 and Appendix A.2)

$$\begin{aligned} \lambda^{n,s}_i &\sim A(\xi)^{2y_i - d}, \\ \lambda^{n,a}_i &\sim A b^{n(2y_i - d)}. \end{aligned} \tag{3.30}$$

To ensure that the Fisher information is strictly decreasing in every direction upon coarsening[9], $g^a_{\mu\nu}$ must be negative semidefinite in the subspace of scaling variables where $2y_i - d > 0$. For these relevant directions, $\lambda_i^n \sim A\xi^{2y_i-d} - Ab^{2y_i-d}n$, with $i = 0,\ 1$. Here, the second term only becomes significant when $b^n \sim \xi$ (i.e. when the lattice spacing is comparable to the correlation length). For irrelevant directions, or relevant ones with $0 < 2y_i < d$ (corresponding to $i \geq 2$ in the Ising model), the analytic piece will eventually dominate as the critical point is approached, yielding $\lambda_i \sim Ab^{2y_i-d}$. These results are in quantitative agreement with those plotted in fig. 3.3 assuming that the variables project onto irrelevant and marginal scaling variables with leading dimensions of $y = 0$ (blue line in fig. 3.3), $y = -2$ (green line) and $y = -4$ (purple line) consistent with the theoretical predictions for the irrelevant eigenvalue spectrum made in [70].

This shrinkage of the FIM is reparameterization invariant in an important way. Although a coordinate system can always be chosen in which the metric is locally the identity, the shrinkage, which can be seen in any coordinate system, quantifies the contraction of the invariant distance between nearby points as observables are coarsened. For example, if we choose a coordinate system in which the metric is the identity when examining microscopic observables, we find that the metric eigenvalues become widely spread after coarsening[10].

It is helpful to contrast the results of this information geometry analysis to those of a more standard RG one. Both can be used to explain the experimental findings of universality: a wide class of microscopic models have identical macroscopic behavior. In an RG picture, one

[9] In each coarsening step $g^n_{\mu\nu} - g^{n+1}_{\mu\nu}$ must be a positive semidefinite matrix. This is because no parameter combinations can be more measurable from a subset of the data available at level n than from its entirety.

[10] Least-Squares models that do not have a concept of coarsening still have a reparameterization invariant manifestation of sloppiness [4, 57]. These models are typically finite in extent, at least in most directions and contain 'edges' where some metric eigenvalues are zero and where parameters take extreme values (for example a rate constant being either zero or infinity). Although a coordinate change can locally set the metric to the identity, the reparameterization invariant shape of the manifold has a 'hyper-ribbon' structure, with a geometric hierarchy of widths. It is unknown if the Ising model has a similar structure on coarsening.

considers a hypothetical large dimensional space of possible Hamiltonians that includes microscopically disparate systems (for example including both ferromagnets and binary fluids). As the renormalization group proceeds the Hamiltonians of their effective models flow towards the same saddle point. The Hamiltonian of this saddle point thus describes the effective interactions of coarsened degrees of freedom. This explains how binary fluids and ferromagnets could have similar effective models for the coarsened observables.

This same hypothetical large dimensional space of Hamiltonians can be considered from an information theory perspective, by adding step (c), calculating the Fisher Information for the effective Hamiltonian and (d), transforming back to microscopic coordinates. Information geometry clarifies that the microscopic Hamiltonians describing binary fluids and ferromagnets produce indistinguishable results for coarsened variables. Although the parameter space distance between microscopic models for binary fluids and ferromagnets is quite large, the 'proper distance' between them defined through the FIM rapidly vanishes upon coarsening. Models for ferromagnets and binary fluids (for which t and h values are identical) differ from each other only along sloppy directions and hence their long-wavelength behaviors become nearly identical. The evolution of the FIM under coarsening tracks the information lost about microscopic details in these physics models. In this information geometry picture of universality, the high dimensional parameter space manifold of systems near Ising critical points collapses onto a two dimensional manifold when its observables are coarsened. This analysis completes what might be seen as a trivial step in RG arguments for universality—demonstrating that nearness in effective model space implies indistinguishability of coarsened observables. The mapping from parameter space distance to metric distance in the space of distinguishability clarifies some confusing points. For example, while FIM distinguishability along relevant parameter directions remain roughly fixed under coarsening, their parameter space distance appears to grow in a usual RG picture. Similarly considering

an enlarged hypothetical parameter space likely explains why many models with sloppy FIMs, for example in systems biology, can be predictive even when important components are entirely absent from their microscopic models.

3.8.6 SIMULATION DETAILS

As described above, $M = 10,000$–$100,000$ independent members from each ensemble x_p are generated using the standard Wolff algorithm [68] implemented on 64×64 periodic square lattices, and are used to calculate the FIM before coarsening.

A variation of the 'Compatible Monte Carlo'[11] method introduced in [67] was employed to generate members of the coarse-grained ensemble defined by equation 3.24. In this method, a Monte Carlo chain is run and any move proposing a switch to a configuration $x'_{p,r}$ for which $C^n(x'_{p,r}) \neq C^n(x_p)$ is rejected. For the mapping $C^n(x_p) = C^n(x_{p,r})$, the simplest implementation equilibrates using Metropolis moves by proposing only the spins not in level n. Additional tricks to speed up convergence are described below.

Consider the task of generating a random member $x'_{p,r}$ for a given x_p at level 1. Because the spins which are free to flip only couple with fixed spins, each one can be chosen independently. As such, choosing each free spin according to its heat bath probability generates an uncorrelated member $x_{p,r}$ of the ensemble defined by x_p in a single step. This idea can be further exploited to exactly calculate the contribution to a metric element at level 1 from a level 0 configuration x. In particular, replacing all of the spins that are not in level 1 with

[11] Ron, Swendsen and Brandt used this technique to generated large equilibrated ensembles close to the critical point, essentially by starting from a small 'coarsened' lattice and iteratively adding layers to generate a large ensemble.

their mean field values defined by $\tilde{s}_{i,j}(x) = \{s_{i,j}\}_{C^n(x)}$ leads to

$$\begin{aligned} \{\Phi_{\alpha\beta}\}_{C^n(x)} &= \sum_{i,j} \tilde{s}_{i,j}(x)\tilde{s}_{i+\alpha,j+\beta}(x), \\ \{\Phi_h\}_{C^n(x)} &= \sum_{i,j} \tilde{s}_{i,j}. \end{aligned} \tag{3.31}$$

It is therefore possible to exactly calculate the level 1 quantities $\{\Phi_\mu\}_{C^1(x)}\{\Phi_\nu\}_{C^1(x)}$ for any microscopic configuration x and the corresponding checkerboard configuration $C^1(x)$. The metric at level 1 can now be written

$$g^1_{\mu\nu} = \tfrac{1}{M^2-M} \sum_{p,q=1,p\neq q}^{M} \Big(\{\Phi_\mu\}_{C^1(x_p)} \{\Phi_\nu\}_{C^1(x_p)} - \{\Phi_\mu\}_{C^1(x_p)} \{\Phi_\nu\}_{C^1(x_q)} \Big). \tag{3.32}$$

Beyond level 1 it becomes necessary to use Compatible Monte Carlo. Because of the independence of free spins at level 1, spins at all levels $n \geq 1$ only interact with spins that are already absent at level 1. Therefore, the spins that are free at level 1 (termed the red sites of the checkerboard) are left integrated out. The partition function for a level 1 configuration is most conveniently written in terms of the number of up neighbors, $n^{up}_{i,j}$ that each red site has

$$\begin{aligned} \log \tilde{Z}(C_1(x)) &= \sum_{i,j \text{ not in level 1}} \log\left(z(n^{up}_{i,j})\right), \\ z(n^{up}) &= \cosh\left((\beta J)(2 - n^{up})\right), \end{aligned} \tag{3.33}$$

Additional spins that are not integrated out at level n are flipped using a heat bath algorithm with the ratio of partition functions in an 'up' vs 'down' configuration used to determine the transition probability. The probability of a spin (at level ≥ 2) transitioning to 'up' after being proposed from the down state is given by $z^{up}_{i,j}/(z^{up}_{i,j} + z^{down}_{i,j})$ with

$$\begin{aligned} z^{up}_{i,j} &= \sum_{\{k,l\} \text{ n.n. of } \{i,j\}} z(n^{up}_{k,l} + 1), \\ z^{down}_{i,j} &= \prod_{\{k,l\} \text{ n.n. of } \{i,j\}} z(n^{up}_{k,l}). \end{aligned} \tag{3.34}$$

Equilibration is fast as there are effectively no correlations larger than the spacing between fixed spins at level n. This allows generating an ensemble of lattice configurations at level 1, conditioned on the system coarsening to an arbitrary configuration at any level $n > 1$.

Equation 3.25 is thus slightly modified to the following which was used to make fig. 3.3 for data at level 2 and higher

$$g^n_{\mu\nu} = \frac{1}{(M)(M'^2-M')} \sum_{q,r,s=1 r\neq s}^{q=M\ r,s=M'} \left(\{\Phi_\mu\}_{c^1(x'_{q,r})} \{\Phi_\nu\}_{c^1(x'_{q,s})} \quad - \frac{1}{M-1} \sum_{p=1\ p\neq q}^{M} \{\Phi_\mu\}_{c^1(x'_{q,r})} \{\Phi_\nu\}_{c^1(x'_{p,s})} \right) \tag{3.35}$$

CHAPTER 4

CHAPTER 4: CANONICAL SECTORS AND EVOLUTION OF FIRMS IN THE US STOCK MARKETS

4.1 ABSTRACT

A classification of companies into sectors of the economy is important for macroeconomic analysis and for investments into the sector-specific financial indices or exchange traded funds (ETFs). Major industrial classification systems and financial indices have so far largely been developed by empirical methods with questionable objectivity and completeness. In this paper, we show how a broad-level sector decomposition of stocks can be made objectively through a machine learning approach that exploits the emergent low dimensional structure of the space of historical stock price returns. The method automatically identifies emergent "canonical sectors" in the market and assigns every stock a participation weight into each sector. Lastly, by analyzing data from different periods at a time, we show how firms listed in the market have evolved in their decomposition into sectors.

[0]The work described in this chapter is being prepared for publication in a journal as follows: R. Chachra, A. A. Alemi, P. Ginsparg, J. P. Sethna. *Canonical Sectors and Evolution of Firms in the US Stock Markets*. The present author initiated and defined the project, obtained the data, implemented the algorithms, analyzed the results and wrote this report.

4.2 INTRODUCTION

The performance of our economy is often understood in a reductionist way. This entails decomposing the economy into its constituents and then learning how each performed over a given period using the so-called *economic indicators*. These variables measure unemployment rate, housing starts, consumer price index, gross domestic product, etc. allowing for broad macroeconomic analysis and modeling.

In analogy to the broader economy, the performance of the stock markets is reported similarly in terms of aggregated quantities with groups of stocks taken at once. A finer, microlevel analysis quickly becomes impractical because of the plethora of listed stocks—as of this writing, stocks from over 8000 US public companies are available for trading in various markets. These include over 4500 securities listed on the three major domestic US exchanges: NASDAQ, NYSE and NYSE MKT. For convenience of analysis and investment, stocks are grouped into indices such as the market-wide Russell 3000 [71] and S&P 500 [72] comprising stocks from diverse companies to reflect the entirety of the market, and sector-specific indices such as Dow Jones Financials Index [73], CBOE Oil Index [74] and Morgan Stanley High-Tech 35 Index [75] that are more granular indicators of performance in individual named sectors.

In principle, a set of mutually exclusive and collectively exhaustive sector indices could describe the overall stock market as a sum of parts, but for practical applications, this approach is rife with ambiguities. First, to what sector does one assign a conglomerate or diversified company such as General Electric that functions in a variety of businesses across different sectors? Second, how does one account for the participation of non-conglomerates outside their core sectors? For example, if a financial services company is deeply invested in the pharmaceutical sector to an extent that such causal relationship is manifest in strongly correlated returns of the two, should that company be considered part of a financial services'

index or a healthcare one? Third, as economic environments or companies evolve, neither the industrial sectors nor firms' sector association remains static, so how does one account for the dynamic nature of firms comprising an index?

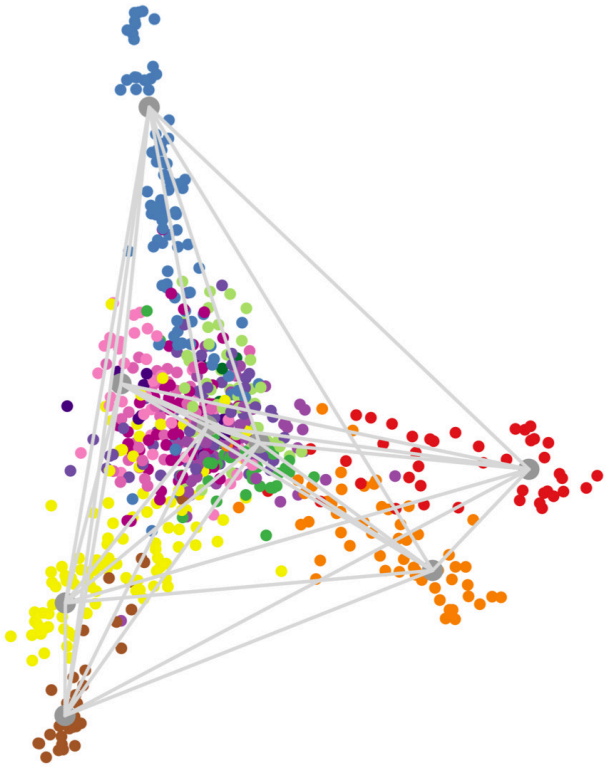

Figure 4.1: Low-dimensional projection of the stock price returns data. Stock price returns are projected onto a plane spanned by two stiff vectors from the SVD of the emergent simplex corners as described in section 4.9.5. Each colored circle corresponds to one of the 705 stocks in the dataset used in the analysis. Colors denote the sectors assigned to companies by Scottrade [11] and the scheme is shown in fig. 4.9. The grey corners of the simplex correspond to sector-defining prototype stocks, whereas all other circles are given by a suitably weighted sum of these grey corners. Projections along other singular vectors are shown in fig. 4.6.

The aforementioned technical issues must be resolved in manner that is grounded with a theoretical framework built to describe the character and properties of the underlying entities. A vast number of studies have previously aimed at finding structure and categories of

stocks in financial markets with a variety of approaches. Recent numerical techniques have included extensive use of the random matrix theory, principal component analysis or the associated eigenvalue decomposition of the correlation matrix [12, 76–79], specialized clustering methods [80–84] or time series analysis [85, 86], pairwise coupling analysis [87], and even topic-modeling of returns [88]. While these methods have yielded important results, a fundamental basis of macro-level analysis, resting upon emergent properties in the markets has so far remained elusive.

In this paper, we demonstrate a new, holistic way of classifying stocks into industrial sectors by utilizing the emergent structure of price returns data space. The method identifies sectors in the market and assigns each stock weights denoting the extent to which its return are comprised of emergent sector returns. Relying purely upon an unsupervised machine learning analysis of historical time series of stock price returns, this method is an objective way of understanding stocks solely through their returns. In the subsequent sections, we show (a) the space of stock price returns has a hyper-tetrahedral (simplex) structure (fig. 4.1), each cell of the simplex is populated by stocks of similar returns time series, (b) the corners of the simplex correspond to emergent "canonical" sectors occupied by stocks of companies that are prototypical (fig. 4.2), every other stock's return decomposes into a weighted sum of returns from the prototypical stocks (fig. 4.3), and (c) the participation weights of the companies are dynamic and provide insights into their evolving nature (fig. 4.4).

