PATTERNS OF STRUCTURAL HIERARCHIES IN COMPLEX SYSTEMS

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

Katherine Nicholson Quinn

August 2019

PATTERNS OF STRUCTURAL HIERARCHIES IN COMPLEX SYSTEMS

Katherine Nicholson Quinn, Ph.D.

Cornell University 2019

Models of complex systems exhibit universal properties: there is a structural hierarchy of parameter importance. Where does this hierarchy come from? What do hierarchies say about model predictions, complex systems, and the way we make sense of phenomena? This thesis explores patterns in the complex, nonlinear models we construct to understand physical and social phenomena, with a focus on the structural hierarchy of parameter importance. Using information geometry, the problem of finding and explaining patterns in models and data is translated to one of finding structure in high-dimensional geometric objects, known as model manifolds (representing the space of all possible model predictions or all data). The structural hierarchy of parameter importance is turned into a geometric hierarchy of lengths and widths of these manifolds.

In the first part of the thesis, we use approximation theory to connect the underlying smoothness of models to bounds on their corresponding model manifolds, explaining global hierarchical structure. Our approach results in universal bounds on model predictions for classes of smooth models, capturing global geometric features that are intrinsic to their model manifolds. We illustrate these ideas using three disparate models from three different fields: exponential decay (physics), reaction rates from an enzyme-catalysed chemical reaction (chemistry), and an epidemiology model of an infected population (biology).

In the second part, we derive a new manifold learning technique called InPCA to obtain low-dimensional visualizations of the manifolds of general,

probabilistic models and data that reveal properties of their corresponding manifolds. Using replicas to tune dimensionality in high-dimensional data, we consider the zero-replica limit to discover a distance metric which preserves distinguishability in high dimensions, and an embedding with superior visualization performance. We apply InPCA to several probabilistic models, including the finite two-dimensional Ising model of atomic spins, a trained convolutional neural network, and the model of cosmology which predicts the angular power spectrum of the cosmic microwave background allowing visualization of the space of model predictions (*i.e.* different universes).

Finally, in the third part of this thesis, we use the tools of dimensional reduction combined with advanced statistical tests to analyse the results of a study in which we quantified student behaviours in the labs of an introductory calculus based physics course. Specifically, we analyzed gendered differences in participation in these labs. We followed 143 students across multiple lab periods in two pedagogically different lab types, and performed a cluster analysis to identify different categories of student behaviour. We found that in labs designed to foster collaborative group work and promote student decision making, there was a task division along gender lines with respect to laptop and equipment use (and found no such divide among students in more guided verification labs). Specifically, women handled laptops more than men and men behaved differently depending on whether they were in mixed-gender or single-gender groups. Students were not overtly assigned tasks, and the only explicit instruction from one student to another was in the form of quick, directed comments: the gendered division of tasks at the class level was not the result of overt task allocation but rather the accumulation of subtle interactions.

BIOGRAPHICAL SKETCH

The author was born in 1991 in Montreal, Canada. After moving around the province of Quebec in childhood, she came back to Montreal to pursue a Bachelors of Science from McGill University with a double major in mathematics and physics, and graduated in 2013. She then moved to Ithaca, New York to pursue her doctoral studies in physics at Cornell University.

PREFACE

The majority of work presented in this thesis is the result of multiple collaborations and is either currently published, in publication, under review, or will be submitted for publication in peer-reviewed journals in the (hopefully) not too distant future.

Chapter 3 reflects work done in collaboration with Heather Wilber, Alex Townsend and James Sethna. The majority of this work has been published in Physical Review Letters [122], and a preprint is available on the arXiv [121]. Katherine Quinn was first author, however her coauthors were responsible for essential components. Heather Wilber is responsible for the derivations in Section 3.1.1, Section 3.3.1, Apppendix A.4 and Appendix A.5, formalized the extension to non-analytic cases in Appendix C, and generated Fig. A.1. Heather Wilber and Alex Townsend generated the proof for Theorem 2. Alex Townsend and James Sethna supervised the project. All collaborators were involved in writing the published manuscript. Funding was provided by NSF DMR-1719490, NSF DMS-1818757, NSF DGE-1650441 and NSERC PGS-D.

Chapter 4 reflects work done in collaboration with Colin Clement, Francesco De Bernardis, Michael Niemack and James Sethna. The majority of this work has been accepted for publication, and a preprint is available on the arXiv [118]. Katherine Quinn was first author, however her coauthors were responsible for essential components. Colin Clement provided expertise in machine learning. Francesco De Bernardis and Michael Niemack provided expertise on the CMB. James Sethna supervised the project. All collaborators were involved in writing the accepted manuscript. Funding provided by the NSF DMR-1312160, NSF DMR-1719490, NSF AST-1454881 and NSERC PGS-D.

Chapter 5 reflects work done in collaboration with Michelle Kelley, Kathryn

McGill, Zachary Whipps, Emily Smith and Natasha Holmes. Katherine Quinn was first author, however her coauthors were responsible for essential components. Michelle Kelley and Kathryn McGill collected data through in-lab observations and were involved in research designed. Zachary Whipps coded single-group videos. Emily Smith recorded all videos and advised on research design. Natasha Holmes supervised the project. Approval for the study was obtained through IRB number 1802007733. A portion of the cluster analysis is currently published in the peer-reviewed proceedings of the 2018 Physics Education Research Conference [119] with a preprint available on the arXiv [120], and a longer manuscript will be submitted with all collaborators involved in writing. Funding was provided by PCCW's Affinito-Stewart Grant and Cornell's College of Arts and Sciences ALI.

ACKNOWLEDGEMENTS

There are a great many people who supported me during my time as a graduate student and made this thesis possible. To quote one of my students (who was an inmate at Five Points Correctional Facility, and for numerous reasons needs to remain anonymous):

> You gotta thank people who give you their time. It's the one thing they ain't never getting back.

First and foremost, I want to thank my advisers, Jim Sethna and Natasha Holmes, for their invaluable support in this process and patience when I invested my time in non-physics pursuits. Jim's unwavering enthusiasm for cool science is highly contagious, and his passion for connecting seemingly disparate fields has drastically shaped my view of research in physics. Natasha opened up new, profound avenues of research for me and her perfect combination of brilliance and dedication to the field are an inspiration. I want to thank the members of Jim's research group with whom I overlapped, Danilo Liarte, Alex Alemi, Matt Bierbaum, Lorien Hayden, Colin Clement, Archishman Raju, Jaron Kent-Dobias, David Hathcock, and Alen Senanian, as well as the members of Natasha's research group with whom I overlapped, Emily Smith, Cole Walsh, Martin Stein, Jack Madden, Ryan Tapping, Michelle Kelley, Kathryn McGill, Scott Allen, Saaj Chattopadhyay, Tim Rehm, and Zachary Whipps, for their interesting and illuminating conversations. I have collaborated on research projects with many of Jim's and Natasha's group members, and it has been a great pleasure to do so.

I want to thank my collaborators on physics project, including Michael Niemack, Francesco De Bernardi, Heather Wilber, Alex Townsend, Qingyang Xu, Ruoshui Wang, and Han Kheng Teoh. I have also had fruitful conversa-

tions with Mark Transtrum, Ben Machta, and Christopher Gosling, as well as members of the CPERL journal club, in particular Peter Lepage, Michelle Smith, Claire Madders, and Frank Castelli. I wish to thank my committee members Liam McAllister and Julia Thom for their thought-provoking A-exam questions.

A formative experience for me during my time at Cornell has been teaching. I wish to thank Alan Giambattista and Dorothy Holland-Minkley for helping me in my first formal teaching position in the auto-tutorial when I arrived at Cornell as a graduate student. I thank all members of Cornell's Prison Education Program, in particular Rob Scott, Jessica Brewer, and Nicole Peppin for handling the intricate nuances of coordinating prison logistics, and my TAs Oscar Boochever, Quincy Blair, and Stephen Wiley for improving my course. I especially thank my students, at Cornell, Five Points Correctional Facility and Cayuga Correctional Facility, for revealing to me simultaneously how much and how little I know about physics.

I have been involved in many student organizations at Cornell. I wish to thank all members of Cornell's OISE, in particular Sara Xayarath Hernandez, Alyssa Lopez, and Anitra Douglas-McCarthy, for their tireless efforts to improve campus climate. I wish to thank oSTEM members, in particular Hao Shi and Kasia Oleske, for getting me involved in student organizations on campus. I wish to thank all GPSA and UA members, in particular Elena Michel and Ekarina Winarto for their superhuman abilities. I wish to thank all GWIS members, in particular Natalie Hofmeister, Gwendolyn Beacham, Michelle Kelley, Aubrie James, Lauren McLeod, Lauren Genova, Lorien Hayden, and Jocienne Nelson, for many memorable organizing meetings and ambitious projects. I wish to thank WIP members for great coffee hours and conversations, and in particular Jenny Wurster, Ritchie Patterson, Patricia Sparks, Lauren McLeod, and Julia

Thom for organizing the 40th Anniversary WIP Reunion conference.

The six years I spent in graduate school were not easy, and so I thank my therapist Dr. Chang for our many sessions. Her professional insight helped me through some very difficult times.

I want to thank Philip Burnham, Andre Frankenthal, Hao Shi, Archishman Raju and Ti-Yen Lan for their friendship and lunches. I wish to thank Lauren McLeod for being an amazing roommate and friend. I thank Lorien Hayden and Jacob Boyett for their friendship and good times.

Finally, I thank Brendan Faeth, for his unwavering support of me over the past several years. He helped me grow into a better physicist and a better person.

I myself have never been able to find out what feminism is; I only know that people call me a feminist whenever I express sentiments that differentiate me from a doormat or a prostitute.

Rebecca West

TABLE OF CONTENTS

Biographical Sketch . . . iii
Preface . . . iv
Acknowledgements . . . vi
Table of Contents . . . x
List of Tables . . . xiii
List of Figures . . . xiv
List of Abbreviations . . . xvii
List of Symbols . . . xviii

1 Introduction **1**
1.1 Modeling Physical Phenomena . . . 4
1.2 Power and Authority in Institutions of Science . . . 6

2 Information Geometry **11**
2.1 Model Manifolds . . . 11
2.2 The Fisher Information Matrix . . . 16
2.2.1 Connection to Hessians . . . 17
2.2.2 Reparametrization . . . 17
2.3 Divergences . . . 18
2.4 Geometric Interpretations and Sloppy Models . . . 19
2.4.1 Local Properties . . . 20
2.4.2 Global Properties . . . 21
2.5 Manifold Learning . . . 24

3 Bounding Model Predictions **26**
3.1 Polynomial Approximations . . . 26
3.1.1 Chebyshev Expansions . . . 29
3.1.2 Taylor Expansions . . . 32
3.2 Examples . . . 33
3.2.1 Physics: Exponential Curves . . . 34
3.2.2 Chemistry: Reaction Velocities . . . 36
3.2.3 Biology: Infected Fraction of a Population . . . 36
3.3 Two-Dimensional Extension of Model Predictions . . . 37
3.3.1 Bounds on the 2D Extension . . . 40
3.4 Summary . . . 41

4 Visualizing Probability Distributions and the Intensive Embedding **43**
4.1 Model Manifolds of Probability Distributions . . . 44
4.2 Hypersphere Embedding . . . 45
4.3 Replica Theory . . . 47
4.3.1 Connection to Least Squares . . . 51
4.4 Intensive Embedding . . . 52

4.4.1 InPCA Algorithm . . . 54
4.5 Properties of the Intensive Embedding and InPCA . . . 55
4.6 Examples . . . 56
4.6.1 Coin Toss . . . 56
4.6.2 Gaussians . . . 59
4.6.3 Ising Model . . . 60
4.6.4 Neural Network MNIST Digit Classifier . . . 63
4.6.5 ΛCDM Predictions of the CMB . . . 63
4.6.6 MNIST Images . . . 72
4.7 Summary . . . 75

5 Quantifying Gendered Behavior in Physics labs 76
5.1 Poststructural Gender Theory . . . 78
5.2 Research Methods . . . 81
5.2.1 Collecting Demographic Information . . . 82
5.2.2 Quantifying Coarse Student Behaviours . . . 83
5.2.3 Describing Detailed Student Behaviour . . . 88
5.2.4 Rescaling Student Profiles . . . 92
5.2.5 Cluster Analysis . . . 93
5.3 Results . . . 97
5.4 Discussion . . . 102

6 Conclusions 106

A Tedious Calculations and Derivations 108
A.1 Fisher Information Matrix for Least-Squares Models . . . 108
A.2 Equivalent Representations of the Fisher Information Matrix . . . 109
A.3 Fisher Information Matrix and the Bhattacharyya Distance . . . 110
A.4 Bounding Non-Analytic Models . . . 111
A.5 Deriving Manifold Bounds from Monomial Basis . . . 112
A.6 Connection Between Intensive Distance and Least-Squares . . . 113
A.7 Deriving the Intensive Cross-Covariance Matrix . . . 114