4.3 Canonical Sectors and Price Returns

Prior work [12] has made it clear that the high-dimensional space[1] of stock price returns has a low-dimensional representation. This implies that only a few dimensions in the space of price

[1] The dimension of data space is the length of the time series analyzed.

returns have signal and the rest can be ascribed to random noise [89]. The key discovery of the present work is that the low-dimensional representation of stock price returns has a well-defined structure (discussed in section 4.9.5) that leads to new insights about individual stocks and the industrial sectors of the economy. This structure is a hyper-tetrahedron (also known as a simplex) that becomes apparent upon visualizing low-dimensional cross-sections of the data as shown in figs. 4.1 and 4.6. A closer examination of the simplex further reveals that each cell is populated by stocks of companies in similar or related business lines implying that every cell corresponds to an identifiable segment of the economy. Moreover, in the zero-centered simplex, data points located near the origin predominantly correspond to stocks of conglomerates or diversified companies (e.g., GE, Walt Disney, 3M, etc.), whereas the corners of every cell consist of companies that are prototypical of known sectors (Texas Instruments (tech), Wells Fargo (financial), Kohl's (retail), etc.). How many emergent sectors are there in the market? The general problem of selecting a signal to noise ratio cutoff or a truncation threshold in high-dimensional data does not always have a clear answer. As is the case with stock price returns, the threshold is generally sensitive to sampling, but nonetheless reasonably robust for qualitative results (section 4.9.3). The dataset used in this analysis (section 4.9.1) consisted of two decades (1993-2013) of daily price returns from 705 US public companies each with a mid-2013 market capitalization of $1 billion or higher, representing a broad section of the economy in a period marked by major crises (section 4.9.1 has more details). This data set has eight emergent sectors (section 4.9.7) which we name as follows (the prefix *c-* signifies "canonical" and distinguishes these names from listed sectors names more commonly used): *c-cyclical* (including retail), *c-energy* (including oil and gas), *c-industrial* (including capital goods and basic materials), *c-financial*, *c-non-cyclical* (including

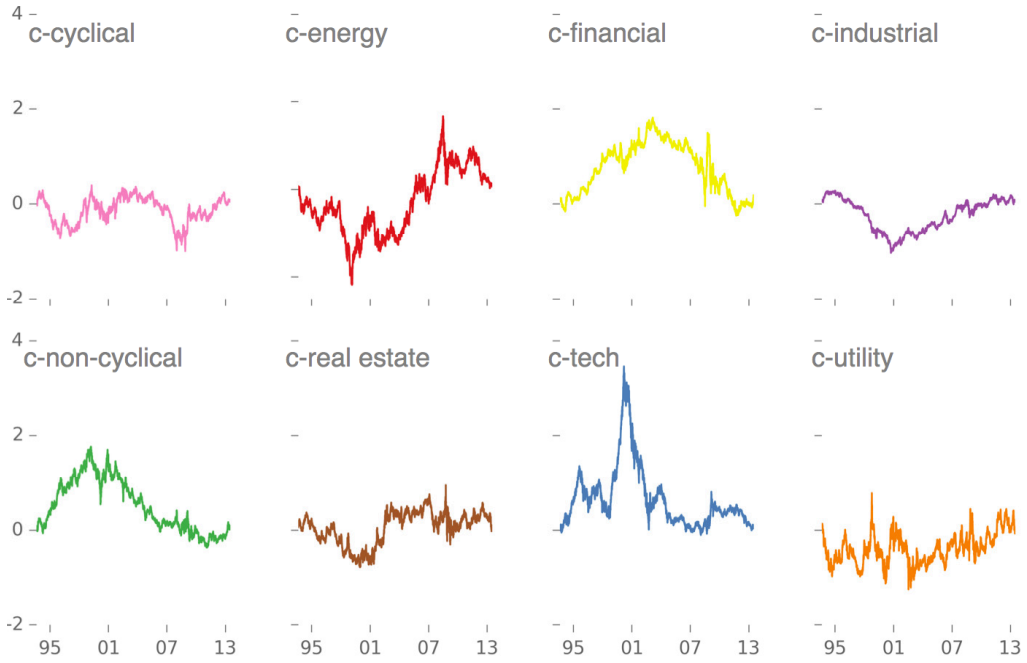

Figure 4.2: Emergent sector time series. Annualized cumulative log price returns of the eight emergent sectors are shown. The time series capture all important features affecting different sectors: dot-com bubble (c. 2000), the energy and financial crises of 2008. Precise definitions is given in equation 4.3; other measures of sector dynamics are in fig. 4.8.

healthcare and consumer non-cyclical goods), *c-real estate*, *c-technology*, and *c-utility*[2]. The prices returns (figs. 4.2 and 4.8) from these sectors show the performance of the different industrial sectors of the economy, including major events that afflicted each such as the few described below.

Dot-com bubble: The building-up of the speculative bubble spanning 1997-2000 and its subsequent crash over two years that followed is clearly seen in the returns of the tech sector. One also sees that the tech bubble was primarily contained within the tech companies' ecosystem with only minor remnants elsewhere.

[2]Major industrial classification schemes (ICB [15], GICS [13], TRBC [14]) discretely separate the broad economy into 10 major sectors roughly as follows: energy, materials, consumer cyclical, consumer non-cyclical, financials, healthcare, industrial, technology, telecom and utilities [90].

Energy crisis: In the period spanning 2003-2008, crude oil price witnessed a four-fold increase (primarily ascribed to disruptions caused by Hurricane Katrina and Iranian nuclear crisis), and then precipitously dropped following the onset of the global recession. Energy stocks also plunged headlong.

Global financial crisis: The financial crisis of 2008 affected the entirety of the market, but had particularly grave implications for the real-estate and the financial sectors.

4.4 Constituent Firms in Canonical Sectors

Canonical sector	Business lines	Prototypical examples
c-cyclical	general and speciality retail, discretionary goods	AutoZone, Kohl's, Nordstrom
c-energy	oil and gas services, equipment, operations	Hess, Schlumberger
c-financial	banks, insurance (except health)	Citigroup, Wells Fargo, M&T Bank
c-industrial	capital goods, basic materials, transport	Dow Chemical Co., Goodyear
c-non-cyclical	consumer staples, healthcare	Pepsi, Procter & Gamble
c-real estate	realty investments and operations	Vornado Realty, Camden Property Trust
c-technology	semiconductors, computers, comm. devices	Intel, Motorola, Oracle
c-utility	electric and gas suppliers	Duke Energy, Edison International

Table 4.1: Canonical sectors and major business lines of primary constituent firms. Examples provided are firms that are strongly associated to these sectors. A full list is available on companion website [10].

As mentioned in the preceding section, eight sectors emerge in analysis of our dataset. Here we list some high-level defining features of each of these sectors. This discussion that follows is summarized in table 4.1.

Firms showing strong association to what we call *c-cyclical* sector include speciality and general retail outlets; well known names include Best Buy, Kohl's, Target, Tiffany, etc. The

canonical sector *c-energy* firms are either integrated oil and gas firms (eg. Exxon), or are involved in operations (eg. Hess), or provide services within this sector (eg. Halliburton). *c-financial* sector firms include large and small banks, all kinds of insurance companies with the notable exception of health insurance firms. Bank of America, Citigroup, Wells Fargo, etc. strongly associate with this emergent sector. The *c-industrial goods* sector firms are involved often specialized large-scale manufacturing of basic materials (paper products, chemicals etc.) or capital goods (machineries); as example, Dow Chemical Company is strongly linked to this sector. The *c-non-cyclical* sector is comprised of consumer staple goods (food, beverage) but also healthcare firms. Coca-Cola, Kellogg, Pfizer, Merck and many other household names are all members of this group. *c-Real estate* sector is almost exclusive linked to firms with heavy real estate operations including real estate investment trusts, insurers, etc. The *c-tech* sector primarily comprised of semiconductor, hardware, software and communication equipment manufacturing firms such as Cisco, Intel, Oracle, Motorola, etc. Core *c-utility* firms are in electric or gas supply business; examples include Duke Energy Corp., Edison International, etc.

4.5 Sector Decomposition of All Firms

Each stock return is modeled as a weighted combination of returns from the canonical sectors. In matrix form this is written as: $R_{ts} = E_{tf} W_{fs}$, where matrices R, E and W contain (normalized log) returns at times t for stocks s, returns of the emergent sectors f, and participation weights respectively. The latter are constrained so that for each stock, the participation weights in multiple sectors add to unity. Calculations are described in the sections 4.9.2 and 4.9.3. Here we discuss important insights in fig. 4.3.

Conglomerates decompose into their core constituents. For example, calculations show

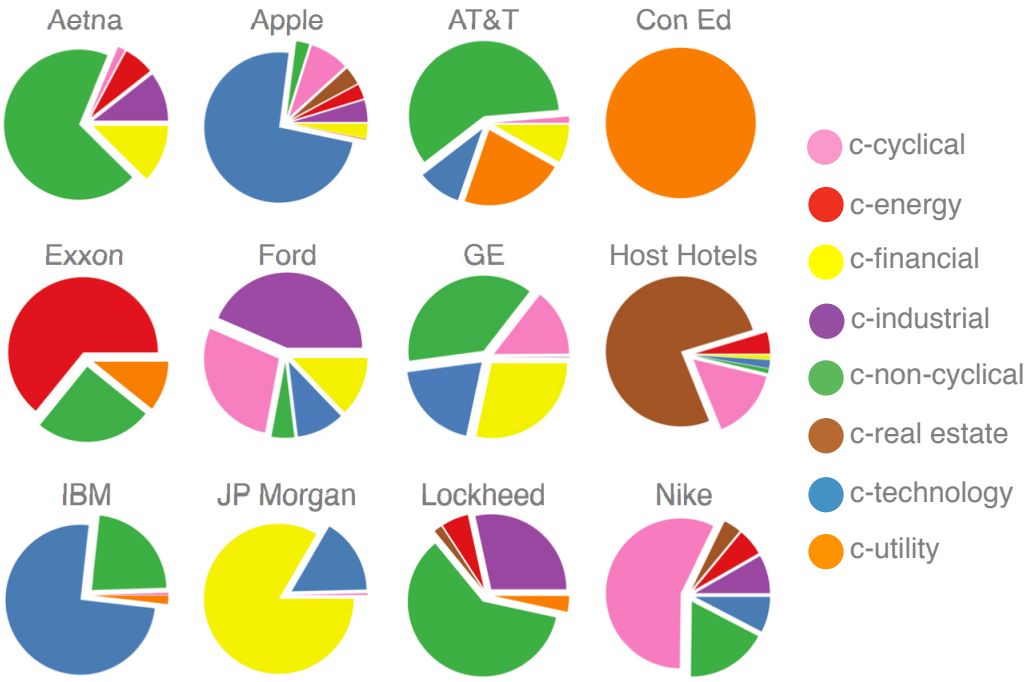

Figure 4.3: Canonical sector decomposition of stocks of selected companies. A complete set of pictures for all 705 stocks is provided on the companion website [10]. Color scheme shown on the right are used in figures throughout the paper except where noted.

that General Electric's returns are comprised of four segments: *c-financials*, *c-non-cyclical*, *c-tech* and *c-cyclical*, while 3M is in the business of *c-industrial* and *c-non-cyclicals*. Technology companies such as Apple that sell mass-market consumer goods also have important fraction in *c-retail* sector in addition to *c-tech*, whereas IBM having significant government contracts and healthcare analytics products has a significant portion of *c-non-cyclical* returns. Telecom companies, for example AT&T and Verizon, are generally classified under a separate major category of their by many classification systems, yet the present analysis shows their returns are described by a combination of *c-non-cyclical* and *c-utility* components. Returns of health insurance providers such as Aetna, United Healthcare, etc. that are commonly classified as financial services firms, are comprised of a major part *c-non-cyclical* and minor part of *c-financial* sector. Defense contractors like Lockheed, Northrop Grumman, Raytheon that are

primarily listed as capital goods companies have their returns comprised of a majority *c-non-cyclical* component and only a smaller share of *c-cyclical* sector.

4.6 Evolution of Sector Weights

The sector decomposition of firms is by no means static. As companies grow, their business foci often change. They may enter or leave different sectors through mergers, acquisitions, spin-offs, new products or target customers. We used the decomposition analysis described above with one-year overlapping windows of time (details in 4.9.3) to get insight into the evolving nature of sector participation of firms.

Major events affecting companies in an idiosyncratic manner show clear signature in this analysis. For example, Corning Inc. not traditionally a tech firm, suffered in the aftermath of the dot-com crisis due to its reliance upon developing products and infrastructure for other tech firms. As such, the company has since then drastically shifted toward *tech* after the bubble burst.

Likewise, a growing company's strategy shift is also seen in the analysis. For example, in the early 1990s, Berry Petroleum grew within its home state of California through development on properties that were purchased in the earlier part of 20th century. In 2003, the company embarked on a transformation [91] by direct acquisition of light oil and natural gas production facilities outside California. Fig. 4.4 shows a clear shift in the distribution of sector weights as the company has moved more squarely toward *c-energy* and away from *c-real estate*. Similarly, as Plum Creek Timber converted to a real estate investment trust (REIT) in the late 1990s [92], its sector weights have also significantly shifted toward *c-real estate* sector as shown.

Lastly, for stable and focused companies such as Pacific Gas & Electric or IBM (fig. 4.4), one sees no significant shifts in sector weights. Wal-Mart's returns, on the other hand, have moved from significantly from *c-cyclical* to *c-non-cyclicals* (consumer staples) in the post-financial crises years as shown. This is also true of other low-price consumer commodities retailers such as Costco, but not true of higher price retailers such as Whole Foods, Macy's, etc.

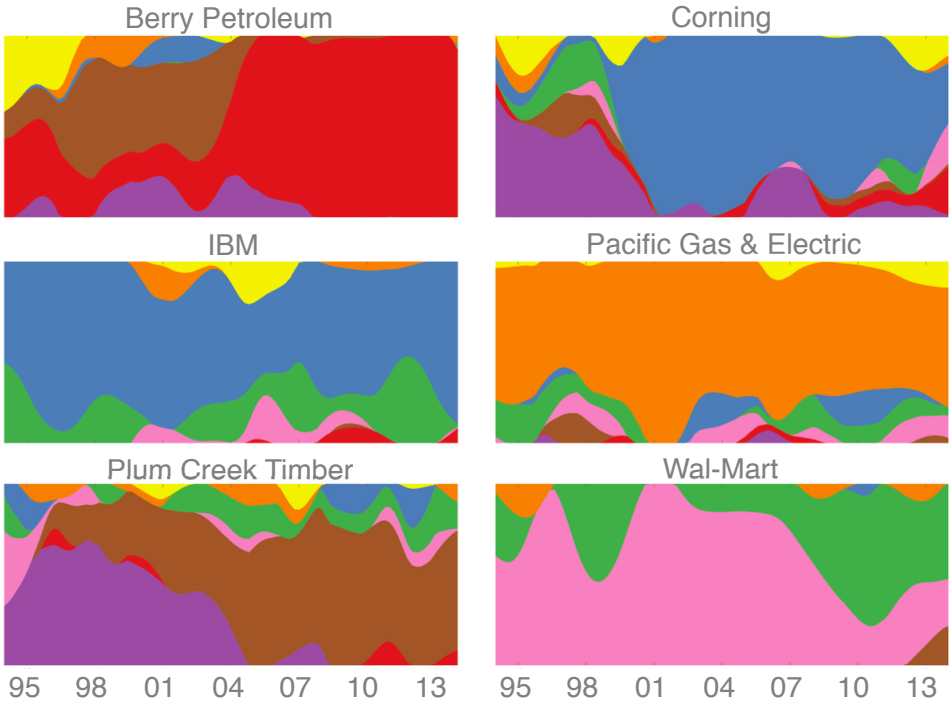

Figure 4.4: Evolving sector participation weights. Results from the sector decomposition made with rolling two-year Gaussian windows are shown for selected stocks. A complete set of 705 pictures is provided on the companion website [10]. Color scheme is as in fig. 4.3

4.7 Discussion

The dataset we analyzed was comprised of daily returns for a 20 year period for 705 US companies with $1 billion or or more in market capitalization. While only a small subset of the

business are publicly traded and even fewer have market caps as high as a billion, our dataset nonetheless represents an excellent segment of our economy by including a broad diversity of firms and the conditions they witnessed in the previous two decades including at least three major domestic crises and their aftershocks.

We first saw that space of stock price returns has a hyper-tetrahedral structure. This structure is inherent in data and has emerged out of a multitude of microscopic interactions (trades) between a plethora of participants. The simplex is not only a low-dimensional manifold representation of this high-dimensional data, but it also has a meaningful sub-structure: Each cell of the simplex is populated by stocks of companies in related businesses, and each corner of the hyper-tetrahedron represents "pure types" of companies that are strongly associated with one individual sector. Stocks populating the center of the tetrahedron are conglomerates or diversified companies.

We further saw that the emergent structure is amenable to a matrix factorization ("archetypal analysis") that identifies the simplex corners as canonical sectors returns and decomposes each stock time series as a weighted sum of returns from the emergent sectors. This decomposition yielded new high-level insights about the nature of stocks returns and their quantifiable participation across sectors, in addition to granular insights about specific firms, revealing their exposure to returns from different sectors of the economy.

We also gained a vivid insight into the evolving character of the sector participation of firms with different windows of time in the last two decades. As firms evolve and become exposed to different industrial sectors, this information is represented in its stock price returns which will show greater correlations with those industrial sectors. Therefore, any sector index should account for the dynamic nature of constituent firms and rebalance the portfolio allocation accordingly.

Future work remains to address survivorship bias, effects of sampling at different frequencies, and incorporating smaller market cap firms. The framework of understanding stock returns via an emergent structure of their data space also suggests development of a generative model. Lastly, investors and governments alike would benefit from the development of new investable sector indices that measure the health of our industrial sectors in a more principled manner as propounded in this study (section 4.9.8).

4.8 Acknowledgements

The research described in this chapter was done in collaboration with A. Alemi and P. Ginsparg, and was directed by J. Sethna. The project was financially supported in part by NSF grants DMR 1312160 and OCI 0926550.

4.9 Supplementary Information

4.9.1 Dataset Particulars

A wealth of financial data is freely available online via multiple sources. For the analysis described in this paper, we obtained names, tickers, listed-sectors and market caps of US-based publicly traded companies from Scottrade [11]. The following criteria were applied to company selection:

- July 2013 market capitalization over $1 billion.
- Registered domicile in US or Caribbean countries.
- Listed for trading on NASDAQ, NYSE, or NYSE MKT (formerly AMEX).

- Continuously traded for 20 years (beginning mid-1993).

The search filters yielded a list of $N = 705$ tickers for which adjusted daily closing prices [3] were obtained from Yahoo! Finance [93] using their API; the rare cases of missing or corrupted data points in the time series were replaced with linear interpolated values. A brief summary of listed sectors and number of companies in each is provided in table 4.2 and a full list of company names, tickers, market caps and listed-sector info is available on the companion website [10].

Listed sector	**Companies**
Basic materials	58
Capital goods	61
Consumer cyclical	41
Consumer non-cyclical	40
Energy	42
Financial (+Real estate)	138
Healthcare	53
Services (+Retail)	101
Technology	93
Telecom	6
Utility	57
Transport	15
TOTAL	**705**

Table 4.2: Listed sectors and number of companies dataset analyzed. Tickers for each company were obtained from [11].

4.9.2 Returns Factorization and Sector Decomposition

The general problem of matrix factorization has received considerable attention in recent years and a variety of factorization algorithms have been developed with the goals of

[3] Prices at the end of every trading day, corrected for stock splits or dividend issues.

dimensional reduction, classification or clustering. Examples include archetypal analysis (AA) [94], heteroscedastic matrix factorization [95], binary matrix factorization [96], K-means clustering [97], simplex volume maximization [98], independent component analysis [99], non-negative matrix factorization (NMF) [100, 101] and its variants such as the semi- and convex-NMF [102], convex hull NMF [103] and hierarchical convex NMF [104], among others. Each method has a unique interpretation [105] and therefore, a successful application of any of these methods is contingent upon the underlying structure of the data.

The hyper-tetrahedral structure of log price returns seen in our analysis motivates a decomposition so that each stock returns is a weighted mixture of canonical sectors:

$$R_{ts} = E_{tf} W_{fs}. \tag{4.1}$$

Columns of E_{tf} are the emergent sector time series (basis vectors) representing the n corners of the hyper-tetrahedron, and W_{fs} are the participation weights ($W_{fs} \geq 0$) in sector f so that $\sum_f W_{fs} = 1$ for each stock s. The first factorization of this kind was developed in 1994 and named "archetypal analysis" (AA) [94], and improvements were proposed more recently [106, 107]. The algorithm reduces dimensionality by representing each sample (here, each stock) as convex combinations of extremes (called archetypes). The archetypes are the columns in the basis matrix E_{tf} and these can be found in multiple ways:

- Minimizing the squared error with convex constraints in factorization as originally proposed [94].
- Making a convex hull of the dataset and choosing one or more of its vertices to be basis vectors, but this method would have serious computational limitations in high-dimensional data.
- Making a convex hull in low-dimensions and choosing one or more of its vertices to be basis vectors [106].

- Minimize after initializing with candidate archetypes that are alternatively guaranteed to lie in the minimal convex set of the data. This technique was proposed in [107].
- Fitting the smallest possible hyper-tetrahedron on the dataset.

In all but the last case above, the archetypes are themselves chosen from the data: $E_{tf} = R_{ts'}C_{s'f}$, such that $\sum_{s'} C_{s'f} = 1$[4]. The columns of the C matrix are shown in figure 4.11.

In sum, AA is defined as a factorization with these properties:

$$\begin{aligned} &R_{ts} \sim R_{ts'}C_{s'f}W_{fs}, \\ &C_{s'f} \geq 0, \;\; \textstyle\sum_{s'} C_{s'f} = 1, \\ &W_{fs} \geq 0, \;\; \textstyle\sum_{f} W_{fs} = 1, \end{aligned} \tag{4.2}$$

in which one minimizes the square of the Frobenius matrix norm: $||R_{ts} - R_{ts'}C_{s'f}W_{fs}||_F^2$.

4.9.3 Calculations and Convergence

Numerical computations were performed using an in-house Python language implementation of the principal convex hull analysis (PCHA) algorithm as described in [107]. For the full dataset, the factorization $R = EW$, with $E = RC$ as defined in eq. 4.2 converged in 35 iterations to a predefined tolerance value of $\Delta_{SSE} < 10^{-7}$, where Δ_{SSE} is the average difference in sum of square error per matrix element in $R - EW$ from one iteration to the next. The resulting columns of E_{tf} are shown in fig. 4.8 (top row). Annualized cumulative log returns obtained by summing in rows of E_{tf}:

$$Q_f(\tau) = \tfrac{1}{\sqrt{250}} \sum_{t=0}^{t=\tau} E_{tf}. \tag{4.3}$$

[4] A factorization with a relaxed version of this constraint, $1 - \delta \leq \sum_{s'} C_{s'f} \leq 1 + \delta$, is described in [107]

The time series $Q_f(\tau)$ are shown in fig. 4.2 and middle row of fig. 4.8. Weights W_{fs} for selected stocks are shown in fig. 4.3, the remainder are available on companion website [10]. In each canonical sector f, the component of weights for companies are shown in fig. 4.9.

The analysis of evolving sector weights was performed similarly, but with a sliding Gaussian time window. We decomposed the local normalized log returns for each stock into the canonical sectors determined from the entire time series. Each column (time series) of the returns matrix R_{ts} was multiplied with a Gaussian, $G_\mu(\tau) = exp(-\frac{(\tau-\mu)^2}{2\times 250^2})$ of standard deviation 250 centered at μ to obtain R^μ_{ts}. With $C_{s'f}$ found using the full dataset[5] as in eq. 4.2, R^μ_{ts} is factorized to obtain new weights W^μ_{fs} that describe sector decomposition of stocks in that period focused at $t = \mu$: $R^\mu = R^\mu_{ts'} C_{s'f} W^\mu_{fs}$. μ is increased in steps of 50 starting at $\mu = 0$ and ending at $\mu = 5000$ and W^μ is calculated at each μ with the corresponding R^μ. These results are plotted in fig. 4.4 for a select group of companies, and the remainder are available on the companion website [10].

4.9.4 Dimensionality of Space of Price Returns

The stock price returns have a dimension given by number of returns in the dataset for each stock. For the dataset used for the analysis described in this paper, 20 years of returns amount to a dimensionality of 5001 (there are about 250 trading days per year). It is often the case with large datasets that the effective dimensionality of the data space is much lower when one filters out the noise. A number of dimensional reduction methods exist; the singular value decomposition (SVD) [45] (c.f. principal component analysis) which is a deterministic matrix factorization, is one of the most commonly used method. We discuss it in more detail in

[5] Fixed C corresponds to keeping the sector defining simplex corners fixed.

order to draw a contrast with previous results, and to apply it for explaining some of our results. First we introduce the following variables names and definitions.

Let matrices $\tilde{p}_{\tau s}$ and $p_{\tau s}$ represent prices and log prices respectively of stocks s at times τ. Log returns are then given by $r_{ts} = p_{(\tau+\delta)s} - p_{\tau s}$[6], where δ is the interval length over which returns are calculated. Define another matrix $R_{\tau s}$ of normalized log returns: with zero mean and unit standard deviation, $R_{\tau s} = (r_{\tau s} - \langle r_{\tau s} \rangle_\tau)/\sigma_s$, where $\sigma_s^2 = \langle r_{\tau s} \odot r_{\tau s} \rangle_\tau - \langle r_{\tau s} \rangle_\tau^2$ is the variance (squared volatility) of log returns.

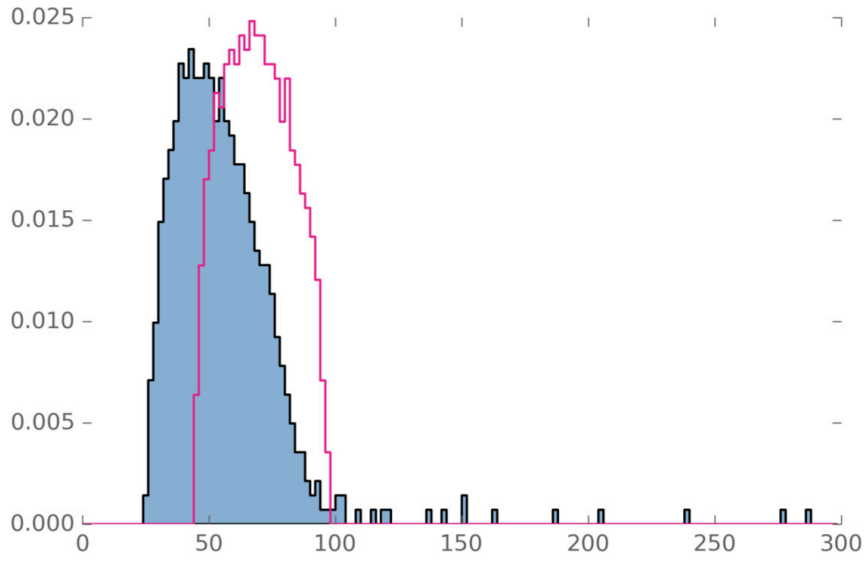

Figure 4.5: Normalized distribution of singular values. Filled blue histogram corresponds to distribution of singular values of returns from the dataset R_{ts}—one notices a clear separation of the hump-shaped bulk of singular values ascribed to random Gaussian noise, and about 20 stiff singular values (the largest singular value ∼952, corresponding to the *market mode* is not shown). Pink line histogram outline shows the distribution of singular values of a matrix of the same shape as R but containing purely random Gaussian entries.

An SVD of R_{ts} is matrix factorization [45] $R_{ts} = U_{tf}\Sigma_{ff'}V^T_{f's}$ such that matrices U and V

[6] Assuming the standard lognormal distribution of stock price returns, log of price returns have a normal distribution—a feature that makes them easily amenable to standard statistical analysis. The heavy tail distribution assumption can be accommodated with more work.

are orthonormal [7], Σ is a diagonal matrix of "singular values". If the goal were purely of rank-reduction, n entries of Σ can be chosen to lie above "noise threshold" are retained and the rest truncated so that $0 \leq f,\ f' \leq n$. This effectively reducing the dimension of R to n. The choice of n can be informed by the distribution of singular values. The distribution of singular values of a rectangular matrix A_{xy} with purely Gaussian random entries (alternatively, of the eigenvalues of a square matrix $A^T A$) has a well characterized shape [108] and has been previously used to filter noise from financial datasets [89]. As shown in the fig. 4.5, most singular values of the returns matrix R are associated with the random noise, whereas only $\sim$20 fall outside that cutoff[8]. The largest singular value of R_{ts} corresponds to what we will refer to as the "market mode" as this represents overall simultaneous rise and fall of stocks. In the analysis presented in this paper, this mode has been filtered from the returns matrix by projecting the R matrix into the subspace spanned by all non-market mode eigenvectors. This is equivalent to filtering the market mode using simple linear regression (as done commonly [12]), although more convenient.

4.9.5 Low-Dimensional Projections of Price Returns

A key discovery of our work is that the high-dimensional space of stock returns has an emergent low-dimensional hyper-tetrahedral (simplex) structure. One of the ways this structure can be seen is projecting the dataset into stiff "eigenplanes". Eigenplanes are formed by pairs of right singular vectors from the SVD. Here, we construct an SVD of the simplex corners, $E_{tf} = X_{tk} Y Z^T_{kf}$ so that in a low-dimensional representation, simplex

[7] The rows of V^T are also the eigenvectors of the stock-stock returns correlation matrix, $\xi_{ss'} \sim R^T_{st} R_{ts}$. It was previously reported that some components of the stiff eigenvectors of this stock-stock correlation matrix loosely corresponded to firms belonging to the same conventionally identified business sector [12]

[8] The singular value bounds of a random Gaussian rectangular matrix of size $\alpha \times \beta$ can be shown to be $\sqrt{\alpha} \pm \sqrt{\beta}$.

corners are given by columns of YZ^T because $YZ^T_{kf} = X^T_{kt}E_{tf}$ (in other words, X^T_{kt} is a projection operator). We now note that the plots in fig. 4.6 are the projections of the dataset, $X^T_{kt}R_{ts} = v_{ks}$. The rows of v taken in pairs form the axes of the projections in figs. 4.1 and 4.6. With those plots, it becomes clear that the eigenplanes represent projections of a simplex-like data into two-dimensions. Secondly, we note that the simplex structure becomes less clear as one looks at planes corresponding to smaller singular value directions.

An alternative way is to visualize eigenplanes directly from the SVD from returns matrix R_{ts} (section 4.9.4), and project the simplex corners E_{tf} in those planes. The resulting projections (not shown) will be similar to the structures seen in fig. 4.6 but in a different basis.

Similarly, the results of the factorization can be seen in eigenplanes from the SVD of $E_{tf}W_{sf} = L_{tk}MN^T_{ks}$. These results (rows of MN^T_{ks}) are shown in fig. 4.7, where we notice that the data is now perfectly residing in simplex region as expected due to constraints.

4.9.6 Proportion of Variation Explained (PVE)

We measure the goodness of the returns decomposition $R = EW$ by measuring the proportion of variation explained (PVE) as follows:

$$PVE \quad = 1 - SSE/SST. \tag{4.4}$$

Here, SSE is denotes the sum of square errors $||R - EW||^2_F$, and SST is the total sum of squares $||R||^2_F$. For the full dataset factorized according to eq. 4.2, we obtain PVE $= 11.6\%$ using the equation above. To put this number is context for the returns dataset, one must separate the variation in R ascribable to signal, and that to Gaussian fluctuations. The SVD of R with singular values shown in fig. 4.5 provides a convenient way for doing so as follows. Only 20 singular values (excluding the market mode) were above the cut-off that was

predicted by the random matrix theory for a matrix of purely random Gaussian entries. For any matrix M with elements m_{ij}, the norm $||M||_F^2 = \sum_{i,j} m_{ij}^2 = \sum_i s_i^2$, where s_i are the singular values [45]. Thus, the fraction of intrinsic variation in R not attributable to noise is the sum of squares of the 20 singular values (not including market mode) divided by SST, $\sum_{i=1}^{i=20} s_i^2 / ||R||_F^2 = 19.8\%$. Therefore, as a first approximation, the factorization of eq. 4.2 explains $11.6/19.8 = 59\%$ of the total variation. We also note for completeness that if R is rank-reduced to eight stiffest components found by SVD (not including market mode), then the factorization of eq. 4.2 explains 85% of the the total variation in R with overall results in good accord with the analysis presented here. This implies that sector decomposition information was already contained in the stiff modes from SVD of R, however SVD is not the appropriate tool for the decomposition.