B Random Matrix Theory and Fisher Information 115
B.1 Correlated Random Matrices . . . 118
B.1.1 Uncorrelated Entries . . . 121
B.1.2 Correlated Rows and Columns . . . 123
B.2 Sequential Random Matrices . . . 124

C Numerical Observations for High Dimensional Manifolds 132

D Generating and Visualizing Model Manifolds 135
D.1 Least Squares Models . . . 135
D.2 2D Ising Model . . . 139
D.3 Convolutional Neural Network with TensorFlow . . . 140

D.4 Cosmic Microwave Background 141

E InPCA Comparisons with t-SNE and Diffusion Maps **144**
E.1 Ising Model Manifold 144
E.2 Neural Network 145

F Parameter Degeneracies and Combinations in the CMB **149**

G Statistical Tests for Quantifying Behaviour **152**
G.1 Error Estimates 152
G.2 Statistical Tests 152
G.3 Example of Student Profile Rescaling 154
G.4 Effect of Student Group Sizes 156

H Different Cluster Analysis **159**

I InPCA Visualization of Student Behaviours **164**

J Video Coding **173**

Bibliography **179**

LIST OF TABLES

3.1 **Manifold relationships** between models and polynomial approximations. 27

5.1 **Student demographics** of participants in the study of quantified lab behaviours. 83
5.2 **Action codes** used to quantify lab bahaviours. 85
5.3 **Demographic breakdown** of student profiles observed in study of quantified lab behaviours. 87
5.4 **Event codes** used in video observations of lab behaviours. 91

G.1 **Sample contingency table** of two observers for inter-rater reliability. 153
G.2 **Sample contingency table** of the distribution of men's profiles in the inqiry labs. 154

LIST OF FIGURES

2.1 **Model manifolds** of two simple nonlinear models. 13
2.2 **The prediction space** of complex systems. 20
2.3 **Eigenvalues of local metric** for many disparate, nonlinear models. 21
2.4 **Normalized, stacked histogram** of ordered eigenvalues of local metrics. 22
2.5 **Manifold lengths** for many disparate, nonlinear models. 23

3.1 **Model manifolds** of three disparate models with universal bounds. 35
3.2 **Model manifolds** of three disparate models extended to two experimental conditions. 39

4.1 **Hypersphere embedding** of two dimensional Ising model. . . . 48
4.2 **Replicated Ising model** illustrating the derivation of the intensive embedding. 50
4.3 **InPCA visualization** of biased coins. 58
4.4 **Hypersphere embedding** of biased coins. 58
4.5 **InPCA visualization** of Gaussians. 60
4.6 **InPCA visualization** of the two-dimensional Ising model. 62
4.7 **InPCA visualization** of the stages of training a convolutional neural network (CNN). 64
4.8 **Different visualizations** of the model manifold of the six parameter dark energy cold dark matter (ΛCDM) cosmological model. 66
4.9 **ΛCDM parameter correlations** with InPCA components. 68
4.10 **ΛCDM parameter correlations** with t-SNE components. 69
4.11 **ΛCDM parameter correlations** with diffusion map components. 70
4.12 **Changes in CMB** temperature (TT) spectra with different parameters. 73
4.13 **Different visualizations** of the raw MNIST data. 74

5.1 **Schematic of the theoretical framework** used in study of lab behaviours. 81
5.2 **Set diagram** of the breakdown of labs used in study of gendered behaviour. 86
5.3 **Box plots** of raw data used in the study of lab behaviours. 88
5.4 **Breakdown of codes** and a decomposed student profile illustrating detailed behaviours. 90
5.5 **Elbow plot** used to determine optimal number of clusters. 95
5.6 **Two-dimensional visualization** of lab behaviour clusters and centers. 96
5.7 **Fraction of groups** with members in different clusters. 98
5.8 **Composition of lab behaviour clusters.** 99

A.1 **Singular values** of non-analytic models. 112

B.1 **Eigenvalue distributions** of local metrics for three disparate models. 116
B.2 **Numerical approximations** of the the FIM for three disparate models. 120
B.3 **Eigenvalue distributions** compared to randomly generated eigenvalues. 122
B.4 **The number of local peaks** in the eigenvalue distributions of sequential random matrices. 127
B.5 **The spacing between peaks** in the eigenvalue distributions of sequential random matrices. 128
B.6 **The spread in the peaks** in the eigenvalue distributions of sequential random matrices. 129
B.7 **Eigenvalue distributions of local metrics** compared to those of sequential random matrices. 131

C.1 **Bounds on the hyperellipsoid lengths** for high-dimensional manifolds. 134

D.1 **Histograms of valid parameter values** used to generate least-squares model manifolds. 138
D.2 **Ising parameter ranges** used to generate model manifolds. . . . 139
D.3 **Convolutional neural network** used to classify MNIST handwritten digits. 140

E.1 **Different visualizations** of the model manifold of the Ising model. 146
E.2 **Different visualizations** of the outputs of the trained neural network. 148

F.1 **ΛCDM parameter combination correlations** with InPCA components. 150

G.1 **Bar plot of code counts** from two observers in study of lab behaviours. 153
G.2 **Coding sheets** used to quantify student behaviour in labs. 155
G.3 **Stacked histogram** of the amount of coded time students spent in lab. 156
G.4 **Sample student profile** illustrating quantified student behaviour. 157
G.5 **Stacked histogram** of group sizes in labs. 158

H.1 **Elbow plot** of different groupings of student profiles. 160
H.2 **Cluster centers** of different groupings of student profiles. 162

I.1 **Qualitative and quantitative visualization** of student behaviour profiles. 165

I.2 **Ordered InPCA eigenvalues** of quantitative visualization of student behaviour profiles. 166
I.3 **InPCA visualization of student behaviour profiles** as compared to random data. 167
I.4 **InPCA visualization of student behaviour profiles** coloured by extracted regions from DBSCAN. 169
I.5 **Cluster breakdown** of student behaviour profiles overlapping with regions extracted from DBSCAN. 171

J.1 **Decomposed student profiles** in Bouncing Ball lab (Friday Group).174
J.2 **Decomposed student profiles** in Bouncing Ball lab (Friday Group).174
J.3 **Decomposed student profiles** in Bouncing Ball lab (Thursday Group). 175
J.4 **Decomposed student profiles** in Bouncing Ball lab (Thursday Group). 175
J.5 **Decomposed student profiles** in Hooke's law lab. 176
J.6 **Decomposed student profiles** in Pendulum lab (Thursday Group).176
J.7 **Decomposed student profiles** in Pendulum lab (Thursday Group).177
J.8 **Decomposed student profiles** in Project Lab (Thursday Group). 177
J.9 **Decomposed student profiles** in Project Lab (Friday Group). . . 178

LIST OF ABBREVIATIONS

ΛCDM Λ Cold Dark Matter.

AAPT American Association of Physics Teachers.
ALI Active Learning Initiative.

BCS Bardeen-Cooper-Schrieffer.

CAMB Code for Anisotropies in the Microwave Background.
CMB Cosmic Microwave Background.
CNN Convolutional Neural Network.

FIM Fisher Information Matrix.

InPCA Intensive Principal Component Analysis.

KL Kullback-Leibler.

MC Monte Carlo.
MDS Multidimensional Scaling.
MNIST Modified National Institute of Standards and Technology.

NSERC National Science and Engineering Research Council.
NSF National Science Foundation.

PCA Principal Component Analysis.
PCCW President's Council of Cornell Women.
PER Physics Education Research.

QCD Quantum Chromodynamics.

SIR Susceptible, Infected and Recovered.

t-SNE T-Distributed Stochastic Network Embedding.
TA Teaching Assistant.

UMAP Uniform Manifold Approximation and Projection.
UTA Undergraduate Teaching Assistant.

LIST OF SYMBOLS

A_α Amplitude term, specifically associated with exponential decay curves.

A_s Primordial fluctuation amplitude.

E_α Activation energy.

E_ρ Bernstein ellipse (ellipse in the complex plane).

H_0 Hubble constant.

H_P High-dimensional hyperellipsoid, bounding polynomial manifolds.

H_Y High-dimensional hyperellispoid, bounding model manifolds.

J Strength of nearest neighbor coupling in an Ising Model.

N_{tot} Total population.

$T_j(t)$ Chebyshev polynomial of degree j evaluated at point t, explicitly expressed as $\cos(j \arccos(t))$.

$\Omega_b h^2$ Physical baryon density.

$\Omega_c h^2$ Physical cold dark matter density.

β Rate of infection.

χ^2 Cost function.

ℓ_i Ordered hyperellipsoid lengths.

η Scalar spectral index.

γ Rate of recovery.

λ_α Decay rate, specifically associated with exponential curves.

λ_i Ordered Eigenvalue.

C_ℓ Correlation matrix for CMB fluctuations.

$\mathcal{F}$ Free energy of a system (in statistical physics) related to the partition function, given as $\frac{1}{T} \log \mathcal{Z}$.

$\mathcal{I}_{\alpha\beta}$ Fisher Information.

$\mathcal{L}$ Probability distribution or likelihood function.

$\mathcal{N}$ Normal or Gaussian distribution.

$\mathcal{P}$ Polynomial manifold, consisting of the set of polynomial predictions for all possible polynomial coefficients (model manifold for polynomials).

$\mathcal{Y}$ Model manifold, consisting of the set of model predictions for all possible model parameters.

$\mathcal{Z}$ Partition function.

∂_α Partial derivative with respect to model parameter θ^α, given explicitly as $\frac{\partial}{\partial \theta^\alpha}$.

ϕ_μ Function of spin states in Ising Hamiltonian.

ϕ_i Polynomial basis.

ρ Sum of semi-major and semi-minor axes of a Bernstein ellipse.

σ^2 Variance.

σ_i Ordered singular value.

τ Optical depth at reionization.

θ^α Parameters or inputs.

ξ Noise.

d^2 Squared Euclidean distance between points.

d_H^2 Hellinger divergence between two distributions (squared distance function).

d_I^2 Intensive distance between two probability distributions (squared distance function).

d_N^2 Hellinger divergence per N replicated distributions (squared distance function).

$g_{\alpha\beta}$ Metric.

h External magnetic field in an Ising Model.

p p-Value used to determine statistical significance.

p_{N-1} Polynomial of degree $N - 1$.

x_i Data or data points.

y_θ Model prediction for fixed parameters θ.

z Complex number.

z_i Square root of a state vector or probability distribution, explicitly expressed as $z_i = 2\sqrt{\mathcal{L}(x_i)}$.

CHAPTER 1

INTRODUCTION

Physics relies on an interplay between *reductionism* and *constructivist epistemology*. To understand a complicated system, break it down into its component parts and then see how the parts fit together, specifically through the construction of falsifiable models (that are then empirically tested). This philosophy forms the foundation for model construction in physics, one so widely accepted that we as physicists apply it without much question [7, 123]. However, even though we view systems as ultimately explainable in terms of their reduced parts, we are still able to construct practical models of complex systems without needing to understand all of their elementary pieces. Curiously, we need not fully understand the microscopic complexity of a system in order to practically model its behaviour (*e.g.* we do not need to know the positions and velocities of every individual water molecule to usefully predict a fluid's motion through a city's water supply[1]). In many ways, the entire field of statistical physics serves as an example of this fact. This raises several questions: what makes a system understandable and predictable? Is the fact that we can model systems without understanding their full complexity a reflection of the systems themselves, or us the researchers? Similarly, do common patterns in our models reflect an underlying pattern in the world around us, or in the way we construct models, or are such patterns a statement about systems we can understand?

To fruitfully bridge the scales between constituent parts and the complex systems they collectively create, the concept of *emergence* has been gaining popularity in recent decades [7]. Emergence is described by Kim [84] as the follow-

[1]Commendations are in order to the researcher who actually derives the Navier-Stokes equations for fluid mechanics directly from quantum field theories [52].

ing:

> As systems acquire increasingly higher degrees of organizational complexity they begin to exhibit novel properties that in some sense transcend the properties of their constituent parts.

An understanding of the elementary properties of a system, alone, is incomplete. I would also argue that, as one considers increasingly complex systems, it is in fact these emergent properties that matter most for effectively modelling the system (*e.g.* the electron coupling in the BCS theory of superconductivity does not hinge on the nature of the effective attractive force between electrons, merely that there is one). Because of this, there is a notion of *hierarchy* in complex systems, at least from the perspective of 'important features for pragmatic modelling'.

In this thesis, I explore patterns of structural hierarchies in complex systems through the lens of *information geometry*. This area of mathematics is used to translate the problem of finding patterns in models and data to one of finding structure in high-dimensional, geometric objects. I outline the important elements of information geometry for this thesis in Chapter 2.

Complex nonlinear models are typically ill-conditioned or *sloppy*; their predictions are significantly affected by only a small subset of parameter combinations, and parameters are difficult to reconstruct from model behavior. Despite forming an important universality class and arising frequently in practice when performing a nonlinear fit to data, formal and systematic explanations of sloppiness are lacking. By unifying geometric interpretations of sloppiness with

Chebyshev approximation theory in Chapter 3, we[2] rigorously explain sloppiness as a consequence of model smoothness. Our approach results in universal bounds on model predictions for classes of smooth models, capturing global geometric features that are intrinsic to their model manifolds, and characterizing a universality class of models. We illustrate this universality using three models from disparate fields (physics, chemistry, biology): exponential curves, reaction rates from an enzyme-catalysed chemical reaction, and an epidemiology model of an infected population.

In using information geometry to better understand model properties, we construct geometric objects know as *model manifolds* whose geometric features we then analyze. Unsupervised learning makes manifest the underlying structure of manifolds (and data more generally) without curated training and specific problem definitions. However, the inference of relationships between data points is frustrated by the 'curse of dimensionality' in high-dimensions. Inspired by replica theory from statistical mechanics, in Chapter 4 we consider replicas of the system to tune dimensionality and take the limit as the number of replicas goes to zero. The result is the intensive embedding, which is not only isometric (preserving local distances) but allows global structure to be more transparently visualized. We develop the Intensive Principal Component Analysis (InPCA) and demonstrate clear improvements in visualizations of the Ising model of magnetic spins, a neural network, and the dark energy cold dark matter model as applied to the Cosmic Microwave Background.

Emergence in complex systems describes how microscopic constituents come together to yield macroscopic phenomena. Groups of people are no ex-

[2]Because the majority of the work presented in this thesis is the result of collaborative projects, I will mostly use "we" to refer to "we who did this work".

ception to this, and so a natural research question is how collective human behaviour emerges from individual interactions. Physics education research offers a fruitful foundation from which to explore this question, where we look at how patterns in behaviour in physics labs is impacted by the interactions between individuals. In Chapter 5, we use poststructural gender theory and cluster analyses to identify patterns in student behaviors during lab instruction, patterns which relate to students' gender identities as well as those of their group members. We conduct additional analyses to understand those behaviors and further explore how they are impacted by the instructional context.

1.1 Modeling Physical Phenomena

John Von Neumann famously said [45]:

> With four parameters I can fit an elephant, and with five I can make him wiggle his trunk.

Complex nonlinear models used to simulate and predict experimentally observed phenomena often exhibit a structural hierarchy; perturbing a few model parameter combinations drastically impacts predictions, whereas most others can vary widely without effect. Such ill-conditioned models are called *sloppy*. Sloppy models appear to be common, arising in many areas of physics. In critical phenomena, this hierarchy of importance explains the parameter scaling with coarsening for diffusion and the Ising model of magnetism [99]. In accelerator physics, linear combinations of the multitude of tunable beam-line parameters exhibit a geometric hierarchy of importance [61]. Exponential curve fitting,

a notoriously ill-conditioned problem, poses a significant challenge, *e.g.* finding correlators in lattice QCD [92, 76]. Sloppy models are not confined to physics and in fact appear in systems biology [23, 22, 62], insect flight [17], power systems [151], machine learning [111], and many other areas [146]. Understanding why sloppiness occurs can therefore connect models used across disparate fields.

To better understand sloppiness, there are many well-studied cases for insensitivity of model predictions to particular combinations of parameters. *Structural identifiability* describes systems for which parameters can be analytically exchanged for one another [30, 125]. Separation of scales, singular perturbations, and continuum limits can make the behavior at a particular time or distance region depend only on a subset of the underlying parameters [83, 43, 96]. Universal critical behavior can yield effective parameter compression on long length scales near continuous transitions [99]. However, these comprehensible sources of sloppiness do not explain the generality of the phenomenon, nor do they offer a rigorous framework by which to quantify the hierarchy of parameter importance. In Chapter 3, we address the generic sloppiness of multiparameter nonlinear models in the absence of particular mechanisms or small parameters. We unify recently developed geometric descriptions of sloppiness [146] with classical ideas from polynomial approximation theory [154]. We posit that in many cases, sloppiness is fundamentally linked to the smoothness of the underlying model, and provide a rigorous description of this connection. Specifically, we use the smoothness of the underlying model to characterize a high-dimensional hyperellipsoid, which bounds the model manifold, and are able to quantify the various widths of this hyperellipsoid – and thus the allowed space for model manifolds.

1.2 Power and Authority in Institutions of Science

A great strength of physics is its foundation on empirical evidence. Researchers construct falsifiable theories, laws, and models that are subjected to rigorous empirical testing. As Richard Feynman famously said in his 1964 Messenger lectures at Cornell [50],

> It doesn't matter how beautiful your theory is, it doesn't matter how smart you are. If it doesn't agree with experiment, it's wrong.

By relying so strongly on experiments, physicists have crafted a rigorous framework by which to build practical models of physical phenomena. The widely successful nature of our technological advancements has led many to argue that this implies a *correspondence truth*: our models are getting closer and closer to a pragmatic *reflection* of the world as it really is [116, 75], or at least getting better and better at providing a pragmatic *interpretation* of the 'real world' relative to a certain purpose [129].

While this philosophy can be appealing, it is easy to forget that research is a human endeavour, and conducted within human institutions. In holding data detached from human elements as "objective" authority[3], we can fall into the trap of ignoring, dismissing, or even denying the existence of the subjective nature of research [49]. This can not only limit areas of research, ultimately undermining the very goals scientists set out the achieve, but damages the community as we conflate the authority of empirical evidence with those who present and interpret said evidence.

[3] As anyone who encounters the phrase "scientists claim" in news articles can attest.

A human element which guides research is *culture*. Many physicists describe the field as having a "culture of no culture"[152], that by conducting fundamental research that aims to remove the human element as much as possible, we have removed any cultural element from the field itself. However, there is a culture in physics, and it is constructed and reconstructed by those who do physics [58, 26]. Importantly, it is hierarchical, with senior researchers in a given discipline holding enormous prestige and sway in the community. This isn't to say social hierarchies serve no purpose: *functional accounts* [63] argues that hierarchical differentiation can facilitate group coordination by clearly defining roles [137], allowing the rapid flow of uniform information [6, 9], and creating patterns of deference. Given the highly technical and complex nature of physics fields, having designated experts can be advantageous for training, guiding and mentoring novel researchers (*e.g.* graduate school) and for assessing research in that field (*e.g.* who to reach out to for peer review).

However, there are tangible consequences to research when deference to authority occurs in science. Senior scientists have abused their positions of power to stifle progress in the field when it conflicts with their own work (such as Russell convincing his graduate student, Payne-Gaposchki, that her conclusions regarding the chemical compositions of stars were wrong [161]) or assumed credit for discoveries made by peers with less social standing (such as Watson and Crick assuming sole credit for the discovery of the DNA double helix and sidelining Franklin [14], Hewish and Ryle being awarded the Nobel prize for Burnell's discovery of recurring signals in pulsars [155], or Yang and Lee being awarded the Nobel prize for observing parity violations and sidelining the vital contributions of Wu [131]).

Moreover, there is a *scarcity of resources* in physics. There are limited faculty positions, as the proportion of people with science Ph.D.'s who manage to obtain a tenure-track position is dropping [36]. With the increase in supply of highly qualified candidates and demand for investment in infrastructure rather than labor, many universities are seeing an increase in adjunct positions instead of tenure-track positions[4] resulting in decreasing job security. In addition, there is limited funding for research (*e.g.* only about 23% of NSF funding applications are granted [108]). The increased pressure on academics fuels the "publish or perish" mantra, to the detriment of research[5] and researchers[6]. The combination of strong hierarchical systems with a scarcity of available resources for community members does *not* create egalitarian communities, facilitate the free and open exchange of ideas, nor promote moral and ethical research. The full impact of such consequences are beyond the scope of this discussion, as I would like to focus on a particular consequence, that of representation within the field.

What Feynman failed to note in his Messenger lectures was that *who you are* also has a huge impact on whether or not your theories and ideas are seen as 'correct' by the physics community. There is a large emphasis on innate ability and raw talent[7] in physics [93] (the so-called "lone genius effect" [105]), and strong cultural stereotypes abound regarding who has innate talent (in particular, white men [15, 59, 136]). An extreme notion of this philosophy, where

[4]The American Association of University Professors reports than now over 70% of full-time faculty are non-tenure track, up from 55% in 1975 https://www.aaup.org/issues/contingency/background-facts

[5]Not only does this fuel predatory, for-profit journals but stifles creativity. Nobel Laureate Higgs doubts he would have been productive enough to survive in the current research climate [2].

[6]There is a rising mental health crisis in academia, with graduate students more than six times likely to experience depression and anxiety as compared to the general population [48, 162]

[7]For a longer discussion on *fixed* versus *growth* mindsets in the context of STEM and educational systems, see [44, 73].

researchers actively set out to prove the supposed "natural" superiority of particular segments of the population, has yielded scientific blunders such as the IQ controversy [135]. As Traxler explains [153],

> Currently, many stereotypes abound in Western technological culture that relate to both science and sex differences; good scientists, and good men, are knowers, rational, and predictable. Women are framed as emotional, unpredictable, and thus irrational and poorly suited to science.

The culture of physics is highly masculine and androcentric [53, 66] and laden with masculine connotations [57]. Importantly, general attitudes of researchers in a discipline correlate with representation in the field [93]. In Leslie's study, 1820 faculty, postdoctoral fellows, and graduate students across 30 disciplines were asked to rate their agreement with statements such as "Being a top scholar of [discipline] requires a special aptitude that just can't be taught". Representation and belief in innate ability were correlated (in both STEM, and the humanities and social sciences): the stronger a discipline's association with the requirement of innate ability (as determined by members of that field), the fewer women were granted Ph.D.'s.

The composition of a research community has an effect on the research conducted, as a more diverse set of people allows for more diverse modes of thinking [58, 24, 65, 127, 128]. By limiting the diversity of physicists, we undermine the field and stifle growth. More importantly, can we truly call the physics community an equitable meritocracy that supports the free and open exchange of ideas if vast segments of the population are effectively barred from entering? This is not to say that physicists should abandon our field's foundation on rig-

orous empirical evidence, but rather that we need to apply our tools of rigorous criticality to our institutions and cultures as well as our immediate research. By rigorously analyzing our morals, values, and decision-making processes, and view this analysis not as an impediment to good science but rather an integral part of it, we can evolve our culture in a directed, deliberate manner, and fundamentally address the problems in the field. In doing so, we can decouple (and re-evaluate) notions of "natural" and "cultural" authority, and "natural" and "culturally expected" and ability, to address the hierarchical nature of the field and great imbalance in representation.

In exploring some of the underlying reasons and mechanisms for this imbalance in representation, in Chapter 5 we explore how gendered behavior manifested in physics labs. We noticed that groups in physics labs, when afforded the opportunity to divide tasks and roles between members, in fact did so along gender lines (with women handling laptops and performing data analysis more than men, and men handling equipment far more when working with other men). As universities implement pedagogical changes in their classrooms (such as the AAPT recommendations for undergraduate physics laboratory curriculum [86]), it is vital that we are as aware of the culture and dynamics of our institutions, and actively account for these effects.

CHAPTER 2

INFORMATION GEOMETRY

In this chapter, we introduce notation and concepts used throughout the rest of the thesis. Much of the methods and analysis rely on applications of *information geometry* [5, 4], an area of mathematics in which the problem of pattern recognition in complex data and models is translated to one of finding structure in high-dimensional, geometric objects. Here, topology and differential geometry meld with information theory and statistics in an elegant manner, revolving around *model manifolds* (Section 2.1) and metrics defined by the *Fisher information matrix* (Section 2.2).

In particular, the phenomena of ill-posed and well-constrained parameter combinations becomes manifest in 'long' and 'short' directions along the model manifold. We discuss the tangible connections in Section 2.4. Finally, in Section 2.5 we discuss different manifold learning techniques, and highlight the ones used in this thesis.