4.9.7 Determining the Number n of Canonical Sectors

It is an open problem to determine the effective dimensionality (optimal rank) of a general dataset (matrix). One could select among models of different dimensions using statistical tests such as the PVE discussed above, or information theory based criteria such as Akaike Information Criterion (AIC) or the Bayesian Information Criterion (BIC), but the choice of the selection criterion is itself generally made on an *ad hoc* basis. Therefore, the most direct observation of results is also the most reliable. In the dataset used for analysis described here, a factorization with $n > 8$ yielded results where both the emergent time series E_{tf} and weights in W_{fs} showed qualitative signs of overfitting. The high-level results of factorization with different values of n are discussed below.

- $n = 9$: Results were in good agreement to $n = 8$, except one resulting sector involved participation from only 11 seemingly unrelated stocks (PVE$= 12.4\%$).

- $n = 7$: Results were similar to $n = 8$, except *c-real estate* and *c-financial* merged into one canonical sector (PVE$= 10.7\%$).
- $n = 6$: Results are similar to $n = 7$, except listed retail companies divide into *c-cyclical* and *c-non-cyclical* sectors (PVE$= 9.9\%$).
- $n = 5$: Results are similar to $n = 6$, except *c-cyclical* and *c-non-cyclical* merge into one canonical sector (PVE$= 8.7\%$).
- $n = 4$: Results are similar to $n = 5$, except *c-energy* and *c-utility* merge into one canonical sector (PVE$= 6.9\%$).
- $n = 4$: All sectors overlap and there is no clear separation of companies (PVE$= 5.2\%$).

In general, a factorization analysis of the returns dataset would be sensitive to the following factors and care must be taken in order to interpret results:

- Number of stocks in the dataset.
- Criteria applied for picking stocks.
- Period over which historical prices are obtained.
- Frequency at which returns are computed.

A robust macroeconomic analysis would therefore require a large number of stocks chosen without sampling bias, with returns calculated over the period of interest and sensitivity checked for frequency of returns calculation (In general, the number of time points should exceed the number of stocks.). On the other hand, an equity fund manager faces a less daunting task for an analysis that is limited the universe of her portfolio of stocks: either to find its canonical sectors, or to analysis the exposure of her holdings to the core sectors of the economy.

4.9.8 CANONICAL SECTOR INDICES

The matrix C_{sf} in decomposition in equation in eq. 4.2 represents how returns R of stocks s must be combined to make canonical sector returns $E_{tf} = R_{ts}C_{sf}$. Since an canonical sector is defined as a combination of stocks, an investment in the sector f can made via buying a basket of constituent stocks s in proportions given by C_{sf} or through an index I_{tf}:

$$I_{tf} \quad = \tilde{P}_{ts'}C_{s'f}. \tag{4.5}$$

where, $\tilde{P}$ are stocks prices suitably weighted by market cap or other divisor as common practice for common indices [109]. An unweighted index of this kind is shown in bottom row of fig. 4.8 for results corresponding to the analysis described in this paper. Conversely, a pre-defined basket of stocks such as the S&P500 can be unbundled to find its exposure to the canonical sectors. With an investment strategy employing longs and shorts at the same time in correct proportions, it is conceivable to invest in, for example, the c-tech component of S&P500.

The desirable features of an index include completeness, objectivity and investability [110]. The *c-indices* constructed using the ideas outlined here would not only be of value to investors through investment vehicles such as ETFs, Futures, etc., but also serve as important macroeconomic indicators.

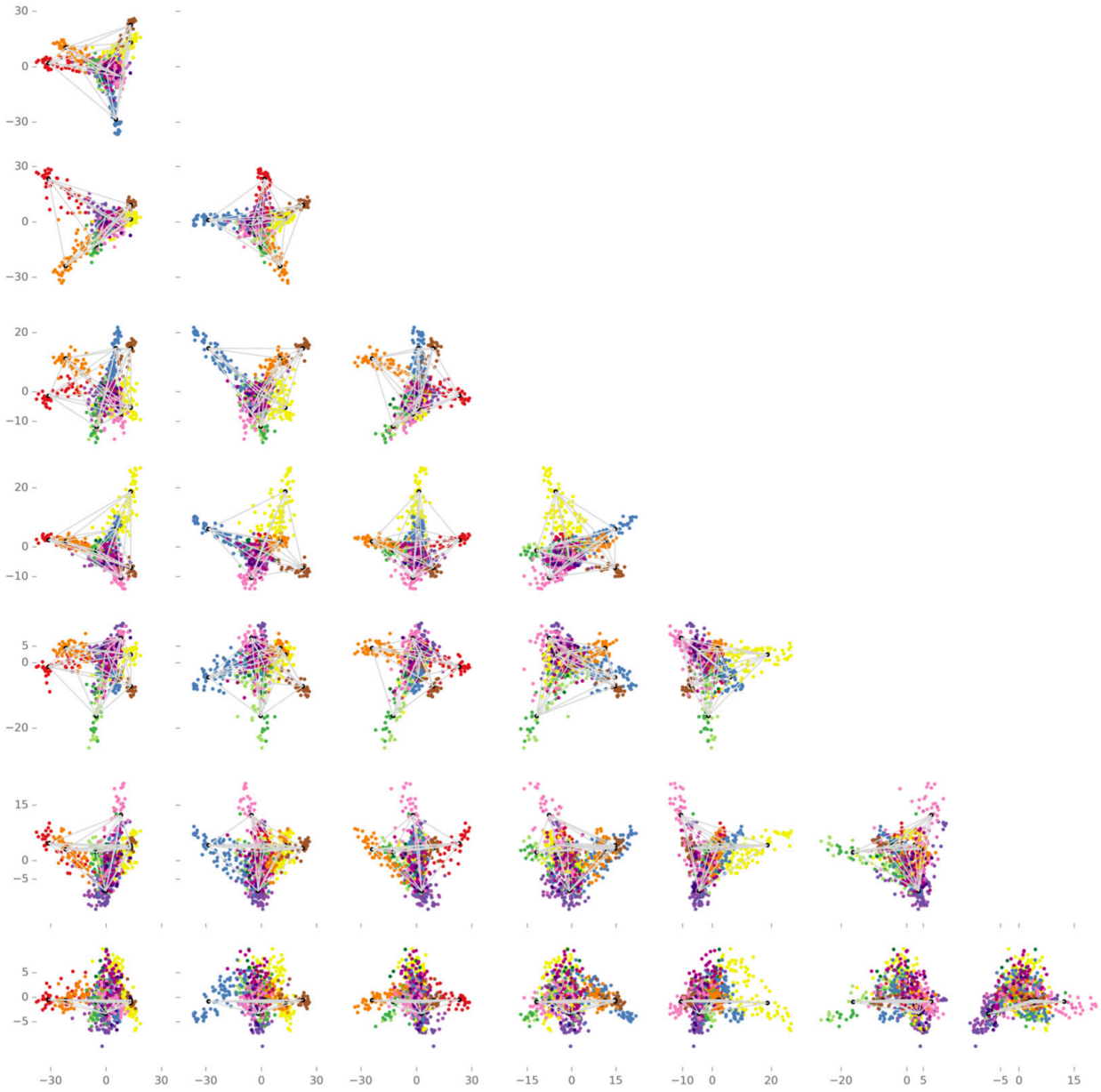

Figure 4.6: Low-dimensional projections of stock returns data. Each colored circle represents a stock in our dataset is colored according to listed sectors scheme in fig. 4.9 according to sectors assigned by Scottrade [11]. The first row is repeated from fig. 4.1. Black circles represent are the archetypes found with our analysis. The $(i, j)^{th}$ figure in the grid is a plane spanned by singular vectors i and $j + 1$ (rows of $X^T R$) from the calculations described in section 4.9.5. Projections after the factorization are shown in fig. 4.7.

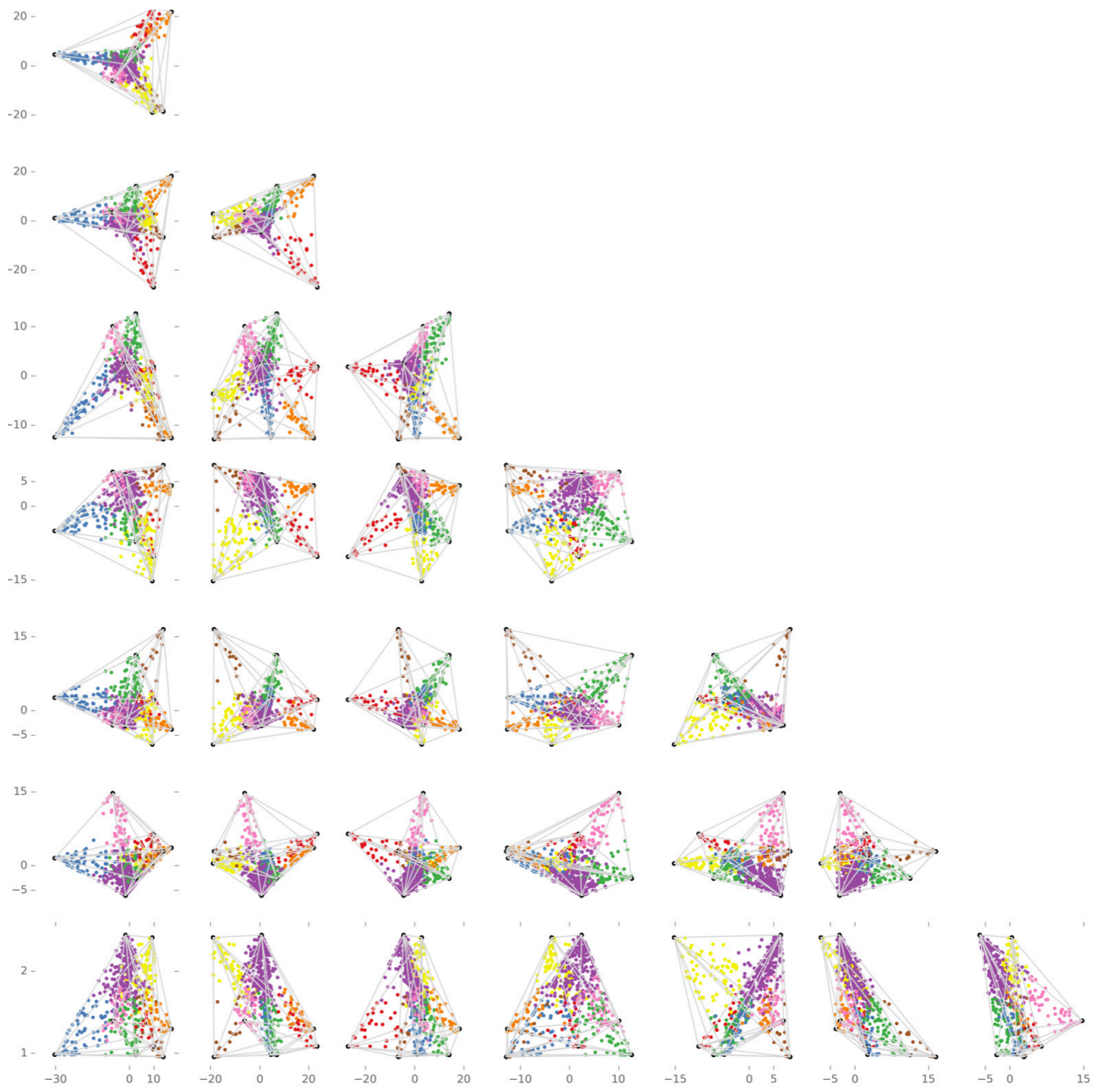

Figure 4.7: Cross-sections along eigenplanes of the factorized returns. Each colored circle represents a stock in our dataset is colored according to scheme in fig. 4.3 based on the primary sector association found after calculations described in this paper. Black circles represent the archetypes found with our analysis. The $(i, j)^{th}$ figure in the grid is a plane spanned by singular vectors i and $j + 1$ (rows of MN^T) from the calculations described in section 4.9.5. Projections of raw data (before the factorization) are shown in fig. 4.6.

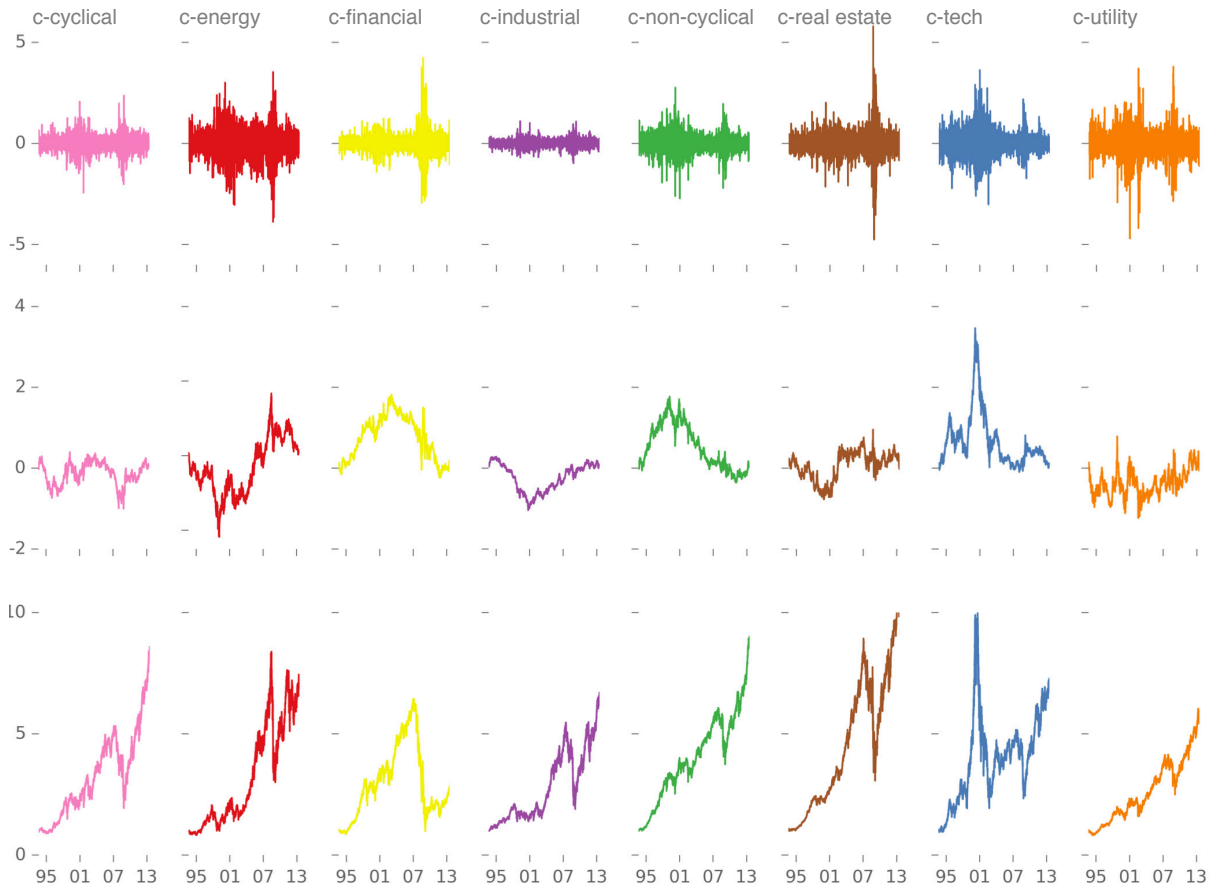

Figure 4.8: Canonical sector time series. Top row: normalized log returns (columns of E_{tf}), middle row: cumulative log returns (same as fig. 4.2 as defined in equation 4.3, and bottom row: unweighted price index of canonical sectors (eq. 4.5).