2.1 Model Manifolds

Model predictions have an elegant geometric interpretation. Given input parameters, models produce predictions. These predictions can be written out as a high dimensional vector (of finite or infinite length). Specifically, given some nonlinear model $y_{\boldsymbol{\theta}}(t)$ with model N parameters $\boldsymbol{\theta} = \{\theta^{\alpha}\}$ evaluated at points $(t_1, t_2, \dots)$, the set of all possible model predictions for all possible model param-

eters defines the model manifold[1]:

$$\mathcal{Y} = \{Y(\boldsymbol{\theta}) \mid \boldsymbol{\theta}\} \quad \text{where} \quad Y(\boldsymbol{\theta}) = (y_\theta(t_1), y_\theta(t_2), \dots). \tag{2.1}$$

The dimension of the manifold is determined by the number of input parameters (and so is an N dimensional space), and it is embedded in a space whose dimension is determined by the number of predictions (*i.e.* the number of points t_i, which may be finite or infinite). In particular, we can see the model as a mapping from parameter space to prediction space.

As an illustration, consider a trivially simple nonlinear model with two input parameters θ_1 and θ_2, given as

$$y_\theta(t) = \cos(\theta_1 t + \theta_2), \tag{2.2}$$

that we evaluate at three time points (t_1, t_2, t_3). The model manifold is therefore a two-dimensional object embedded in a three dimensional space, as shown in Fig. 2.1. The geometry of the manifold reflects properties of the model. Here, because of the periodic nature of cos, there are degeneracies in the model predictions (*e.g.* $\cos(\theta_1 t + \theta_2) = \cos(\theta_1 t + (\theta_2 + 2\pi))$) which is reflected by a winding in the manifold itself.

As a second illustration, consider another trivially simple nonlinear model with two input parameters θ_1 and θ_2 given as

$$y_{\boldsymbol{\theta}} = e^{-\theta_1 t} + e^{-\theta_2 t} \tag{2.3}$$

which we evaluate at three time points, (t_1, t_2, t_3). The boundaries of this manifold reflect important limits of the model, with either θ_1 or θ_2 going to infinity,

[1]This definition of a manifold does not satisfy the strict, mathematical requirement that every point on the manifold have a neighborhood that is homeomorphic to Euclidean space because we allow for features like cusps.

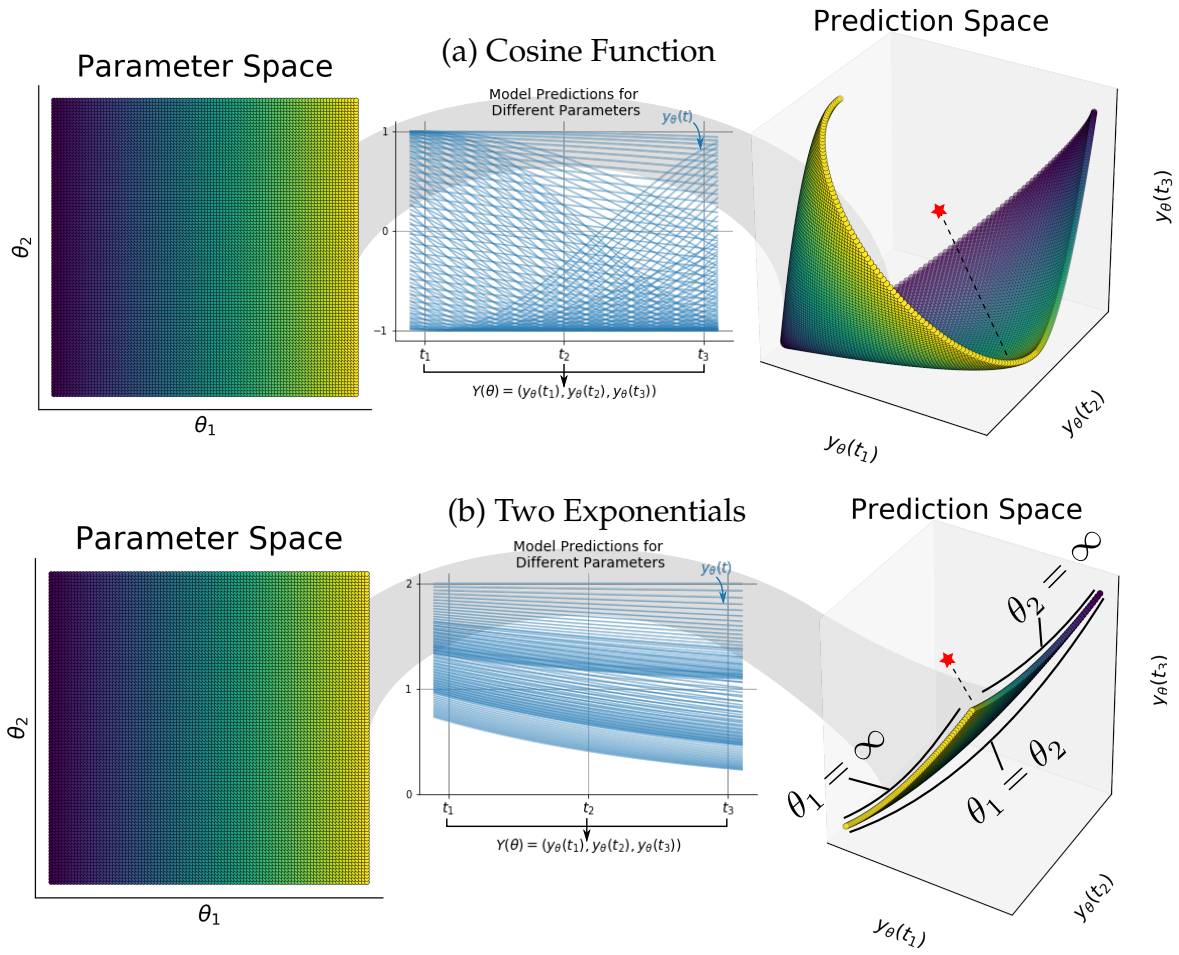


Figure 2.1: **Model manifolds** for the toy nonlinear models in (a) a simple cosine function (from Eq. (2.2)), and (b) the sum of two exponential curves with different decay rates (from Eq. (2.3)). Model manifolds are constructed as the set of all possible predictions for all possible model parameters and seen as a mapping from parameter space to prediction space. Fitting model parameters to data can be interpreted as projecting data onto the manifold (red star).

or when $\theta_1 = \theta_2$. We will explore the geometric interpretations of boundaries further in Section 2.4.2.

Fitting model parameters to data has a tangible interpretation in this framework; it can be seen as projecting data onto the model manifold. When fitting parameters to data, the model predictions are compared to experimental measurements. The measured data $(x_1, x_2, \dots)$ are compared to the corresponding model predictions $(y_\theta(t_1), y_\theta(t_2), \dots)$, and so form a vector in the embedding space of the model manifold. The parameters corresponding to the best fit are ones which minimize the cost function,

$$\chi^2 = -2 \log \mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x}), \tag{2.4}$$

where $\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x})$ represents a general likelihood function that the model predictions for parameters $\boldsymbol{\theta}$ fit the data $\boldsymbol{x}$. The cost function serves as a measure of distance, and can be directly translated to distances in the embedding space. Minimizing the cost is equivalent to minimizing the distance in this embedding space (up to a rescaling, discussed next), and so fitting the model parameters to data can be seen as finding the projection of the data onto the manifold. Figure 2.1 illustrates this with a red star being projected onto the manifold.

To see this effect explicitly, consider a canonical least-squares model. Here, the model predictions at different time points $y_\theta(t_i)$ have Gaussian noise $\xi \sim \mathcal{N}(0, \sigma_i^2)$. The cost function is expressed as[2]:

$$\chi^2 = -2 \log \mathcal{L}(\boldsymbol{x} \mid \boldsymbol{\theta}) = \sum_i \frac{(y_{\boldsymbol{\theta}}(t_i) - x_i)^2}{\sigma_i^2} + 2 \log \mathcal{Z} \tag{2.5}$$

[2]Note that we are expressing the cost in terms of $\mathcal{L}(\boldsymbol{x} \mid \boldsymbol{\theta})$ instead of $\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x})$, *i.e.* the likelihood of observing data $\boldsymbol{x}$ given parameters $\boldsymbol{\theta}$ rather than the likelihood that $\boldsymbol{\theta}$ fit for observed data $\boldsymbol{x}$. The two are related through Bayes' theorem, which states that $\mathcal{L}(\boldsymbol{\theta} \mid \boldsymbol{x})\mathcal{L}(\boldsymbol{x}) = \mathcal{L}(\boldsymbol{x} \mid \boldsymbol{\theta})\mathcal{L}(\boldsymbol{\theta})$. where $\mathcal{L}(\boldsymbol{x})$ and $\mathcal{L}(\boldsymbol{\theta})$ are priors on the data and the parameters respectively. In assuming uniform priors, the two likelihood functions are interchangeable. The impact of non-uniform priors is a rich and fruitful area of research, but will not be the focus of work presented here.

where $\mathcal{Z}$ is a normalization term, given as $\prod_i \frac{1}{\sqrt{2\pi\sigma_i^2}}$. If we drop the normalization term (which is a constant independent of chosen parameters) then we are left with a variance-scaled Euclidean distance between predictions $y_\theta(t_i)$ and data x_i. We use the cost as a basis for a distance function in this space, which we give as:

$$d^2(\boldsymbol{y}, \boldsymbol{x}) = \sum_i \frac{(y_i - x_i)^2}{\sigma_i^2}. \tag{2.6}$$

With a distance function, we can now find the metric on the manifold by considering the distance between parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta} + \delta\boldsymbol{\theta}$:

$$\begin{aligned} d^2(\boldsymbol{\theta}, \boldsymbol{\theta} + \delta\boldsymbol{\theta}) &= \sum_i \frac{(y_{\boldsymbol{\theta}}(t_i) - y_{\boldsymbol{\theta}+\delta\boldsymbol{\theta}}(t_i))^2}{\sigma_i^2} \qquad (2.7) \\ &= \sum_i \frac{(y_{\boldsymbol{\theta}}(t_i) - y_{\boldsymbol{\theta}}(t_i))^2}{\sigma_i^2} - 2\sum_\alpha \sum_i \frac{1}{\sigma_i^2}(y_{\boldsymbol{\theta}}(t_i) - y_{\boldsymbol{\theta}}(t_i)) \frac{\partial y_{\boldsymbol{\theta}}(t_i)}{\partial \theta^\alpha}\delta\theta^\alpha \\ &\quad - \sum_{\alpha,\beta}\sum_i \frac{1}{\sigma_i^2}\left[(y_{\boldsymbol{\theta}}(t_i) - y_{\boldsymbol{\theta}}(t_i)) \frac{\partial^2 y_{\boldsymbol{\theta}}(t_i)}{\partial\theta^\alpha \partial\theta^\beta} - \frac{\partial y_{\boldsymbol{\theta}}}{\partial\theta^\alpha}\frac{\partial y_{\boldsymbol{\theta}}}{\partial\theta^\beta}\right]\delta\theta^\alpha\delta\theta^\beta + O(\delta\boldsymbol{\theta}^3) \\ &= \sum_{\alpha,\beta} \underbrace{\sum_i \frac{1}{\sigma_i^2}\frac{\partial y_{\boldsymbol{\theta}}(t_i)}{\partial\theta^\alpha}\frac{\partial y_{\boldsymbol{\theta}}(t_i)}{\partial\theta^\beta}}_{g_{\alpha\beta}} \delta\theta^\alpha\delta\theta^\beta + O(\delta\boldsymbol{\theta}^3). \qquad (2.8) \end{aligned}$$

The metric $g_{\alpha\beta}$ can be expressed in terms of the Jacobian of the model. We can write this out explicitly as

$$g_{\alpha\beta} = \sum_i J_{i\alpha}J_{i\beta} \quad \text{where} \quad J_{i\alpha} = \frac{1}{\sigma_i}\frac{\partial y_{\boldsymbol{\theta}}(t_i)}{\partial\theta^\alpha}. \tag{2.9}$$

This metric is identical to a known object in statistics, known as the *Fisher information matrix*, which we discuss in the next section.

2.2 The Fisher Information Matrix

The Fisher Information Matrix (FIM) is a measure of the amount of information about parameters that can be extracted from experimental data [51]. It is fundamentally connected to the covariance matrix of model parameters, and so is widely used in optimal experimental design as a way of determining the expected accuracy of parameter estimates [87]. In other words, it determines the best parameter uncertainty you could hope for given a particular experiment. Given some probability distribution $\mathcal{L}(\boldsymbol{x} \mid \boldsymbol{\theta})$ that depends on parameters $\boldsymbol{\theta}$ for predictions $\boldsymbol{x}$, the FIM is expressed in two equivalent ways[3]:

$$(i) \quad \mathcal{I}_{\alpha\beta}(\boldsymbol{\theta}) = \sum_{\boldsymbol{x}} \frac{\partial \log \mathcal{L}(\boldsymbol{x} \mid \boldsymbol{\theta})}{\partial \theta^{\alpha}} \frac{\partial \log \mathcal{L}(\boldsymbol{x} \mid \boldsymbol{\theta})}{\partial \theta^{\beta}} \mathcal{L}(\boldsymbol{x} \mid \boldsymbol{\theta}); \tag{2.10}$$

$$(ii) \quad \mathcal{I}_{\alpha\beta}(\boldsymbol{\theta}) = -\sum_{\boldsymbol{x}} \frac{\partial^2 \log \mathcal{L}(\boldsymbol{x} \mid \boldsymbol{\theta})}{\partial \theta^{\alpha} \partial \theta^{\beta}} \mathcal{L}(\boldsymbol{x} \mid \boldsymbol{\theta}) \tag{2.11}$$

It is widely seen as the appropriate metric for probability distributions [4], because it sets a lower bound on the possible variance of parameter estimates for an unbiased prior through the Cramér–Rao bound [34, 126],

$$\mathrm{Cov}(\hat{\boldsymbol{\theta}}) \geq \mathcal{I}(\boldsymbol{\theta})^{-1}, \tag{2.12}$$

where $\hat{\boldsymbol{\theta}}$ emphasizes that this is the covariance matrix of an *estimator* of $\boldsymbol{\theta}$.

For the simple least-squares models discussed in Section 2.1 (where model predictions $y_{\boldsymbol{\theta}}(t)$ at points t_i with Gaussian noise of variance σ_i^2 are compared to data x_i), the likelihood functions are expressed as

$$\mathcal{L}(\boldsymbol{x} \mid \boldsymbol{\theta}) = \prod_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(y_{\boldsymbol{\theta}}(t_i) - x_i)^2}{2\sigma_i^2}\right). \tag{2.13}$$

[3]The equivalence between the two expressions of the FIM can be derived by taking the second derivative of the log-likelihood in the second term, and taking advantage of the normalized nature of $\mathcal{L}$. This is shown in Section A.2.

By plugging in the above expression for the likelihood function into Eq. (2.10), a tedious but straightforward calculation in Section A.1 shows that the FIM for least squares models is given by $\mathcal{I} = J^T J$ where $J_{i\alpha} = \frac{\partial y_{\boldsymbol{\theta}}(t_i)}{\sigma_i \partial \theta^\alpha}$ is the Jacobian of the model, thus matching the metric for model manifolds found in Eq. (2.9).

2.2.1 Connection to Hessians

The Hessian of the cost function at the best fit for least-squared models, using Eq. (2.5), is expressed as:

$$\frac{\partial^2 \chi^2(\boldsymbol{\theta})}{\partial \theta^\alpha \partial \theta^\beta} = 2 \sum_i \underbrace{\frac{(y_{\boldsymbol{\theta}}(t_i) - x_i)}{\sigma_i^2}}_{\text{very small near best fit}} \frac{\partial^2 y_{\boldsymbol{\theta}}(t_i)}{\partial \theta^\alpha \partial \theta^\beta} + \frac{\partial y_{\boldsymbol{\theta}}(t_i)}{\sigma_i \partial \theta^\alpha} \frac{\partial y_{\boldsymbol{\theta}}(t_i)}{\sigma_i \partial \theta^\beta} \approx 2 \underbrace{\sum_i J_{i\alpha} J_{i\beta}}_{\mathcal{I}_{\alpha\beta}}. \tag{2.14}$$

The Fisher information is therefore equivalent to the Hessian of the cost function at the best fit (up to a constant factor) for least-squares models.

2.2.2 Reparametrization

An important feature of the local metric is that it's heavily parameter dependent. Under reparametrization $\theta^\alpha \to \theta^{\tilde{\alpha}}$, the FIM is given by

$$\mathcal{I}_{\tilde{\alpha}\tilde{\beta}} = \frac{\partial \theta^\alpha}{\partial \theta^{\tilde{\alpha}}} \frac{\partial \theta^\beta}{\partial \theta^{\tilde{\beta}}} \mathcal{I}_{\alpha\beta}, \tag{2.15}$$

which for our purposes can be interpreted as a matrix multiplication, of the original FIM with the Jacobian of the coordinate transformation.

2.3 Divergences

How does one measure the 'distance' between two probability distributions? Ideally, if two distributions are indistinguishable they should have distance 0, and if they have no overlap they should be very far apart. A class of measures known as f-divergences [35, 106, 3] emerged in the mid-1900s to answer this question. Given two normalized probability distributions $P(x)$ and $Q(x)$, and convex function f with $f(1) = 0$, the divergence is defined as:

$$D_f(P \parallel Q) = \int dx f\left(\frac{P(x)}{Q(x)}\right) Q(x) \tag{2.16}$$

Importantly, the metric for all such divergences is proportional to the FIM [5].

The most commonly used divergence is the Kullback-Leibler (KL) divergence [89], where $f(x) = x \log x$, and written out as:

$$D_{KL}(P \parallel Q) = \int dx P(x) \log\left(\frac{P(X)}{Q(X)}\right) \tag{2.17}$$

The KL-divergence is a measure of the relative entropy between two distributions.

Another notable divergence is the Hellinger distance [70], where $f(x) = \left(\sqrt{x} - 1\right)^2$ and written out as:

$$D_H(P \parallel Q) = 1 - \int dx \sqrt{P(x)Q(X)} \tag{2.18}$$

There is a direct geometric interpretation of the Hellinger divergence, in particular with its relation to the unit sphere which we will explore in Section 4.2.

A less well-known yet very important distance function for probability distributions is the Bhattacharyya distance [19], expressed as

$$D_B(P \parallel Q) = -\log \int dx \sqrt{P(x)Q(x)} \tag{2.19}$$

While not technically an f-divergence, we will show in Chapter 4 that it is in fact the limit of an f-divergence. It measures the difference between two distributions in terms of their overlap. Importantly, the metric is also proportional to the Fisher Information, as shown in a tedious calculation in Section A.3.

2.4 Geometric Interpretations and Sloppy Models

A vital observation made through the study of model manifolds is that they are *hierarchical*: there is a structural hierarchy in both their *local* and *global* properties. Specifically, there are longer and shorter directions along the manifold. By finding the parameter combinations these relate to, one can extract *stiff* directions in parameter space (*i.e.* parameter combinations that drastically impact model predictions) and *sloppy* directions in parameter space (*i.e.* parameter combinations that have little impact on predictions) [146]. Figure 2.2 shows an illustration of how different directions in parameter space translate to different directions (of varying lengths) in prediction space.

Studying the model manifold yields fruitful information for a variety of reasons. From a pragmatic perspective, an understanding of local features of the manifold such as curvature can lead to more efficient data fitting methods [148] due to the geometric connection with data (discussed in Section 2.1, where fitting model parameters to data can be seen as projecting onto the model manifold).

From a theory perspective, understanding the dominant components of the manifold can yield a better understanding of *emergence*, *i.e.* how microscopic features of the model yield simple macroscopic behaviour [99].

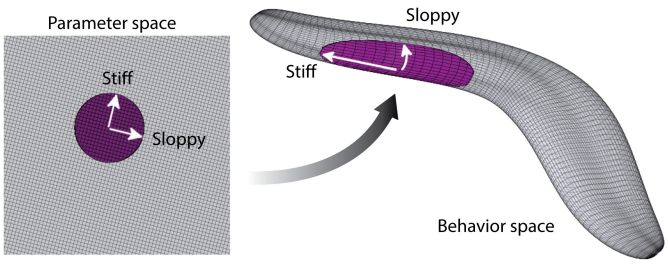


Figure 2.2: **The prediction/behavior space of complex systems** is primarily controlled by a few parameter combinations, even when the model has numerous parameters. The hierarchy of parameter importance directly translates to a geometric hierarchy in the local properties (through geometric decay in the eigenvalues of the metric) and global features (through a geometric decay in the hierarchy of manifold widths). Figure taken from [134].

2.4.1 Local Properties

Local features on the model manifold are determined by the metric, the FIM. The structural hierarchy at the local levels is revealed through eigenvalues of the metric. They are geometrically distributed (evenly distributed on a log scale). Figure 2.3 shows the eigenvalue spread of the FIM for different nonlinear models. Larger eigenvalues correspond to *stiff* parameter combinations (parameter combinations that heavily affect model predictions) whereas the smaller eigenvalues correspond to *sloppy* parameter combination that can vary wildly with minimal impact on model predictions.

While the sloppy spectra shown in Fig. 2.3 are the result of the FIM evaluated at a single point, the spread appears to be a feature at every (or at least, nearly every) point on the model manifold. Figure 2.4 shows the normalized distribution of the eigenvalues of the metric computed at every point on the model manifold for the three different models considered in Chapter 3. The eigenvalue spectra of the FIM for sloppy models, with speculations regarding a connection

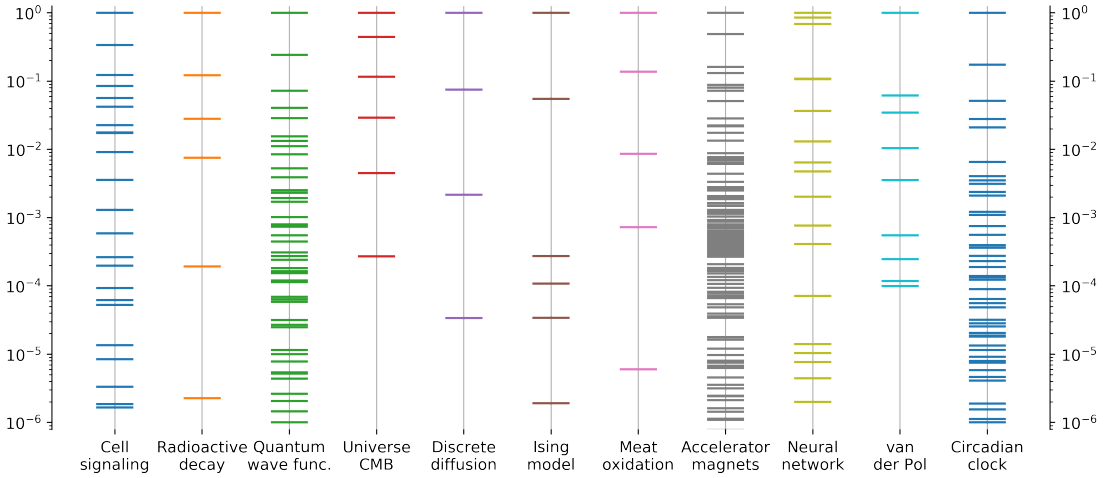


Figure 2.3: **Eigenvalues of the local metric** for many disparate, nonlinear models (rescaled by the largest natural eigenvalue for each model) that span many orders of magnitude. Note the enormous range in the vertical axis. Cell signaling data from [22], radioactive decay and neural network are taken from [148], quantum wavefunction are taken from [157], diffusion model and Ising model are taken from [99], meat oxidation is from [142], CMB data are explained in Section D.4, accelerator model taken from [61], van der Pol oscillator taken from [28] and circadian clock model from [37].

to random matrix theory, is explored in Appendix B.

2.4.2 Global Properties

Global properties of the model manifold are determined by the size of the manifold in the embedding space, which reflect the full range of predictions allowed by the model. There are interesting topological considerations. For instance, the boundaries represent reduced-model approximations [149], *i.e.* models with reduced complexity that still retain much predictive power.

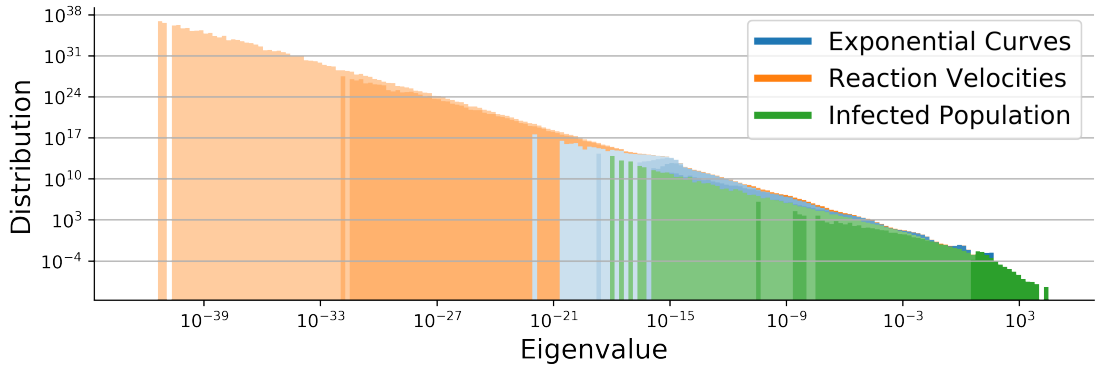


Figure 2.4: **Normalized, stacked histogram of ordered eigenvalues** of the metric computed at every point on the model manifold for the thee nonlinear models considered in Chapter 3. Note the enormous ranges, both in the horizontal and vertical axes (both are log-scaled). Color reflects the model (blue is exponential curves discussed in Section 3.2.1, orange is reaction velocity discussed in Section 3.2.2, green is the infected fraction of a population in an SIR epidemiology model discussed in Section 3.2.3). Opacity reflect order of eigenvalues in the local metric (darkest color is largest eigenvalue in metric). Eigenvalues follow a geometric decay, indicative of sloppy spectra. While the three models appear quite different, their eigenvalue spectra follow very similar geometric decays, a feature explored briefly in Appendix B.