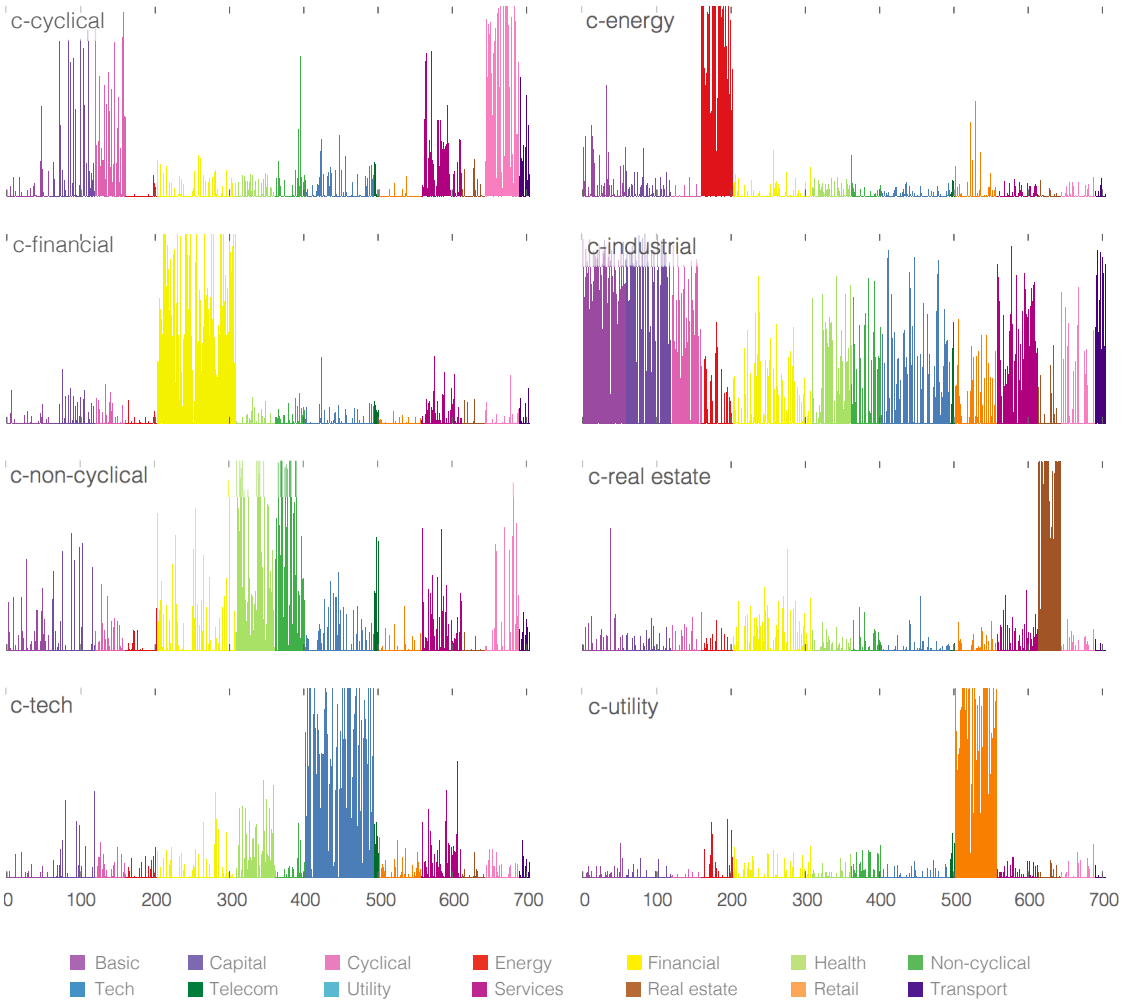

Figure 4.9: Weight distribution in canonical sectors. Each of the eight subplots shows the constituent participation weights of all 705 companies in an canonical sector (rows of W_{fs}). Stocks are colored by listed sectors as shown at the bottom. Listed sector information was obtained from [11]. Y-axis range is from 0 to 1.

Figure 4.10: Singular vectors V_{fs}^T of SVD of returns R_{ts}. The orthonormal right singular vectors (rows of V_{fs}^T) of SVD of R_{ts} are equivalent to the eigenvectors of the stock-stock correlation matrix $\xi_{ss'} \sim R^T R$. Eight of these stiffest eigenvectors including the *market mode* are shown in rows of two at a time. Each has 705 components corresponding to stocks in an the dataset. The *market mode* with all components in the same direction describes overall fluctuations in the market; it was excluded from the analysis described in the paper. Previous work [12] has suggested that each eigenvector of the stock-stock correlation matrix describes a listed sector, however as seen above, a more correct interpretation is that each eigenvector is a mixture of listed sectors with opposite signs in components. For example, the stiffest direction (after market mode) has positive components in real estate and utility, but negative in tech. Less stiff eigenvectors (including the last one shown here), do not contain sector-relevant information. Stocks are colored by listed sectors as shown at the bottom. Listed sector information was obtained from [11]. Y-axis range is from –0.5 to 0.3.

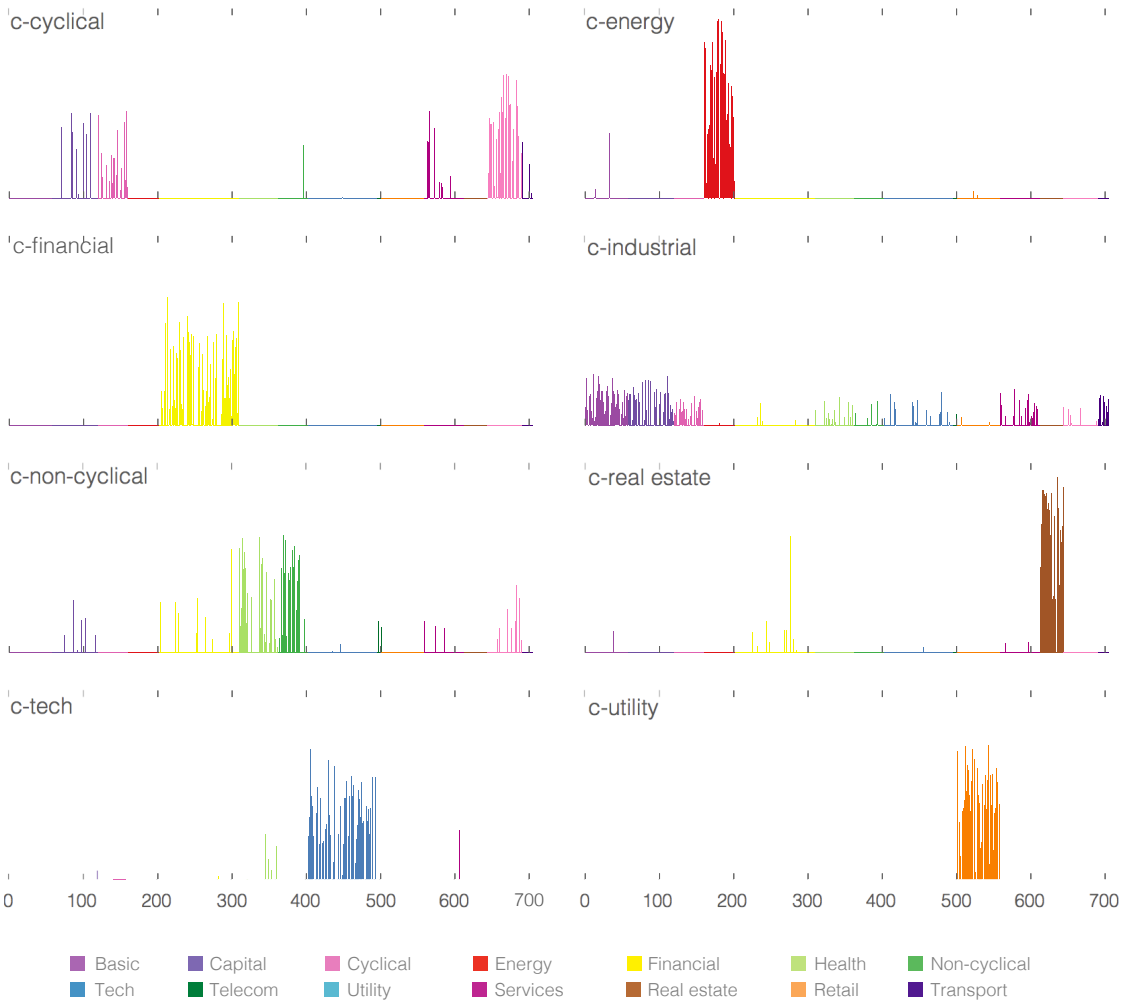

Figure 4.11: Canonical Sector Constituents (shown as columns of the C_{sf}). C_{sf} represents a weighted combination stocks that defines of the canonical sector each of which has a time series represented by E_{tf} that is given by $E_{tf} = R_{ts}C_{sf}$. The eight subplots show the constituent participation component of stocks in each canonical sector f. Canonical sectors are labeled on the plot; their names were chosen according to the listed sectors of firms that comprise them. Noteworthy features seen above include the co-association of listed sectors: basic, capital, transport and part of cyclicals into *industrial goods*. Similarly, healthcare and non-cyclicals are coupled together in what we call *non-cyclicals*. Canonical *retail* goes primarily with listed retail and cyclicals. Stocks are colored by listed sectors as shown at the bottom. Listed sector information was obtained from [11]. Y-axis range is from 0 to 0.05.

APPENDIX A

MISCELLANEOUS RESULTS AND INFORMATION

Several individuals have contributed toward the results collected herein. Sections A.1, A.2 are mostly due to Benjamin B. Machta and James P. Sethna. Section A.4 was written with the assistance of Alexander A. Alemi.

A.1 FISHER INFORMATION MATRIX AS A METRIC ON PARAMETER SPACE

This section gives an overview of the information theoretic approach used throughout [58, 111, 112] motivated by the following questions: how different are two probability distributions, $P_1(x)$ and $P_2(x)$, and what is an appropriate measure of distance between them? Can one test the hypothesis that a set of independent data points $\{x_1, x_2, ..., x_N\}$ (unbeknownst to us generated by P_1) was instead generated by P_2? The probability that P_1 would have generated the data is given by its likelihood:

$$P_1(\{x_1, x_2, ...x_N\}) = \prod_i P_1(x_i) = \exp\left(\sum_i \log P_1(x_i)\right) \tag{A.1}$$

To determine which of two candidate models more probably generated this sequence of data, one considers the log likelihood ratio:

$$\lambda(\{x_1, x_2, ...x_N\}) = \log\left(\frac{P_1(\{x_1, x_2, ...x_N\})}{P_2(\{x_1, x_2, ...x_N\})}\right) \tag{A.2}$$

If λ is large and positive (negative) than the data suggests P_1 (P_2). Alternatively, if λ is close to zero than either model could be valid and the data is inconclusive. How much data is needed before one should expect that one model distinguishes itself? In a given distribution, λ is a stochastic variable. However, one can define the expectation value for $\lambda(x)$ in distribution P_1, giving the log likelihood per sample that an ensemble drawn from P_1 could have instead been drawn from P_2.

This defines the Kullback-Liebler Divergence, D_{KL}, a statistical measure of how distinguishable P_1 is from P_2 from its data x [112, 113]

$$D_{KL}(P_1||P_2) = \sum_x P_1(x)\lambda(x) = \sum_x P_1(x) \log\left(\frac{P_1(x)}{P_2(x)}\right). \tag{A.3}$$

Because D_{KL} does not necessarily satisfy $D_{KL}(P_1||P_2) = D_{KL}(P_2||P_1)$, it is not a mathematically proper distance metric[1]. However, D_{KL} becomes symmetric for two 'nearby' models. For a continuously parameterized set of models P_θ where θ is a set of N parameters θ^μ, the infinitesimal D_{KL} between models P_θ and $P_{\theta+\Delta\theta}$ is[2]

$$D_{KL}(P_\theta, P_{\theta+\Delta\theta}) = g_{\mu\nu}\Delta\theta^\mu\Delta\theta^\nu + \mathcal{O}\Delta\theta^3, \tag{A.4}$$

where $g_{\mu\nu}$ is the Fisher Information Matrix (FIM), given by[3]

$$g_{\mu\nu}(P_\theta) = -\sum_x P_\theta(x)\frac{\partial}{\partial\theta^\mu}\frac{\partial}{\partial\theta^\nu}\log P_\theta(x). \tag{A.5}$$

[1] A distance measure should also satisfy some sort of generalized triangle inequality- at the very least $D(A,B) + D(B,C) \geq D(A,C)$ which is also not necessarily satisfied here.

[2] It is an interesting exercise to show that there is no term linear in $\Delta\theta$. The crucial step uses that P_θ is a probability distribution so $\partial_\mu \sum_x P_\theta(x) = 0$.

[3] Although the KL-divergence is a common measure of statistical distinguishability among probability distributions, it is not unique. In fact it is a member of a broader class of divergences known as the f-divergences, which take the form $D_f(P_1, P_2) = \sum_x P_1(x) f\left(\frac{P_2(x)}{P_1(x)}\right)$ for some function $f(t)$ that is convex and satisfies $f(1) = 0$. The KL-divergence therefore corresponds to the choice $f(t) = -\log(t)$. Other common choices are $f(t) = 2(1-\sqrt{t})$, corresponding to the Hellinger Distance, and $f(t) = |t-1|$, corresponding to the total variation distance. For our purposes, this distinction is unimportant: the Fisher Information is the lowest order contribution to any f-divergence for infinitesimally separated probability distributions.

The quadratic form of the KL-divergence at short distances motivates using the FIM as a metric on parameter space. The FIM is symmetric, positive-definite, and transforms like a covariant rank-2 tensor under parameter transformations, endowing it with all the properties of a Riemannian metric, the study of which is known as information geometry [58]. In fact, the FIM is the *unique* natural Riemannian metric that is consistent with the additional structure that each point specifies a probability distribution[4].

Information geometry provides a framework for understanding more generalized Bayesian inference. It gives an immediate derivation of Jeffreys' 'uninformative' prior [116]: the invariant volume element in any Riemannian geometry is given by $\sqrt{\det(g)}d\theta^1 d\theta^2 \dots d\theta^N$. In a Bayesian inference scheme, choosing a prior on parameter space equal to $\sqrt{\det(g)}/Z$ ensures that model predictions are reparameterization invariant. The normalization constant, Z, is the invariant volume of the manifold that quantifies the amount of information expected to be gained from a single measurement of x.

The FIM is well defined for any models that predict stochastic data. The next sub-sections derive the form of the FIM for two special cases used in this work, the case of Gaussian models, and the case of exponential families familiar from statistical physics. The similarity of parameter space structure in these seemingly very different classes of models suggests that it is not an artifact of the particular choice of stochastic model employed.

[4]Riemannian metrics have more structure than other metric spaces since the metric tensor defines an inner product on the tangent space at each point on the manifold. The FIM is the only inner product that is invariant under specific probabilistically important mappings. The basic argument considers partitions on the domain of the probability distribution, known as Markov mappings. Requiring that the inner product be invariant under these mappings is a rigid constraint that is only satisfied by the FIM [114, 115]

A.1.1 THE METRIC OF A GAUSSIAN MODEL

Nonlinear least squares models output a vector of data, y_0^i (for $1 < i < M$), that is generated assuming that the observations y^i are normally distributed with widths σ^i around prediction $\vec{y}_0(\theta)$. The fitting 'cost' or sum of squared residuals is proportional to the negative log likelihood (plus a constant), hence the probability distribution of data is

$$P_\theta(\vec{y}) \sim \exp\left(-\sum_i (y^i - y_0^i(\theta))^2/2\sigma^{i2}\right). \tag{A.6}$$

Defining the Jacobian between parameters and scaled data as

$$J_{i\mu} = \frac{1}{\sigma^i}\frac{\partial y_0^i(\theta)}{\partial \theta^\mu}, \tag{A.7}$$

the Fisher Information Matrix for least squares problems is given by[5] [4, 57]

$$g_{\mu\nu} = \sum_i J_{i\mu} J_{i\nu}. \tag{A.8}$$

The Euclidean distance between nearby points in prediction space

$$\begin{aligned} \sum(\Delta y_i)^2 &= \sum_{i,\mu,\nu}\left(\tfrac{\partial y^i}{\partial \theta^\mu}\Delta\theta^\mu \tfrac{\partial y^i}{\partial \theta^\nu}\Delta\theta^\nu\right) \\ &= g_{\mu\nu}\Delta\theta^\mu\Delta\theta^\nu \end{aligned} \tag{A.9}$$

is the metric tensor contracted with corresponding displacements $\Delta\theta^\mu$ in parameter space. Thus the FIM has a geometric interpretation: distance is locally the same as that measured by embedding the model in the space of scaled data according to the mapping $y_0(\theta)$ (it is *induced* by the Euclidian metric in data space). This metric was shown to be sloppy in seventeen models from the systems biology literature [1] and in several other contexts. See fig. 3.1 and [8].