Model manifolds typically form striking *hyperribbons* [147], so-called because, like ribbons, successive widths follow a geometric decay. They are much longer than they are wide, much wider than they are thick, etc., yielding effective low-dimensional representations. Figure 2.5 shows the hierarchy in manifold lengths for the models considered in this thesis.

Importantly, because directions along the model manifold correspond to specific parameter combinations, there is a direct connection between the hyperribbon nature of model manifolds and the structural hierarchy of model parameters. In other words, understanding why model manifolds form hyperribbons leads to an understanding of why structural hierarchies in parameter

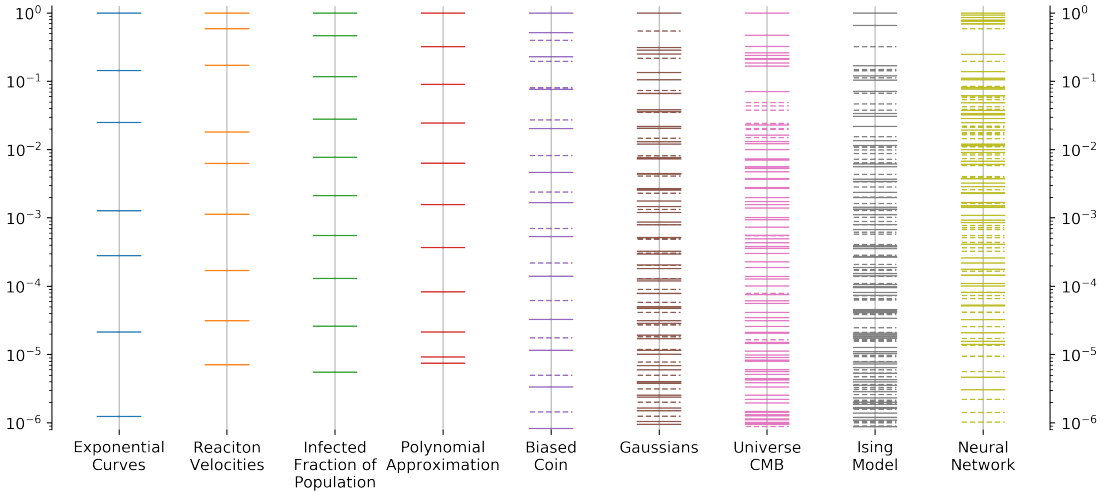


Figure 2.5: **Manifold lengths** for many disparate, nonlinear models (rescaled by the largest length for each model) illustrating the hyperribbon structure that characterizes model manifolds. Note the enormous range in vertical axis. Exponential curves are discussed in Section 3.2.1, reaction velocities of an enzyme-catalysed reaction are discussed in Section 3.2.2, and infected fraction of a population from an SIR epidemiology model are discussed in Section 3.2.3. Polynomial approximations reflects the numerically established bounds from Chapter 3. Biased coin is discussed in Section 4.6.1, Gaussians are discussed in Section 4.6.2, ΛCDM predictions of the CMB are discussed in Section 4.6.5, and the Ising model is discussed in Section 4.6.3. For probabilistic models, imaginary lengths (*i.e.* negative squared distances) are reflected by dashed lines and are discussed in Chapter 4.

importance exists. In Chapter 3, we explore this geometric structure in greater detail and use polynomial approximation theory to better understand where the hyperribbon structure comes from.

2.5 Manifold Learning

Visualizing high-dimensional data is a cornerstone of machine learning, modeling, big data, and data mining. These fields require learning faithful and interpretable low-dimensional representations of high-dimensional data, and, almost as critically, producing visualizations which allow interpretation and evaluation of what was learned [41, 97, 91, 163]. Unsupervised learning, which infers features from data without manually curated data or specific problem definitions [107], is especially important for high-dimensional, big data applications in which specific models are unknown or impractical. For high dimensions, the relative distances between features become small and most points are orthogonal to one another [88]. A trade-off between preserving local and global structure must often be made when inferring a low-dimensional representation. Generally, there is seen to be two kinds of manifold learning techniques: (1) linear methods, which preserve global features, and (2) nonlinear methods, which capture local features.

Classic manifold learning techniques include linear methods such as Principal Component Analysis (PCA) [77] and multidimensional scaling (MDS) [144], which preserve global structure but at the cost of obscuring local features. Such methods aim to project the data into an orthogonal coordinate system, determined by the eigenbasis of the covariance matrix of points.

Existing nonlinear manifold learning techniques, such as t-distributed Stochastic Network Embedding (t-SNE) [98] and diffusion maps [31], preserve the local structure while only maintaining some qualitative global patterns such as large clusters. The Uniform Manifold Approximation (UMAP) [102] better

preserves topological structures in data, a global property. We will use t-SNE in Chapter 5 to qualitatively visualize the results of a clustering method applied to behaviour data.

In Chapter 4, we develop a new nonlinear manifold learning technique which achieves a compromise between global vs. local trade-off, by embedding manifolds into a Minkowski-like space that allows them to be 'unwound' in a natural way. We call this new method the Intensive Principal Component Analysis, or InPCA.

CHAPTER 3

BOUNDING MODEL PREDICTIONS

In this chapter[1], we quantify the underlying smoothness of a model and combine it with polynomial approximation theory to set geometric constraints on the space of allowed model predictions. We use this to rigorously quantify the hyperribbon structure of model manifolds (see Section 2.4.2 for description of hyperribbons) for regular, least-squares models.

3.1 Polynomial Approximations

Consider a nonlinear model that depends continuously on K input parameters $\theta = (\theta^1, \ldots, \theta^K)$ to generate predictions $y_\theta(t)$. If we consider the model predictions at N fixed points, $\{t_0, \ldots, t_{N-1}\}$, then our predictions for parameters θ form an N-dimensional vector $Y(\theta) = (y_\theta(t_0), \ldots, y_\theta(t_{N-1}) = (Y_0, \ldots, Y_{N-1})$. We use $\mathcal{Y}$ to represent the model manifold, defined as the space of all possible predictions for all possible parameter combinations (so all allowed $Y(\theta)$). Specifically, model manifold $\mathcal{Y}$ is a K-dimensional surface embedded in an N-dimensional prediction space.

To bound the model manifold $\mathcal{Y}$ and study its geometry, we consider polynomial approximations of model y_θ. Without loss of generality, we shift and rescale

[1]Work in this chapter was done in collaboration with Heather Wilber, Alex Townsend and James Sethna. The majority of this chapter has been published in PRL [122], and a preprint is available on the arXiv [121]. We thank Mark Transtrum for suggestions related to selecting models used in this chapter, John Guckenheimer for suggesting that there could be a connection between the third and fourth author's research areas, and Peter Lepage for his expertise and insight into the connection with lattice QCD. KNQ was supported by a fellowship from the Natural Sciences and Engineering Research Council of Canada (NSERC), and JPS and KNQ were supported by the National Science Foundation (NSF) through grant DMR-1719490. AT was supported by NSF grant no. DMS-1818757, and HW was supported by NSF grant no. DGE-1650441.

the points so that $\{t_k\}_{k=0}^{N-1} \subset [-1, 1]$. Let $\{\phi_j\}_{j=0}^{\infty}$ be a complete polynomial basis, and suppose that model $y_\theta(t)$ is decomposed into this basis: $y_\theta(t) = \sum_{j=0}^{\infty} b_j(\theta)\phi_j(t)$. We can consider the predictions of the truncated series at fixed points, just as we considered the model predictions at fixed points to generate the original model manifold $\mathcal{Y}$. Let $p_{N-1}(t;\theta)$ be the *truncated* series representing the polynomial approximation to model $y_\theta(t)$. Note that the truncation is set by the number of sampled points, N. We can view the coefficients $(b_0(\theta), \ldots, b_{N-1}(\theta))$ as a set of N parameters. Now, let $P(\theta) = (p_{N-1}(t_0), \ldots, p_{N-1}(t_{N-1})) = (P_0, \ldots, P_{N-1})$ define the polynomial manifold $\mathcal{P}$. Thus, we have model manifold $\mathcal{Y}$ and a polynomial manifold $\mathcal{P}$. Both of these manifolds have associated parameters, given respectively by the original model parameters θ and polynomial coefficients $(b_0, \ldots, b_{N-1})$. Table 3.1 illustrates the relation between $\mathcal{Y}$ and $\mathcal{P}$.

Table 3.1: **Manifold relationships** between models and their polynomial approximations at fixed points $\{t_0, \ldots, t_{N-1}\}$ for polynomial basis $\{\phi_j\}_{j=0}^{\infty}$.

	Physical Model	**Polynomial Approx.**
	y_θ	p_{N-1}
Series	$\sum_{j=0}^{\infty} b_j(\theta)\phi_j(t)$	$\sum_{j=0}^{N-1} b_j\phi_j(t)$
Parameters	θ	$(b_0, \ldots, b_{N-1})$
Predictions	$Y(\theta) = (y_\theta(t_0), ..., y_\theta(t_{N-1}))$	$P(\theta) = (p_{N-1}(t_0), ..., p_{N-1}(t_{N-1}))$
Manifold	$\mathcal{Y}$	$\mathcal{P}$
Hyperellipsoid	H_Y	H_P
Widths	$2r\sigma_j(X) + 2\|y - p_{N-1}\|_\infty$	$2r\sigma_j(X)$

By definition, $P(\theta) = X\mathbf{b}$, where $X_{ij} = \phi_{j-1}(t_{i-1})$ and $\mathbf{b} = (b_0, \ldots, b_{N-1})^T$. Here, X forms a linear map from the space of polynomial coefficients to the space of possible predictions, and is determined by the chosen polynomial basis and fixed

points t_i. The singular values of X can be used to understand the hyperribbon structure of the polynomial manifold $\mathcal{P}$. Suppose, for example, that $\|\mathbf{b}\|_2 < r$, so that the coefficient space is bounded in S, an n-sphere of radius r. The action of X on S distorts it into a hyperellipsoid H_P. If $\ell_j(H_P)$ is the diameter of the jth cross-section of hyperellipsoid H_P, then

$$\ell_j(H_P) = 2r\sigma_j(X), \tag{3.1}$$

where $\sigma_j(X)$ are the ordered singular values of X. When X has rapidly decaying singular values, H_P has a hyperribbon structure because there is a strict hierarchy in successive widths. Accounting for the polynomial approximation error $\|y_\theta - p_{N-1}\|_\infty$, where $\|\cdot\|_\infty$ is the L^∞ norm on $[-1, 1]$, we can define a hyperellipsoid H_Y that must enclose model manifold $\mathcal{Y}$, where the cross-sectional widths are given by

$$\ell_j(H_Y) = 2r\sigma_j(X) + 2\|y - p_{N-1}\|_\infty. \tag{3.2}$$

In this way, we find that *any* model manifold $\mathcal{Y}$ is bounded within a hyperribbon whenever $\sigma_j(X)$ decays geometrically and $\|y - p_{N-1}\|_\infty$ is small enough. A fundamental question is whether it matters which polynomial basis or which set of time points are chosen to define H_P and H_Y. The hyperribbon structure of $\mathcal{Y}$, of course, does not depend on our representation of y_θ, but rather on intrinsic properties of the model, such as its smoothness. For example, if for every $t_0 \in [-1, 1]$, the Taylor expansion of y_θ at t_0 has a large enough radius of convergence, any sequence of polynomial interpolants with N distinct interpolating points converges to y_θ at a geometric rate with N [154]. This fact underpins the qualitative observation in [147, 146] that certain analytic models have manifolds bounded within hyperribbons. Here we make that observation rigorous.

We consider two such choices. First, we choose our basis functions $\{\phi_j\}_{j=0}^{\infty}$ as

the Chebyshev polynomials. Truncated Chebyshev expansions converge to y_θ at an asymptotically optimal rate for a polynomial approximation [154]. As we show below, this rate controls the magnitude of $\sigma_j(X)$ in Eq. (3.2), and can be used to explicitly bound the cross-sectional widths of H_Y. Our bounds deliver an outright description of a hyperribbon that must contain $\mathcal{Y}$.