[5] This assumes that the uncertainty σ^i does not depend on the parameters, and that errors are diagonal. Both of these assumptions seem reasonable for a wide class of models if measurement error dominates. The more general case is still tractable, but less transparent

A.1.2 THE METRIC OF A STATISTICAL MECHANICAL MODEL

Exponential models familiar from statistical mechanics are defined by a parameter set θ dependent Hamiltonian H that assigns an energy to every possible configuration x. Each parameter θ^μ controls the relative weighting of some function of the configuration, $\Phi_\mu(x)$, which together define the probability distribution on configurations through the following (with temperature and Boltzmann's constant set to 1)

$$\begin{aligned} P(x|\theta) &= \exp(-H_\theta(x))/Z, \\ Z(\theta) &= \exp(-F(\theta)) = \sum_x \exp(-H_\theta(x)), \\ H_\theta(x) &= \sum_\mu \theta^\mu \Phi_\mu(x) \end{aligned} \tag{A.10}$$

Here F is the Helmholtz free energy. Many models can be put into this exponential form. For example, the 2d Ising model of section 3.5 has spins $s_{i,j} = \pm 1$ on a square $L \times L$ lattice with the configuration, $x = \{s_{i,j}\}$, being the state of all spins. The magnetic field, $\theta^0 = h$ multiplies $\Phi_0(\{s_{i,j}\}) = \sum_{i,j} s_{i,j}$, and the nearest neighbor couplings, $\theta^{01} = \theta^{10} = -J$ multiplies $\Phi_1(\{s_{i,j}\}) = \sum_{i,j} s_{i,j}s_{i+1,j} + s_{i,j}s_{i,j+1}$. This form is chosen for convenience in calculating the metric, which is written [64, 66, 117][6]

$$\begin{aligned} g_{\mu\nu} &= \langle -\partial_\mu \partial_\nu \log(P(x)) \rangle , \\ &= \langle \partial_\mu \partial_\nu H(x) \rangle + \partial_\mu \partial_\nu \log(z), \\ &= \partial_\mu \partial_\nu \log(z) = -\partial_\mu \partial_\nu F. \end{aligned} \tag{A.11}$$

In the last equation we have taken advantage of the fact that the Hamiltonian is linear in parameters θ^μ so that $\langle \partial_\mu \partial_\nu H(x) \rangle = 0$.

[6] Several seemingly reasonable metrics can be defined for systems in statistical mechanics and all give similar results in most circumstances [66]. Most differences occur either for systems not in a true thermodynamic (N large) limit, or for systems near a critical point. As far as we are aware, Crooks [64] was the first to stress that the one used here can be derived from information theoretic principles, perhaps making it the most 'natural' choice. Crooks showed [64] that when using this metric 'length' has an interesting connection to dissipation by way of the Jarzynski equality [118].

A.2 DERIVATION OF FIM EIGENVALUES

Here we discuss the way the eigenvalues of the FIM scale near the Ising critical point, deriving the results quoted in equation 3.18. Our formula for the FIM is given by

$$\begin{aligned} g^s_{\mu\nu} &= A\sum_{\alpha,\beta}(\tfrac{\partial u^\alpha}{\partial\theta^\mu}\tfrac{\partial u^\beta}{\partial\theta^\nu})(\tfrac{\partial}{\partial r^\alpha}\tfrac{\partial}{\partial r^\beta}\mathcal{U})\xi^{y_\alpha+y_\beta-d} \\ &= J^\alpha_\mu \hat{g}^s_{\alpha\beta} J^\beta_\nu \end{aligned} \tag{A.12}$$

where $\hat{g}^s_{\alpha\beta}$ is the metric tensor in the scaling variable coordinates $u^\alpha(\vec{\theta})$ for which the renormalization-group flows expand by a factor b^{y_α}, and $J^\beta_\nu = \partial u^\beta/\partial\theta^\nu$ is the Jacobian transforming the natural coordinates θ^ν to the scaling variable coordinates. Our job is to show that the ordered eigenvalues λ^s_i of g^s scale like

$$\lambda^s_i \sim A\xi^{2y_i-d} \tag{A.13}$$

(equation 3.18). To do so, we first demonstrate that the eigenvalues $\hat{\lambda}_i$ of the FIM $\hat{g}^s$ in scaling variable coordinates satisfies this bound, and then show that this scaling is preserved by the transformation J to bare coordinates.

We make use of Weyl's inequality for matrix eigenvalues, which implies that if B and M are real, symmetric matrices and $B-M$ is nonnegative definite, then each ordered eigenvalue of B is greater than or equal to the corresponding one of M. Let us write

$$\hat{g}^s_{\alpha\beta} = A\xi^{-d}\left(\frac{\partial}{\partial r^\alpha}\frac{\partial}{\partial r^\beta}\mathcal{U}\right)\xi^{y_\alpha+y_\beta} = A\xi^{-d}EME \tag{A.14}$$

where $M_{\alpha\beta} = \partial^2\mathcal{U}/\partial r^\alpha\partial r^\beta$ and $E_{\sigma\rho} = \delta_{\sigma\rho}\xi^{y_\sigma}$. This form of $\hat{g}^s$ is similar to that of matrices studied in [8].

Let C be the maximum eigenvalue of M, and let $B_{\alpha\beta} = C\delta_{\alpha\beta}$, so in particular $B-M$ is nonnegative definite, and hence $W^T(B-M)W \geq 0$ for any vector W.

Conclusion: $(A\xi^{-d}EBE - \hat{g}^s)$ is nonnegative definite, and thus $\hat{g}^s$ has sorted eigenvalues $\hat{\lambda}_i \leq CA\xi^{2y_i-d}$.

Argument: Because $\hat{g}^s = A\xi^{-d}EME$, for any vector V,

$$V^T(A\xi^{-d}EBE - \hat{g}^s)V = V^T(A\xi^{-d}E(B-M)E)V = A\xi^{-d}W^T(B-M)E)W \geq 0, \quad \text{(A.15)}$$

where $W = EV = VE$. Since $B_{\alpha\beta} = C\delta_{\alpha\beta}$ and $E_{\alpha\beta} = \xi^{y_\alpha}\delta_{\alpha\beta}$ are diagonal, the sorted eigenvalues of $A\xi^{-d}EBE$ are just $CA\xi^{2y_i-d}$, which by Weyl's inequality bound the sorted eigenvalues of $\hat{g}^s$.

We now need to transform from the scaling coordinates u^α to the original coordinates θ^ν. The mapping from scaling variable to bare coordinates is non-orthogonal. Let the eigenvector of $\hat{g}^s$ corresponding to $\hat{\lambda}_i$ be $\hat{v}_i$. Each scaling-coordinate eigenvector transforms to a vector in parameter space,

$$V^i_\mu = \sum_\alpha \hat{v}^\alpha_i J^\alpha_\mu = \sum_\alpha \hat{v}^\alpha_i \frac{\partial u^\alpha}{\partial \theta^\mu}. \quad \text{(A.16)}$$

The V^is are neither orthogonal nor normalized. The metric in parameter space can be written as:

$$g^s_{\mu\nu} = \sum_{i=1}^{\infty} \hat{\lambda}_i V^i_\mu V^i_\nu \quad \text{(A.17)}$$

Conclusion: The sorted eigenvalues of g^s, the FIM matrix in the original coordinates, scale as $\lambda_i \sim A\xi^{2y_i-d}$.

Argument: Consider the truncated version of this matrix formed by adding just the first N contributions:

$$g^{s,N}_{\mu\nu} = \sum_{i=1}^{N} \hat{\lambda}_i V^i_\mu V^i_\nu. \quad \text{(A.18)}$$

It is positive semidefinite, with rank N. Also, $g^{s,N+1} - g^{s,N}$ is nonnegative definite, so Weyl's inequality tells us that the sorted eigenvalues of $g^{s,N+1}$ are each greater than or equal to those of $g^{s,N}$, $\lambda^{i,N+1} > \lambda^{i,N}$. As traces of matrices sum, we also have that

$\sum_i \lambda^{i,N+1} - \lambda^{i,N} = \hat{\lambda}_{N+1}|V^{N+1}|^2$. This implies that all eigenvalues must increase, with none increasing by more than $\hat{\lambda}_{N+1}|V^{N+1}|^2$. As the mapping from parameter space to scaling variables is analytic at the critical point, the normalization factor $|V|^2$ is order one (does not diverge as $\xi \to \infty$). Hence the eigenvalue λ_i in parameter space is a sum of positive terms $\sim \hat{\lambda}_j$ for $j \geq i$. Since by the Lemma $\hat{\lambda}_j \leq CA\xi^{2y_j - d}$, as $\xi \to \infty$ the dominant term will be $\hat{\lambda}_i$, so $\lambda_i \sim A\xi^{2y_i - d}$.

A.3 Derivation of α and β of the Linear Regression Equation

Let matrix D_{st} describe returns of stocks s at times t with dimensions (S, T). The goal is to write the return for each stock as sum of the total market return M_t that is common to all stocks (also known as a benchmark index), an outperformance coefficient α and an error ϵ. Each stock couples of the market through a constant β.

$$D_{st} \quad = \alpha_s 1_t + \beta_s M_t + \epsilon_{st} \tag{A.19}$$

A.3.1 A New Elementary Method

The common way to derive expressions for α and β is through a minimization of the sum of the squares of errors ϵ_{st}. A more transparent way presented here fully utilizes two key ideas: (1) $\epsilon_{st} = 0_s$, i.e. errors are distributed around 0 for each stock, and (2) $cov(\epsilon_{st}, M_t) = 0_s$, i.e. error for each stock is statistically independent of the market return.

To derive the expression for β, subtract the time average of both sides of the previous equation from itself:

$$D_{st} - \langle D_{st}\rangle 1_t \quad = \beta_s(M_t - \langle M_t\rangle 1_t) + \epsilon_{st} \tag{A.20}$$

Right multiply both sides by $M_t - \langle M_t\rangle 1_t$. Take mean again and solve for β, using $cov(\epsilon_{st}, M_t) = 0$ to get:

$$\beta_s \quad = \frac{\langle D_{st}M_t\rangle - \langle D_{st}\rangle\langle M_t\rangle}{\langle M_t^2\rangle - \langle M_t\rangle^2} = \frac{cov(D_{st}, M_t)}{var(M_t)} \tag{A.21}$$

Take the time average of both sides of the first equation to get α as a function of β: $\alpha_s = \langle D_{st}\rangle - \beta_s\langle M_t\rangle$ and then use expression of β from previous equation in terms of known quantities.

A.3.2 The Standard Method

To minimize the square error in the linear regression equation, we compute the residuals:

$$r^2 = \sum_{s,t} (D_{st} - \alpha_s 1_t - \beta_s M_t)^2 \tag{A.22}$$

and start by finding the best α:

$$\partial r^2 \alpha_s = \sum_t (D_{st} - \alpha_s 1_t - \beta_s M_t) = 0 \tag{A.23}$$

which gives:

$$\begin{aligned} T\alpha_s &= \sum_t [D_{st} - \beta_s M_t] \\ \alpha_s &= \langle D_s\rangle - \beta_s\langle M\rangle. \end{aligned} \tag{A.24}$$

Here we have introduced the notation:

$$\langle A\rangle \equiv \frac{1}{T}\sum_t A_t. \tag{A.25}$$

Proceeding with β we have,

$$\partial r^2 \beta_s = 0 = \sum_t \left(D_{st} - \alpha_s 1_t - \beta_s M_t\right)\left(-M_t\right) \tag{A.26}$$

So that

$$\beta_x \left\langle M^2 \right\rangle = \left\langle D_s M \right\rangle - \alpha_s \left\langle M \right\rangle \tag{A.27}$$

$$= \left\langle D_s M \right\rangle - \left\langle M \right\rangle \left(\left\langle D_s \right\rangle - \beta_s \left\langle M \right\rangle\right) \tag{A.28}$$

$$\beta_s \left(\left\langle M^2 \right\rangle - \left\langle M \right\rangle^2\right) = \left\langle D_s M \right\rangle - \left\langle D_s \right\rangle \left\langle M \right\rangle \tag{A.29}$$

$$\tag{A.30}$$

Which gives

$$\beta_s = \frac{\left\langle D_s M \right\rangle - \left\langle D_s \right\rangle \left\langle M \right\rangle}{\left\langle M^2 \right\rangle - \left\langle M \right\rangle^2} \tag{A.31}$$

Note that if we define

$$\mu_t \equiv M_t - \left\langle M \right\rangle 1_t \tag{A.32}$$

this simplifies a bit

$$\beta_s = \frac{\left\langle D_s \mu \right\rangle}{\left\langle \mu^2 \right\rangle} \tag{A.33}$$

Now we can express our residual matrix

$$\epsilon_{xt} = D_{xt} - \alpha_x 1_t - \beta_x M_t \tag{A.34}$$

$$= D_{xt} - \left(\left\langle D_x \right\rangle - \beta_x \left\langle M \right\rangle\right) 1_t - \beta_x M_t \tag{A.35}$$

$$= \left(D_{xt} - \left\langle D_x \right\rangle 1_t\right) - \beta_x \left(M_t - \left\langle M \right\rangle\right) \tag{A.36}$$

$$= d_{xt} - \frac{\mu_t \left\langle \mu D_x \right\rangle}{\left\langle \mu^2 \right\rangle} \tag{A.37}$$

where similar to μ, d is defined

$$d_{st} \equiv D_{st} - \left\langle D_s \right\rangle 1_t \tag{A.38}$$

A.4 SVD OF CENTERED DATA

As discussed in previous section we can write our data as,

$$D_{st} = \alpha_s 1_t + \beta_s M_t + \epsilon_{st} \tag{A.39}$$

where data D_{st} is the data of dimension (S, T). In the SVD of D:

$$D_{st} = \sum_f U_{sf} \Sigma_f V_{ft}^T. \tag{A.40}$$

The market mode M_t is the largest singular vector:

$$M_t \equiv V_{0t}^T \tag{A.41}$$

We will now discuss the how SVD changes with centering (removing the means).

$$\begin{aligned} d_{st} &= D_{st} - \langle D \rangle 1_t && \text{(A.42)} \\ &= \sum_f U_{sf} \Sigma_f V_{ft}^T - \frac{1}{T} \sum_t \sum_f U_{xf} \Sigma_f V_{ft}^T && \text{(A.43)} \\ &= \sum_f U_{sf} \Sigma_f \left(V_{ft}^T - \left\langle V_f^T \right\rangle 1_t \right) && \text{(A.44)} \\ &= \sum_f U_{sf} \Sigma_f v_{ft}^T && \text{(A.45)} \end{aligned}$$

where v are simply centering the time-like singular vectors:

$$v_{ft}^T \equiv V_{ft}^T - \left\langle V_f^T \right\rangle 1_t \tag{A.46}$$

therefore,

$$M_t = V_{0t}^T \implies \mu_t = v_{0t}^T \tag{A.47}$$

Note that these centered singular vectors are not orthonormal anymore. So this is not a new SVD precisely. We can see what has happened in this transformation:

$$\epsilon_{st} = d_{st} - \frac{\mu_t \langle \mu D_s \rangle}{\langle \mu^2 \rangle} \tag{A.48}$$

$$= d_{st} - \frac{v_{0t}^T}{\langle (v_0^T)^2 \rangle} \frac{1}{T} \sum_t D_{st} v_{0t}^T \tag{A.49}$$

$$= d_{st} - \frac{v_{0t}^T}{\langle (v_0^T)^2 \rangle} \frac{1}{T} \sum_{t,f} U_{sf} \Sigma_f V_{ft}^T v_{0t}^T \tag{A.50}$$

$$= d_{st} - \frac{v_{0t}^T}{\langle (v_0^T)^2 \rangle} \frac{1}{T} \sum_f U_{sf} \Sigma_f \sum_t V_{ft}^T \left(V_{0t}^T - \langle V_0^T \rangle 1_t \right) \tag{A.51}$$

$$\epsilon_{st} = d_{st} - \frac{v_{0t}^T}{\langle (v_0^T)^2 \rangle} \frac{1}{T} \sum_f U_{sf} \Sigma_f \left(\delta_{0f} - T \langle V_0^T \rangle \langle V_f^T \rangle \right) \tag{A.52}$$

$$= d_{st} - \frac{v_{0t}^T}{T \langle (v_0^T)^2 \rangle} \left(U_{s0} \Sigma_0 - T \langle V_0^T \rangle \sum_f U_{sf} \Sigma_f \langle V_f^T \rangle \right) \tag{A.53}$$