We also analyze the case where $\{\phi_j\}_{j=0}^{\infty}$ are the monomials and p_{N-1} is the truncated Taylor series expansion of y_θ. In this case, we observe that the numerical computation of $\sigma_j(X)$ results in excellent practical and universal bounds on the prediction space for large classes of models.

3.1.1 Chebyshev Expansions

Suppose that the model y_θ has a convergent Chebyshev expansion, so that it is given by $y_\theta(t) = \sum_{j=0}^{\infty} c_j(\theta) T_j(t)$, where $T_j(t) = \cos(j \arccos t)$ is the degree j Chebyshev polynomial [154, Ch. 3]. We can approximate y_θ with a degree $\leq N-1$ polynomial by truncating the Chebyshev series after N terms:

$$p_{N-1}(t;\theta) = \sum_{j=0}^{N-1} c_j(\theta) T_j(t). \tag{3.3}$$

Truncated Chebyshev expansions have near-best global approximation properties. The error $\|y_\theta - p_{N-1}\|_\infty$ is within a $\log N$ factor of $\|y_\theta - p_{N-1}^{best}\|_\infty$ [154, Ch. 16], where p_{N-1}^{best} is the best polynomial approximant to y_θ of degree $\leq N-1$. We cannot directly use p_{N-1}^{best} in our arguments because bounds on $\|y_\theta - p_{N-1}^{best}\|_\infty$ are only known in an asymptotic sense. Fortunately, explicit bounds on $\|y_\theta - p_{N-1}\|_\infty$ are known when y_θ is sufficiently smooth.

We first consider the case where y_θ is analytic in an open neighborhood of

$[-1, 1]$. Such a region contains a *Bernstein ellipse* E_ρ in the complex plane, defined as the image of the circle $|z| = \rho$ under the Joukowsky mapping $(z + z^{-1})/2$. It has foci at ± 1, and the lengths of its semi-major and semi-minor axes sum to ρ. The polynomial in Eq. (3.3) converges to y_θ as $N \to \infty$ at a rate determined by ρ:

Theorem 1 *Let $M > 0$ and $\rho > 1$ be constants and suppose that $y_\theta(t)$, $t \in [-1, 1]$, is analytically continuable to the region enclosed by the Bernstein ellipse E_ρ, with $|y_\theta| \leq M$ in E_ρ, uniformly in θ. Let $p_{N-1}(t; \theta)$ be as in Eq. (3.3). Then,*

$$(i) \ \|y_\theta - p_{N-1}\|_\infty \leq \frac{2M\rho^{-N+1}}{\rho - 1}, \tag{3.4}$$

$$(ii) \ |c_0| \leq M, \ |c_j(\theta)| \leq 2M\rho^{-j}, \qquad j \geq 1. \tag{3.5}$$

Proof 1 *For a proof, see Theorem 8.2 in [154].*

To exploit the decay of the coefficients in Eq. (3.5), we define modified coefficients $\tilde{c}_j = \rho^j c_j$. We then have that polynomial predictions $P(\theta) = X\tilde{\mathbf{c}}$, where $X = JD$, $J_{ij} = T_{j-1}(t_{i-1})$, and D is diagonal with entries $D_{jj} = \rho^{-(j-1)}$. By Eq. (3.5), we have that $\|\tilde{\mathbf{c}}\|_2 < 4M\sqrt{4N - 3}$. This implies that the polynomial manifold $\mathcal{P}$ is bound in a hyperellipsoid H_P. By Eq. (3.1), we have that $\ell_j(H_P) = 8M\sqrt{4N - 3}\sigma_j(X)$. To bound $\sigma_j(X)$ explicitly, we first prove a conjecture proposed in [157]:

Theorem 2 *Let $S \in \mathbb{R}^{N\times N}$ be symmetric and positive definite. Let $E \in \mathbb{R}^{N\times N}$ be diagonal with $E_{ii} = \epsilon^{i-1}$ and $0 < \epsilon < 1$. If $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$ are the ordered eigenvalues of ESE, then $\lambda_{m+1} = O(\epsilon^{2m})$. Specifically,*

$$\lambda_{m+1} \leq \frac{\epsilon^{2m}}{1 - \epsilon^2} \max_{1\leq j,k\leq N} \left|S_{jk}\right|, \qquad 1 \leq m \leq N - 1. \tag{3.6}$$

Proof 2 [2] *Consider the rank m matrix*

$$S_m = S(:, 1:m)S(1:m, 1:m)^{-1}S(1:m, :), \tag{3.7}$$

where $1 \leq m \leq N-1$, *and the notation* $M(:, 1:m)$ *denotes the submatrix of* M *consisting of its first* m *columns. Clearly,* S_m *is well-defined because* $S(1:m, 1:m)$ *is a principal minor of a positive definite matrix and is therefore invertible. Moreover, it can be verified that* $(S - S_m)_{jk} = 0$ *for* $1 \leq j, k \leq m$.

Since ESE *is positive definite and* $\operatorname{rank}(S_m) = m$, *we know that*

$$\lambda_{m+1} \leq \|E(S - S_m)E\|_2,$$

where $\|\cdot\|_2$ *denotes the spectral matrix norm [56, Ch. 2]. Using* $\|\cdot\|_F$ *to denote the Frobenius norm, we have*

$$\begin{aligned}
\lambda_{m+1}^2 &\leq \|E(S - S_m)E\|_2^2 \leq \|E(S - S_m)E\|_F^2 && (3.8)\\
&= \sum_{j=m+1}^{N} \sum_{k=m+1}^{N} \epsilon^{2(j-1)+2(k-1)} \left|S_{jk} - (S_m)_{jk}\right|^2 && (3.9)\\
&\leq \frac{\epsilon^{4m}}{(1-\epsilon^2)^2} \max_{1\leq j,k\leq N} \left|S_{jk} - (S_m)_{jk}\right|^2 && (3.10)\\
&\leq \frac{\epsilon^{4m}}{(1-\epsilon^2)^2} \max_{1\leq j,k\leq N} \left|S_{jk}\right|^2, && (3.11)
\end{aligned}$$

where the last inequality comes from the fact that the block

$$S(m+1:N, m+1:N) - S_m(m+1:N, m+1:N)$$

is the Schur complement of $S(1:m, 1:m)$ *in* S *[56].*

Applying Theorem 2 to $X^TX = DJ^TJD$, we have that for $j > 1$,

$$\sigma_j(X) \leq \frac{\sqrt{N}\rho^{-j+2}}{\sqrt{\rho^2 - 1}}, \tag{3.12}$$

[2] Previous proofs with weaker bounds were provided through private communications with Ari Turner and Yaming Yu. Current proof provided by co-authors Heather Wilber and Alex Townsend.

where we have used the fact that $|T_k(t)| \leq 1$ for $k \geq 0$ and $-1 \leq t \leq 1$. It follows from Equations (3.2) and (3.4) that predictions for $y_\theta(t)$ are bounded by a hyperellipsoid H_Y, with

$$\ell_j(H_Y) \leq \frac{2M\sqrt{4N^2 - 3N}\rho^{-j+2}}{\sqrt{\rho^2 - 1}} + \frac{4M\rho^{-N+1}}{\rho - 1}, \tag{3.13}$$

for $2 \leq j \leq N$, *i.e.*,

$$\ell_j(H_Y) = O(\rho^{-j} + \rho^{-N}). \tag{3.14}$$

These bounds indicate that the hyperribbon structure of H_Y is controlled by ρ, a parameter characterizing the analyticity of the model. As ρ becomes larger (*i.e.* as the model becomes "smoother"), bounds on the widths of the successive cross-sections of H_Y must decay more rapidly. In principle, H_Y becomes successively thinner and more ribbon-like.

When y_θ is not analytic on an open neighborhood of $[-1, 1]$, the decay rate of $\sigma_j(JD)$ is instead controlled by the smoothness of y_θ on $[-1, 1]$. More discussion of bounding non-analytic models is provided in Appendix A.4. Furthermore, when we consider models with two experimental conditions (for instance, time and temperature) these bounds can be extended to the two-dimensional case. We provide more discussion two-dimensional cases in Section 3.3.

3.1.2 Taylor Expansions

For Taylor expansions, the degree $N - 1$ truncated polynomial of y_θ is $p_{N-1}(t) = \sum_{k=0}^{N-1} a_k(\theta)(t - t_0)^j$, where $a_k(\theta) = y_\theta^{(k)}(t_0)/k!$. One could describe the smoothness of the model by finding some $C > 0$ and $R > 1$ such that

$$\left| \frac{1}{k!} \frac{d^k y_\theta(t)}{dt^k} \right| < \frac{C}{R^k} \tag{3.15}$$

for all $k \geq 1$.

We relax this definition, and instead describe the analyticity of y_θ using the following condition: for all $N \geq 1$,

$$\sum_{k=0}^{N-1} \left(\frac{R^k}{k!} \frac{d^k y_\theta(t)}{dt^k} \right)^2 < C^2 N, \tag{3.16}$$

where $C > 0$, $R > 1$ are constants in θ. A straightforward but tedious calculation outlined in Appendix A.5 shows that the lengths of the resulting hyperellipsoid are given by

$$\ell_j(H_P) \leq \frac{2CN}{\sqrt{R^2 - 1}} R^{-j+2}. \tag{3.17}$$

3.2 Examples

To apply our results, we selected three models from quite disparate fields (physics, chemistry, biology). This was done deliberately, to illustrate the universal nature of our results. In all three cases, the context for model construction is different, and yet the underlying smoothness of each can be used to relate them to a single, *universal* bound.

The model manifolds for these three models are shown in Fig. 3.1. They are all contained within the *same* hyperellipsoid, as shown in Fig. 3.1(b), and so share the *same* universal bound. The hyperribbon structure of the manifolds is accurately captured by the numerical bound from Eq. (3.17), and the decay in successive manifold widths are clearly captured by the Chebyshev rate from Eq. (3.13). These three models were derived in very different contexts and exhibit what would appear to be fundamentally different properties, yet they all share a fundamental property: in all cases, there is a structural hierarchy in their

model manifolds as determined by a *universal* bound. Because of the geometric decay in successive manifold widths, low-dimensional representations (as determined by the longest directions) capture the large variance in model predictions. This is because they are all part of the same universality class, that of sloppy models.

Appendix D.1 discusses in greater detail how the models are sampled, and Fig. D.1 shows the parameter ranges considered for each of the three models.

3.2.1 Physics: Exponential Curves

The first model we consider is that of exponential curves, such as for radioactive decay [146, 148] and calculating correlators in lattice QCD [92, 76]. Here, we set

$$y_\theta(t) = \sum_{\alpha=0}^{10} A_\alpha \exp\left(-\lambda_\alpha t\right), \tag{3.18}$$

where model parameters are the amplitudes A_α and decay rates λ_α, and t represents time.

To extend this model to two experimental conditions (discussed further in Section 3.3), we consider temperature dependent decay rates,

$$\lambda_\alpha \quad \rightarrow \quad \lambda_\alpha \exp\left(-E_\alpha s\right), \tag{3.19}$$

$$y(t) \quad \rightarrow \quad y(t,s) = \sum_\alpha A_\alpha \exp\left(-\lambda_\alpha \exp(-E_\alpha s) t\right), \tag{3.20}$$

where $s = 1/T$ is inverse temperature and E_α represents activation energes.