$$= \left[U_{s0} \Sigma_0 v_{0t}^T + \sum_{f>0} U_{sf} \Sigma_f v_{ft}^T \right] - \frac{v_{0t}^T}{T \langle (v_0^T)^2 \rangle} \left(U_{x0} \Sigma_0 - T \langle V_0^T \rangle \left[U_{s0} \Sigma_0 \langle V_0^T \rangle + \sum_{f>0} U_{sf} \Sigma_f \langle V_f^T \rangle \right] \right) \tag{A.54}$$

$$= \left[U_{s0} \Sigma_0 v_{0t}^T - \frac{v_{0t}^T}{T \langle (v_0^T)^2 \rangle} \left(U_{s0} \Sigma_0 - T \langle V_0^T \rangle U_{s0} \Sigma_0 \langle V_0^T \rangle \right) \right] \tag{A.55}$$

$$+ \left[\sum_{f>0} U_{sf} \Sigma_f v_{ft}^T + \frac{v_{0t}^T}{T \langle (v_0^T)^2 \rangle} T \langle V_0^T \rangle \sum_{f>0} U_{sf} \Sigma_f \langle V_f^T \rangle \right] \tag{A.56}$$

$$= \left[U_{s0} \Sigma_0 v_{0t}^T \left(1 - \frac{1 - T \langle V_0^T \rangle \langle V_0^T \rangle}{T \langle (v_0^T)^2 \rangle} \right) \right] \tag{A.57}$$

$$+ \left[\sum_{f>0} U_{sf} \Sigma_f \left(v_{ft}^T + \frac{v_{0t}^T}{\langle (v_0^T)^2 \rangle} \langle V_0^T \rangle \langle V_f^T \rangle \right) \right] \tag{A.58}$$

$$= \left[U_{s0}\Sigma_0 v_{0t}^T \left(1 - \frac{\frac{1}{T} - \left\langle V_0^T \right\rangle^2}{\langle (V_0^T)^2 \rangle - \left\langle V_0^T \right\rangle^2} \right) \right] \tag{A.59}$$

$$+ \left[\sum_{f>0} U_{sf}\Sigma_f \left(V_{ft}^T - \left\langle V_f^T \right\rangle 1_t + \frac{T}{1 - T\left\langle V_0^T \right\rangle} \left(V_{0t}^T - \left\langle V_0^T \right\rangle 1_t \right) \left\langle V_0^T \right\rangle \left\langle V_f^T \right\rangle \right) \right] \tag{A.60}$$

$$= 0 + \left[\sum_{f>0} U_{sf}\Sigma_f \left(V_{ft}^T - \frac{1}{1 - T\left\langle V_0^T \right\rangle^2} \left(\left\langle V_f^T \right\rangle 1_t - TV_{0t}^T \left\langle V_0^T \right\rangle \left\langle V_f^T \right\rangle \right) \right) \right] \tag{A.61}$$

$$= \sum_{f>0} U_{sf}\Sigma_f \left(V_{ft}^T - \left\langle V_f^T \right\rangle \frac{1_t - TV_{0t}^T \left\langle V_0^T \right\rangle}{1 - T\left\langle V_0^T \right\rangle^2} \right) \tag{A.62}$$

Calling this new vector W:

$$W_{ft}^T = V_{ft}^T - \left\langle V_f^T \right\rangle \frac{1_t - TV_{0t}^T \left\langle V_0^T \right\rangle}{1 - T\left\langle V_0^T \right\rangle^2} \tag{A.63}$$

If we compute the norm of these vectors, we discover

$$\sum_t W_{at}^T W_{bt}^T = \sum_t \left[\left(V_{at}^T - \left\langle V_a^T \right\rangle \frac{1_t - TV_{0t}^T \left\langle V_0^T \right\rangle}{1 - T\left\langle V_0^T \right\rangle^2} \right) \left(V_{bt}^T - \left\langle V_b^T \right\rangle \frac{1_t - TV_{0t}^T \left\langle V_0^T \right\rangle}{1 - T\left\langle V_0^T \right\rangle^2} \right) \right] \tag{A.64}$$

$$= \sum_t \left[V_{at}^T V_{bt}^T - \beta \left\langle V_a^T \right\rangle V_{bt}^T - \beta \left\langle V_b^T \right\rangle V_{at}^T + \beta\alpha V_{bt}^T V_{0t}^T + \beta\alpha V_{at}^T V_{0t}^T \right. \tag{A.65}$$

$$\left. + \left\langle V_a^T \right\rangle \left\langle V_b^T \right\rangle \beta^2 \left(1_t - \alpha V_{0t}^T \right) \left(1_t - \alpha V_{0t}^T \right) \right] \tag{A.66}$$

$$= \delta_{ab} - 2\beta T \left\langle V_a^T \right\rangle \left\langle V_b^T \right\rangle + \left\langle V_a^T \right\rangle \left\langle V_b^T \right\rangle \beta^2 \left(T - T^2 \left\langle V_0^T \right\rangle^2 + T^2 \left\langle V_0^T \right\rangle^2 \right) \tag{A.67}$$

$$= \delta_{ab} - T \frac{\left\langle V_a^T \right\rangle \left\langle V_b^T \right\rangle}{1 - T\left\langle V_0^T \right\rangle^2} \tag{A.68}$$

So since the new vectors are not orthonormal, strictly speaking this is not the actual SVD that results. Looking at the formulas above, it becomes clear that if we work with centered data, the affect of the fitting procedure is trivial, it just removes the top singular vector and value, without disturbing the rest of the fit in any way.

Comparing the equations for α (A.24) and β (A.31) we see that if the data were considered centered we have:

$$\alpha_s = \langle D_s \rangle - \beta_s \langle M \rangle = 0 \tag{A.69}$$

$$\beta_s = \frac{\langle D_s M \rangle - \langle D_s \rangle \langle M \rangle}{\langle M^2 \rangle - \langle M \rangle^2} \tag{A.70}$$

$$= \frac{\langle D_s M \rangle}{\langle M^2 \rangle} \tag{A.71}$$

$$= \frac{\frac{1}{T} \sum_{f,t} U_{sf} \Sigma_f V_{ft}^T V_{0t}^T}{\frac{1}{T} \sum_t V_{0t}^T V_{0t}^T} \tag{A.72}$$

$$= U_{s0} \Sigma_0 \tag{A.73}$$

So that looking at equation (A.37), we see that the effect of the fitting procedure gives

$$D_{st} = D_{st} - U_{s0} \Sigma_0 V_{0t}^T \tag{A.74}$$

A.5 The Major Sector Classification Systems

Industries are organized into groups for statistical purposes primarily based on their products sold or services offered. Tables A.1, A.2, and A.3 describe the top-level divisions as defined by major industrial classification systems.

Sector	Industry Groups
Energy	Energy
Materials	Materials
Industrials	Capital Goods, Commercial & Professional Services, Transport
Consumer Discretionary	Automobiles & Parts, Durables & Apparel, Hotels, Media, Retail
Consumer Staples	Food & Drug, Food, Beverage & Tobacco, Household & Personal
Financials	Banks, Diversified Financials, Insurance, Real Estate
Healthcare	Healthcare Equipment & Services, Pharmaceuticals & Biotech
IT	Tech Equipment, Software & Software & Services, Semiconductors
Telecommunication	Telecommunication Services
Utilities	Utilities

Table A.1: The S&P/MSCI Barra Global Industry Classification Standard (GICS) [13].

Economic Sector	Business Sector (or Industry Group)
Energy	(Coal, Oil & Gas, Related Equipment & Services, Renewable Energy)
Basic Materials	Chemicals, Minerals Resources, Applied Resources
Industrials	Industrial Goods, Industrial Services, Conglomerates, Transportation
Cyclical	Automobiles & Parts, Cyclical Products Cyclical Services, Retailers
Non-Cyclical	Food & Beverage, Food & Drug Retail Personal & Household Products and Services
Financials	Banking & Investment, Insurance, Real Estate, Investment Trust
Healthcare	Health Services, Pharmaceuticals & Medical Research
Technology	Tech Equipment, Software & IT Services
Telecommunication	Telecommunication Services
Utilities	(Electric, Natural Gas, Water & Other, Multiline)

Table A.2: The Thomson Reuters Business Classification (TRBC) [14].

Industry	Supersector (or Sector)
Oil & Gas	(Oil & Gas Producers, Oil Equipment, Services & Distribution, Alternative Energy)
Basic Materials	Chemicals, Basic Resources
Industrials	Construction & Materials, Industrial Goods & Services
Consumer Goods	Automobiles & Parts, Food & Beverage, Personal & Household Goods
Healthcare	(Healthcare Equipment & Sevices, Pharmaceuticals & Biotechnology)
Consumer Services	Retail, Media, Travel & Leisure
Telecommunications	(Fixed Line Telecom, Mobile Telecom)
Utilities	Electricity, Gas, Water & Multiutilities
Financials	Banks, Insurance, Real Estate, Financial Services, Equity/Non-Equity Investment Instruments
Technology	(Software & Computer Services, Hardware & Equipment)

Table A.3: The Dow Jones/FTSE Industry Classification Benchmark (ICB) [15].

REFERENCES

1. R. N. Gutenkunst, J. J. Waterfall, F. P. Casey, K. S. Brown, C. R. Myers, and J. P. Sethna, "Universally sloppy parameter sensitivities in systems biology models," *PLoS Comput Biol* **3** no. 10, (10, 2007) e189. viii, x, 2, 3, 13, 28, 31, 86
2. "Nmist database." http://yann.lecun.com/exdb/mnist/. Accessed: 2013-11-27. viii, 9
3. L. Hayden, A. A. Alemi, and J. P. Sethna, "Unpublished results,". viii, 9
4. M. K. Transtrum, B. B. Machta, and J. P. Sethna, "Geometry of nonlinear least squares with applications to sloppy models and optimization," *Phys. Rev. E* **83** (Mar, 2011) 036701. x, 3, 13, 22, 31, 49, 86
5. K. S. Brown, C. C. Hill, G. A. Calero, C. R. Myers, K. H. Lee, J. P. Sethna, and R. A. Cerione, "The statistical mechanics of complex signaling networks: nerve growth factor signaling," *Physical Biology* **1** no. 3, (2004) 184. x, 31
6. R. Chachra, M. K. Transtrum, and J. P. Sethna, "Structural susceptibility and separation of time scales in the van der pol oscillator," *Phys. Rev. E* **86** (Aug, 2012) 026712. http://link.aps.org/doi/10.1103/PhysRevE.86.026712. x, 5, 12, 31
7. J. Locke, P. Westermark, A. Kramer, and H. Herzel, "Global parameter search reveals design principles of the mammalian circadian clock," *BMC Systems Biology* **2** no. 1, (2008) 22. x, 31
8. J. J. Waterfall, F. P. Casey, R. N. Gutenkunst, K. S. Brown, C. R. Myers, P. W. Brouwer, V. Elser, and J. P. Sethna, "Sloppy-model universality class and the vandermonde matrix," *Phys. Rev. Lett.* **97** (Oct, 2006) 150601. x, 3, 28, 30, 31, 86, 88
9. D. Sagan and J. Smith, "The tao accelerator simulation program," in *Proceedings of the Particle Accelerator Conference*, pp. 4159–4161. 2005. x, 31
10. "Project website with additional figures and analyses." www.lassp.cornell.edu/sethna/Finance. Accessed: 2013-11-28. xii, xiv, 60, 62, 64, 67, 70
11. "Scottrade." www.scottrade.com. Accessed: 2013-11-28. xi, xii, xiii, xiv, 56, 66, 67, 77, 80, 81, 82
12. V. Plerou, P. Gopikrishnan, B. Rosenow, L. A. N. Amaral, T. Guhr, and H. E. Stanley, "Random matrix approach to cross correlations in financial data," *Phys. Rev. E* **65** (Jun, 2002) 066126. http://link.aps.org/doi/10.1103/PhysRevE.65.066126. xiii, 57, 72, 81
13. "GICS MSCI." http://www.msci.com/products/indices/sector/gics/. Accessed: 2013-11-27. xiv, 59, 97
14. "TRBC." http://thomsonreuters.com/business-classification/. Accessed:

2013-11-27. xiv, 59, 97

15. "ICB." http://www.icbenchmark.com/. Accessed: 2013-11-27. xiv, 59, 98
16. M. K. Transtrum and P. Qiu *Submitted* (2013) . 4
17. v. d. Pol and J. v. d. Mark, "Frequency demultiplication," *Nature* **120** (1927) 363–364. 4
18. B. van der Pol, "On relaxation-oscillations," *The London, Edinburgh and Dublin Phil. Mag. & J. of Sci.* **2** no. 7, (1927) 978–992. 4, 14
19. R. FitzHugh, "Impulses and physiological states in theoretical models of nerve membrane," *Biophysical Journal* **1** no. 6, (1961) 445–466. 4
20. J. Nagumo, S. Arimoto, and S. Yoshizawa, "An active pulse transmission line simulating nerve axon," *Proceedings of the IRE* **50** no. 10, (1962) 2061–2070. 4
21. J. H. E. Cartwright, V. M. Eguiluz, E. Hernandez-Garcia, and O. Piro, "Dynamics of elastic excitable media," *International Journal of Bifurcation and Chaos* **09** no. 11, (1999) 2197–2202. http://www.worldscientific.com/doi/abs/10.1142/S0218127499001620. 4
22. B. B. Machta, R. Chachra, M. K. Transtrum, and J. P. Sethna, "Parameter space compression underlies emergent theories and predictive models," *Science* **342** no. 6158, (2013) 604–607, http://www.sciencemag.org/content/342/6158/604.full.pdf. http://www.sciencemag.org/content/342/6158/604.abstract. 8, 27
23. P. Zikopoulos, D. deRoos, K. Parasuraman, T. Deutsch, J. Giles, and D. Corrigan, *Harness the Power of Big Data – The IBM Big Data Platform*. Mcgraw-Hill, 2012. 8
24. "Apple iOS Siri." http://www.apple.com/ios/siri. Accessed: 2013-11-27. 8
25. "Sebastian Thrun: Google's driverless car." http://www.ted.com/talks/sebastian_thrun_google_s_driverless_car.html. Accessed: 2013-11-27. 8
26. "IBM Watson." http://www.ibm.com/watson. Accessed: 2013-11-27. 8
27. "Netflix Prize." http://www.netflixprize.com/. Accessed: 2013-11-27. 8
28. F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research* **12** (2011) 2825–2830. 8
29. K. S. Brown and J. P. Sethna, "Statistical mechanical approaches to models with many poorly known parameters," *Phys. Rev. E* **68** (Aug, 2003) 021904. 13
30. F. P. Casey, D. Baird, Q. Feng, R. Gutenkunst, J. J. Waterfall, C. Myers, K. S. Brown, R. A. Cerione, and J. Sethna, "Optimal experimental design in an epidermal growth factor receptor signalling and down-regulation model," *Systems Biology, IET* **1** no. 3, (2007) 190–202. 13, 35
31. R. N. Gutenkunst, F. P. Casey, J. J. Waterfall, C. R. Myers, and J. P. Sethna, "Extracting

falsifiable predictions from sloppy models," *Annals of the New York Academy of Sciences* **1115** no. 1, (2007) 203–211. 13, 24

32. M. K. Transtrum and J. P. Sethna, "Improvements to the levenberg-marquardt algorithm for nonlinear least-squares minimization,". 13

33. J. F. Apgar, D. K. Witmer, F. M. White, and B. Tidor, "Sloppy models, parameter uncertainty, and the role of experimental design," *Molecular Biosystems* **6** no. 10, (2010) 1890–1900. 13, 24, 35

34. M. Secrier, T. Toni, and M. P. H. Stumpf, "The ABC of reverse engineering biological signalling systems," *Mol. BioSyst.* **5** (2009) 1925–1935. 13

35. A. Dayarian, M. Chaves, E. D. Sontag, and A. M. Sengupta, "Shape, size, and robustness: Feasible regions in the parameter space of biochemical networks," *PLoS Comput Biol* **5** no. 1, (01, 2009) e1000256. 13

36. H. Hettling and J. H. van Beek, "Analyzing the functional properties of the creatine kinase system with multiscale 'sloppy' modeling," *PLoS Comput Biol* **7** no. 8, (08, 2011) e1002130. 14

37. C. K. Jones and A. I. K. (Eds.), *Multiple-time-scale dynamical systems*. Springer, 2000. 14

38. J. Grasman, *Asymptotic Methods for Relaxation Oscillations and Applications*. Springer Press, 1987. 14, 15

39. S. H. Strogatz, *Nonlinear Dynamics and Chaos: With Applications to Physics, Biology, Chemistry and Engineering*. Westview Press, 2001. 15

40. A. Katok and B. Hasselblatt, *Introduction to the Modern Theory of Dynamical Systems*. Cambridge University Press, 1997. 17, 18

41. J. M. Guckenheimer and P. Holmes, *Nonlinear Oscillations, Dynamical Systems and Bifurcations of Vector Fields*. Springer Press, 1983. 18

42. J. Murdock, *Normal Forms and Unfoldings for Local Dynamical Systems*. Springer, New York, 2003. 18

43. R. N. Gutenkunst, J. C. Atlas, F. P. Casey, R. S. Kuczenski, J. J. Waterfall, C. R. Myers, and J. P. Sethna, "Sloppycell `http://sloppycell.sourceforge.net`,". 20

44. C. R. Myers, R. N. Gutenkunst, and J. P. Sethna, "Python unleashed on systems biology," *Computing in Science Engineering* **9** no. 3, (2007) 34 –37. 20

45. W. H. Press, S. A. Teukolsky, W. T. Vetterling, and B. P. Flannery, *Numerical Recipes 3rd Edition: The Art of Scientific Computing*. Cambridge University Press, New York, NY, USA, 3 ed., 2007. 22, 70, 71, 74

46. M. Schmidt and H. Lipson, "Distilling free-form natural laws from experimental data," *Science* **324** no. 5923, (2009) 81–85. 24, 28

47. R. Chachra, M. K. Transtrum, and J. P. Sethna, "Comment on "Sloppy models,

parameter uncertainty, and the role of experimental design"," *Molecular Biosystems* **7** no. 8, (2011) 2522. 24

48. B. B. Machta, *Criticality in Cellular Membranes and the Information Geometry of Simple Models*. Cornell University Press, Ithaca, New York, 2013. 27

49. E. P. Wigner, "The unreasonable effectiveness of mathematics in the natural sciences. richard courant lecture in mathematical sciences delivered at new york university, may 11, 1959," *Communications on Pure and Applied Mathematics* **13** no. 1, (1960) 1–14. 28

50. P. W. Anderson, "More is different," *Science* **177** no. 4047, (1972) 393–396. 28

51. U. Alon, "Simplicity in biology," *Nature* **446** no. 7135, (MAR 29, 2007) 497. 28

52. G. J. Stephens, B. Johnson-Kerner, W. Bialek, and W. S. Ryu, "Dimensionality and dynamics in the Behavior of C-elegans," *PLoS Computational Biology* **4** no. 4, (APR, 2008) .

53. G. J. Stephens, L. C. Osborne, and W. Bialek, "Searching for simplicity in the analysis of neurons and behavior," *Proceedings Of the National Academy of Sciences* **108** no. 3, (SEP 13, 2011) 15565–15571.

54. T. Sanger, "Human arm movements described by a low-dimensional superposition of principal components," *Journal of Neuroscience* **20** no. 3, (FEB 1, 2000) 1066–1072.

55. F. Corson and E. D. Siggia, "Geometry, epistasis, and developmental patterning," *Proceedings of the National Academy of Sciences* **109** no. 15, (2012) 5568–5575.

56. T. Mora and W. Bialek, "Are Biological Systems Poised at Criticality?," *Journal of Statistical Physics* **144** no. 2, (JUL, 2011) 268–302. 28

57. M. K. Transtrum, B. B. Machta, and J. P. Sethna, "Why are nonlinear fits to data so challenging?," *Phys. Rev. Lett.* **104** no. 6, (Feb, 2010) 060201. 28, 49, 86

58. S. Amari and H. Nagaoka, *Methods of Information Geometry*. Translations of Mathematical Monographs. American Mathematical Society, 2000. 29, 83, 85

59. I. Myung, V. Balasubramanian, and M. Pitt, "Counting probability distributions: Differential geometry and model selection," *Proceedings of the National Academy of Sciences* **97** no. 21, (OCT 10, 2000) 11170–11175.

60. V. Balasubramanian, "Statistical inference, Occam's razor, and statistical mechanics on the space of probability distributions," *Neural Computation* **9** no. 2, (FEB 15, 1997) 349–368. 29

61. G. Arfken and H. Weber, *Mathematical Methods for Physicists*. International paper edition. Academic Press, 2001. 38

62. P. M. Chaikin and T. C. Lubensky, *Principles of Condensed Matter Physics*. Cambridge University Press, Cambridge, 1995. 34

63. S. L. Veatch, O. Soubias, S. L. Keller, and K. Gawrisch, "Critical fluctuations in domain-forming lipid mixtures," *Proc. Natl. Acad. Sci. USA* **104** no. 45, (2007)

17650–17655. 34

64. G. E. Crooks, "Measuring thermodynamic length," *Phys. Rev. Lett.* **99** no. 10, (2007) . 34, 87

65. J. Cardy, *Scaling and Renormalization in Statistical Physics*. Cambridge University Press, 1996. 34

66. G. Ruppeiner, "Riemanniam geometry in thermodynamic fluctuation theory," *Reviews Of Modern Physics* **67** no. 3, (JUL, 1995) . 34, 41, 87

67. D. Ron, R. Swendsen, and A. Brandt, "Inverse Monte Carlo renormalization group transformations for critical phenomena," *Physical Review Letters* **89** no. 27, (DEC 30, 2002) . 35, 46, 51

68. U. Wolff, "Collective monte carlo updating for spin systems," *Phys. Rev. Lett.* **62** (Jan, 1989) 361–364. 41, 51

69. D. C. Brody and A. Ritz, "Information geometry of finite ising models," *Journal of Geometry and Physics* **47** (2003) 207. 41

70. M. Caselle, M. Hasenbusch, A. Pelissetto, and E. Vicari, "Irrelevant operators in the two-dimensional Ising model," *Journal of Physics A* **35** no. 23, (2002) 4861–4888. 49

71. "Russell 3000 index." www.russell.com/indexes/data/fact_sheets/us/russell_3000_index.asp. Accessed: 2013-11-28. 55

72. "S&P 500 Index." us.spindices.com/indices/equity/sp-500. Accessed: 2013-11-28. 55

73. "Dow Jones US Indices: Industry Indices." www.djindexes.com/mdsidx/downloads/fact_info/Dow_Jones_US_Indices_Industry_Indices_Fact_Sheet.pdf. Accessed: 2013-11-28. 55

74. "CBOE Oil Index." http://www.cboe.com/products/IndexComponentsAuto.aspx?PRODUCT=OIX. Accessed: 2013-11-28. 55

75. "Morgan Stanley High-Tech 35 Index." www.nasdaq.com/options/indexes/msh.aspx. Accessed: 2013-11-28. 55

76. D.-H. Kim and H. Jeong, "Systematic analysis of group identification in stock markets," *Phys. Rev. E* **72** (Oct, 2005) 046133. http://link.aps.org/doi/10.1103/PhysRevE.72.046133. 57

77. D. J. Fenn, M. A. Porter, S. Williams, M. McDonald, N. F. Johnson, and N. S. Jones, "Temporal evolution of financial-market correlations," *Phys. Rev. E* **84** (Aug, 2011) 026109. http://link.aps.org/doi/10.1103/PhysRevE.84.026109.

78. T. Conlon, H. Ruskin, and M. Crane, "Cross-correlation dynamics in financial time series," *Physica A: Statistical Mechanics and its Applications* **388** no. 5, (2009) 705 – 714. http://www.sciencedirect.com/science/article/pii/S0378437108008960.

79. C. Eom, G. Oh, H. Jeong, and S. Kim, "Topological properties of stock networks based on random matrix theory in financial time series," papers, arXiv.org, 2007. http://EconPapers.repec.org/RePEc:arx:papers:0709.2209. 57

80. R. Mantegna, "Hierarchical structure in financial markets," *The European Physical Journal B - Condensed Matter and Complex Systems* **11** no. 1, (1999) 193–197. http://dx.doi.org/10.1007/s100510050929. 57

81. G. Bonanno, N. Vandewalle, and R. N. Mantegna, "Taxonomy of stock market indices," *Phys. Rev. E* **62** (Dec, 2000) R7615–R7618. http://link.aps.org/doi/10.1103/PhysRevE.62.R7615.

82. G. Bonanno, G. Caldarelli, F. Lillo, and R. N. Mantegna, "Topology of correlation-based minimal spanning trees in real and model markets," *Phys. Rev. E* **68** (Oct, 2003) 046130. http://link.aps.org/doi/10.1103/PhysRevE.68.046130.

83. T. Heimo, K. Kaski, and J. SaramÃďki, "Maximal spanning trees, asset graphs and random matrix denoising in the analysis of dynamics of financial networks," *Physica A: Statistical Mechanics and its Applications* **388** no. 2âĂŞ3, (2009) 145 – 156. http://www.sciencedirect.com/science/article/pii/S037843710800839X.

84. N. Basalto, R. Bellotti, F. D. Carlo, P. Facchi, and S. Pascazio, "Clustering stock market companies via chaotic map synchronization," *Physica A: Statistical Mechanics and its Applications* **345** no. 1âĂŞ2, (2005) 196 – 206. http://www.sciencedirect.com/science/article/pii/S0378437104009872. 57

85. B. Podobnik and H. E. Stanley, "Detrended cross-correlation analysis: A new method for analyzing two nonstationary time series," *Phys. Rev. Lett.* **100** (Feb, 2008) 084102. http://link.aps.org/doi/10.1103/PhysRevLett.100.084102. 57

86. A. C. Martins, "Random, but not so much a parameterization for the returns and correlation matrix of financial time series," *Physica A: Statistical Mechanics and its Applications* **383** no. 2, (2007) 527 – 532. http://www.sciencedirect.com/science/article/pii/S0378437107002166. 57

87. T. Bury, "Market structure explained by pairwise interactions," *Physica A: Statistical Mechanics and its Applications* **392** no. 6, (2013) 1375 – 1385. http://www.sciencedirect.com/science/article/pii/S0378437112009685. 57

88. G. Doyle and C. Elkan, "Financial topic models," in *NIPS Workshop on Applications for Topic Models: Text and Beyond*. Whistler, Canada, 2009. 57

89. L. Laloux, P. Cizeau, J.-P. Bouchaud, and M. Potters, "Noise dressing of financial correlation matrices," *Phys. Rev. Lett.* **83** (Aug, 1999) 1467–1470. http://link.aps.org/doi/10.1103/PhysRevLett.83.1467. 58, 72

90. D. Nadig and L. Crigger, "Signal from noise," *Journal of Indexes* **14** (March, 2011) 40–43, 50. 59

91. "Berry Petroleum Company History." http://www.bry.com/pages/history.html. Accessed: 2013-11-28. 63

92. "Plum Creek History." http://www.plumcreek.com/AboutPlumCreek/History/tabid/55/Default.aspx. Accessed: 2013-11-28. 63

93. "Yahoo! Finance." finance.yahoo.com. Accessed: 2013-11-28. 67

94. A. Cutler and L. Breiman, "Archetypal analysis," *Technometrics* **36** (Nov, 1994) 338–347. 68

95. P. Tsalmantza and D. W. Hogg, "A data-driven model for spectra: Finding double redshifts in the sloan digital sky survey," *The Astrophysical Journal* **753** no. 2, (2012) 122. http://stacks.iop.org/0004-637X/753/i=2/a=122. 68

96. Z. Zhang, T. Li, C. Ding, and X. Zhang, "Binary matrix factorization with applications," in *Proceedings of the 2007 Seventh IEEE International Conference on Data Mining*, ICDM '07, pp. 391–400. IEEE Computer Society, Washington, DC, USA, 2007. http://dx.doi.org/10.1109/ICDM.2007.99. 68

97. C. Ding and X. He, "K-means clustering via principal component analysis," in *Proceedings of the Twenty-first International Conference on Machine Learning*, ICML '04, pp. 29–. ACM, New York, NY, USA, 2004. http://doi.acm.org/10.1145/1015330.1015408. 68

98. C. Thurau, K. Kersting, and C. Bauckhage, "Yes we can—simplex volume maximization for descriptive web scale matrix factorization.," in *CIKM*, J. Huang, N. Koudas, G. J. F. Jones, X. Wu, K. Collins-Thompson, and A. An, eds., pp. 1785–1788. ACM, 2010. 68

99. A. Hyvärinen and E. Oja, "Independent component analysis: Algorithms and applications," *Neural Netw.* **13** no. 4-5, (May, 2000) 411–430. http://dx.doi.org/10.1016/S0893-6080(00)00026-5. 68

100. D. D. Lee and H. S. Seung, "Learning the parts of objects by non-negative matrix factorization," *Nature* **401** no. 6755, (Oct., 1999) 788–791. http://dx.doi.org/10.1038/44565. 68

101. Y.-X. Wang and Y.-J. Zhang, "Nonnegative matrix factorization: A comprehensive review," *Knowledge and Data Engineering, IEEE Transactions on* **25** no. 6, (2013) 1336–1353. 68

102. C. H. Q. Ding, T. Li, and M. I. Jordan, "Convex and semi-nonnegative matrix factorizations," *IEEE Trans. Pattern Anal. Mach. Intell.* **32** no. 1, (Jan., 2010) 45–55. http://dx.doi.org/10.1109/TPAMI.2008.277. 68

103. C. Thurau, K. Kersting, M. Wahabzada, and C. Bauckhage, "Convex non-negative matrix factorization for massive datasets," *Knowledge and Information Systems* **29** no. 2, (2011) 457–478. http://dx.doi.org/10.1007/s10115-010-0352-6. 68

104. K. Kersting, M. Wahabzada, C. Thurau, and C. Bauckhage, "Hierarchical convex nmf for clustering massive data.," *Journal of Machine Learning Research - Proceedings Track* **13** (2010) 253–268. http://dblp.uni-trier.de/db/journals/jmlr/jmlrp13.html. 68

105. T. Li and C. Ding, "The relationships among various nonnegative matrix factorization methods for clustering," *Data Mining, 2006. ICDM '06. Sixth International Conference on* (2006) 362–371. 68

106. C. Thurau, K. Kersting, and C. Bauckhage, "Convex non-negative matrix factorization in the wild," in *Data Mining, 2009. ICDM '09. Ninth IEEE International Conference on*, pp. 523–532. 2009. 68

107. M. Mörup and L. K. Hansen, "Archetypal analysis for machine learning and data mining," *Neurocomputing* **80** no. 0, (2012) 54 – 63. 68, 69

108. M. L. Mehta, *Random Matrices*. Academic Press, Boston, MA, USA, 3 ed., 2004. 72

109. M. Tagiliani, *The Practical Guide to Wall Street*. John Wiley & Songs, Inc., Hoboken, NJ, USA, 1 ed., 2009. 76

110. L. Pastor, J. Heaton, and A. Foss, "The index is dead. Long live the index," *Journal of Indexes* **16** (July, 2013) 16–21, 55. 76

111. C. E. Shannon, "A mathematical theory of communication," *Bell System Technical Journal* **27** no. 3, (1948) 379–423. 83

112. T. Cover and J. Thomas, *Elements of Information theory*. Wiley Interscience, New York, NY, 1991. 83, 84

113. S. Kullback and R. Leibler, "On information and sufficiency," *Annals Of Mathematical Statistics* **22** no. 1, (1951) 79–86. 84

114. N. N. Cencov, *Statistical decision rules and optimal inference*. American Mathematical Society, 1981. 85

115. L. Campbell, "An extended äśencov characterization of the information metric," *Proceedings of the American Mathematical Society* **98** no. 1, (1986) 135–141. 85

116. H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proceedings of the Royal Society of London Series A* **186** no. 1007, (1946) 453. 85

117. M. Prokopenko, J. T. Lizier, O. Obst, and X. R. Wang, "Relating fisher information to order parameters," *Phys. Rev. E* **84** (2011) 041116. 87

118. C. Jarzynski, "Nonequilibrium equality for free energy differences," *Phys. Rev. Lett.* **78** no. 14, (1997) . 87