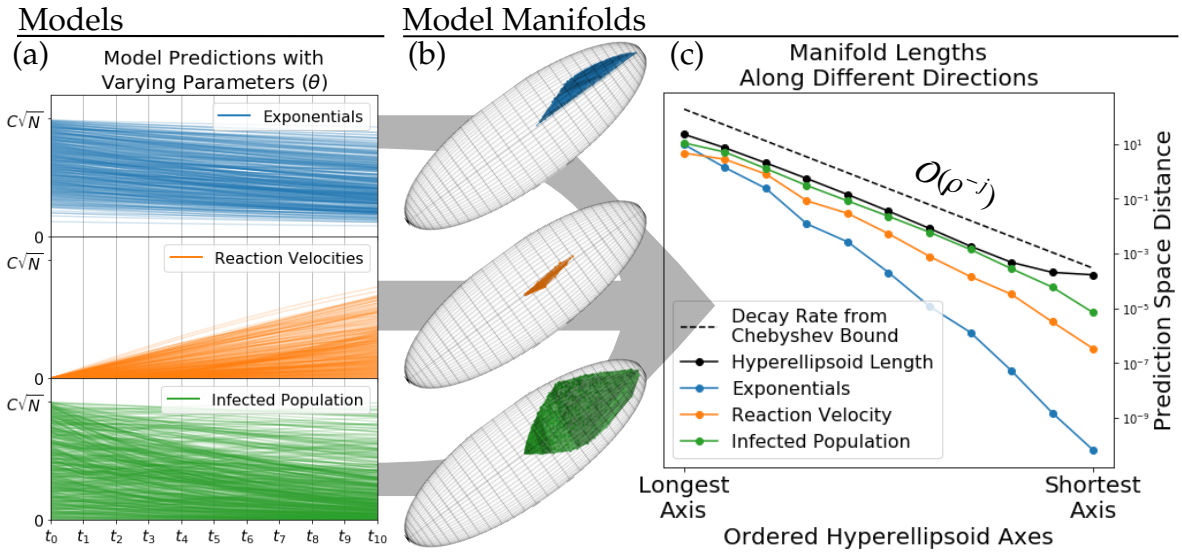


Figure 3.1: **Model manifold** of three disparate models: (1) exponential curves described in Section 3.2.1, (2) reaction velocities of an enzyme-catalysed reaction described in Section 3.2.2, and (3) the infected population in an SIR model described in Section 3.2.3. The models are evaluated at 11 equally spaced points on $[0, 1]$, and obey the smoothness condition in Eq. (3.16), with $C = 1$ and $R = 2$. (a) An illustration of each model, where each line represents the respective model predictions with a different set of parameters. (b) The *model manifolds* are all bounded by the same hyperellipsoid, and so the two axes represent the first and second longest hyperellipsoid axes. Note that, in all three models, only values greater than 0 are physically significant. This constraint manifests itself geometrically through their *location* in the hyperellipsoid. (c) The lengths of each model manifold along the eleven axes of the hyperellipsoid H_P in Eq. (3.17). Black points are the numerically computed lengths of H_P, given by $2C\sqrt{N}\sigma_j(VD)$ in Eq. (3.17), and include the error term from Eq. (3.2) (note the kink at the second to last point), forming an upper bound on possible lengths of the manifolds. The explicit decay rate of the Chebyshev-based bound (black dotted line) is based on the fact that models obeying Eq. (3.16) are analytic in the ellipse $E_\rho(\zeta)$. (Here, $\rho(\zeta) \approx 3.81$.) It captures the decay rate of $\sigma_j(VD)$ for $j < 11$, and closely follows the true decay rate in the successive widths of the various manifolds.

3.2.2 Chemistry: Reaction Velocities

The second model we consider is of an enzyme-catalysed chemical reaction [11, 85]. This model can be expressed as

$$y_\theta(t) = \frac{\theta_1 t^2 + \theta_2 t}{t^2 + \theta_3 t + \theta_4}, \tag{3.21}$$

where t represents the substrate concentration [146]. This model stands in for steady-state behavior of complex chemical reaction networks in engineering and ecology [103].

To extend this model to two experimental conditions (discussed further in Section 3.3), we consider temperature dependent parameters,

$$\theta_\alpha \to \theta_\alpha \exp\left(-E_\alpha s\right), \tag{3.22}$$

where again $s = 1/T$ is inverse temperature, and E_α represent activation energies.

3.2.3 Biology: Infected Fraction of a Population

The infected fraction of a population in the SIR epidemiology model [71] predicts the size of a population that is susceptible to infection ($S(t)$), infected ($I(t)$), and recovered from infection ($R(t)$). These are expressed through three coupled differential equations:

$$\dot{S} = -\beta \frac{IS}{N_{tot}} \tag{3.23}$$

$$\dot{I} = \beta \frac{IS}{N_{tot}} - \gamma I \tag{3.24}$$

$$\dot{R} = \gamma I, \tag{3.25}$$

where model parameters β and γ represent the rates of infection and recovery, and additional parameters include the total population N_{tot}, and initial infected and recovered population. At all times, $S(t)$, $I(t)$ and $R(t)$ sum to N_{tot}, and we set $y_\theta(t) = I(t)$. This model serves to represent classes of models involving numerical ODEs, which occur in power systems [151, 150] and systems biology [23, 22].

To extend this model to two experimental conditions, we introduce infection and recovery rates that vary continuously with an infection parameter s by introducing

$$\beta \rightarrow \beta \exp\left(-E_\beta s\right), \tag{3.26}$$

$$\gamma \rightarrow \gamma \exp\left(-E_\gamma s\right). \tag{3.27}$$

For our purposes, this model has effectively three parameters: the initial fraction of a population which is infected, the rate of recovery, and the rate of infection. We evaluate the model in terms of its reduced parameters (*i.e.* dividing all equations by N_{tot}), and consider situations with only initial infection (*i.e.* there is no initial recovered population in our models).

3.3 Two-Dimensional Extension of Model Predictions

Here, we extend the three models used Section 3.2 to the 2D setting. We do this by adding an extra experimental condition, denoted by s, to each model. In Fig. 3.2, we construct the model manifolds for all three. Just as before, the model manifold is bounded by a hyperellipsoid H_Y with a hierarchy of widths that form a hyperribbon structure.

Fig. 3.2 shows the model manifolds of all three example models, illustrating

their hyperribbon structures. To generate these figures, we consider models that obey an analyticity constraint analogous to Eq. (3.17). Specifically, we assume that for all $0 \leq j+k \leq N-1$, the following condition holds uniformly in θ for a given $2D$ model $y_\theta(t,s)$:

$$\sum_{j+k\leq N-1} \left(\frac{R^{j+k}}{j!k!} \frac{d^{j+k} y_{\boldsymbol{\theta}}(t,s)}{dt^j ds^k} \right)^2 < C^2 n. \tag{3.28}$$

where $R > 1, C > 0$ are constants, and $n = N(N+1)/2$. Under this constraint, it makes sense to bound the prediction space using truncated Taylor expansions of total degree $\leq N-1$ for small to moderate N (see the discussion in Appendix C). This choice results in an $n \times n$ linear system of the form $y_\theta(\mathbf{t},\mathbf{s}) \approx X\tilde{\mathbf{a}}$, where X is a column-scaled 2D Vandermonde matrix, and $\|\tilde{\mathbf{a}}\|_2 < C\sqrt{n}$. The structure of X can be exploited to bound its singular values explicitly [145]. Alternatively, one can apply the 2D analogue to Theorem 2 to find explicit bounds in terms of R. In Fig. 3.2, we simply use the relation $\ell_j(H_Y) = \ell_j(H_P) + 2\|y_\theta - p_{N-1}\|_\infty$, and compute $\ell_j(H_P) = 2r\sigma_j(X)$ numerically.

We compare this with the Chebyshev-based bound (established in the following section),

$$\ell_j(H_P) \leq \sqrt{N} \frac{3\sqrt{C_2}}{2} n \rho^{-\left\lfloor \sqrt{8(j-1)+1}/2 - 1/2 \right\rfloor}, \tag{3.29}$$

where ρ is a characteristic length related to the analyticity of the model, $C_2 = (1+\rho^{-2}+\rho^{-4})/(1-\rho^{-2})^3$, and $\lfloor \cdot \rfloor$ represents the floor function. This bound captures the subgeometric decay rate of the model manifold lengths for all three examples, illustrated through the dashed line in Fig. 3.2.

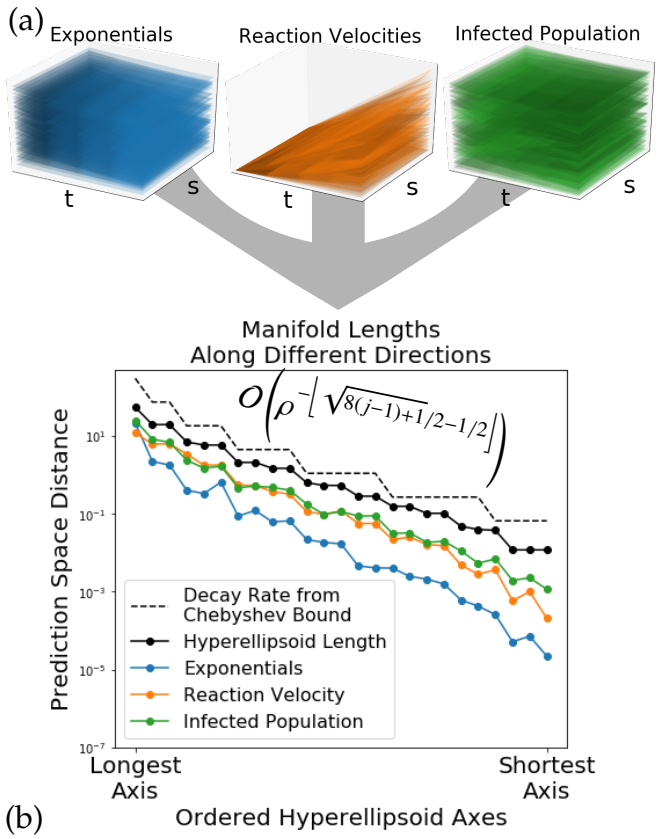


Figure 3.2: **Model manifold** of three models with two experimental conditions: (1) exponential decay with temperature dependent decay rates, (2) reaction velocities of an enzyme-catalysed reaction with temperature dependent reaction rates, and (3) the infected population in an SIR model with infection and recovery rates that vary with parameter s. (a) The models are evaluated at 25 equally spaced points $(t_i, s_i) \in [0, 1]^2$ (shifted and rescaled from the interval $[-1, 1]^2$) with different model parameters. All models obey the analyticity condition in Eq. (3.28) with C=1 and $R = 2$. (b) The explicit lengths of the three models are shown along the twenty-five axes of the hyperellipsoid H_P. The upper bounds on the possible lengths (black dots) are given by $\ell_j(H_P) = 2C\sqrt{n}\sigma_j(X)$, where X is described in Section 3.3. They exhibit subgeometric decay, with a rate that is captured by the bound in Eq. (3.33) (dashed line) with $\rho \approx 4.1$. The hierarchy of widths coming from the explicit bounds suggests that the manifolds are hyperribbons.

3.3.1 Bounds on the 2D Extension

We can again use polynomial approximation to constrain the geometry of the resulting model manifold $\mathcal{Y}$ for 2D extensions of models. In this case, we assume without loss of generality that $(t, s) \in [-1, 1]^2$, and we assume y_θ can be expressed as a 2D Chebyshev expansion: $y_\theta(t, s) = \sum_{j=0}^{\infty} \sum_{k=0}^{\infty} c_{jk}(\theta) T_{jk}(t, s)$, where $T_{jk}(t, s) = T_j(t) T_k(s)$. The following 2D polynomial of total degree $N-1$ approximates y_θ:

$$p_{N-1}(t, s; \theta) = \sum_{0 \le j+k \le N-1} c_{jk}(\theta) T_{jk}(t, s). \tag{3.30}$$

Let $\rho > 1$ and $M > 0$ be constants. For all fixed choices of $s = s^*$, suppose that the 1D function of t, $y_\theta(t, s^*)$, is analytic in t and bounded $\le M$ uniformly with respect to both s and θ, and that an analogous condition holds for $y_\theta(s, t^*)$. A result similar to Theorem 1 can be proven by adapting the ideas in [154, Ch. 8] to the 2D setting. Specifically, we have that

$$(i)\ \|y - p_{N-1}\|_\infty \le 4MNC_1\rho^{-N+1}, \tag{3.31}$$

$$(ii)\ |c_{jk}(\theta)| \le 4M\rho^{-(j+k)}, \tag{3.32}$$

where $C_1 = (2\rho - 1)/(1 - \rho)^2$.

As in the 1D case, we study the model manifold $\mathcal{P}$ associated with p_{N-1} as an approximation to $\mathcal{Y}$, the manifold for y_θ. We parameterize $\mathcal{P}$ using a vector of blocks, $P(\theta) = (B_0, \ldots, B_{N-1})^T$, where $B_j = (P_{0j}, P_{1(j-1)}, \ldots, P_{j0})$ and $P_{jk} = p_{N-1}(t_j, s_k; \theta)$. Since each block B_j has $j + 1$ entries, $P(\theta)$ is of length $n = N(N+1)/2$. Corresponding vectors of sample locations $\mathbf{t}$ and $\mathbf{s}$ are defined so that $P(\theta) = p_{N-1}(\mathbf{t}, \mathbf{s}; \theta)$.

As before, we exploit the decay of the bounds in Eq. (3.32) to show that P lies in the range of a matrix with strongly decaying singular val-

ues. To see this, define $\tilde{\mathbf{c}}$ as an appropriately ordered $n \times 1$ vector of the scaled coefficients $\tilde{c}_{jk} = \rho^{-(j+k)} c_{jk}$, and form the linear map $P(\theta) = X\tilde{\mathbf{c}}$. Here, $X = [X_{B^0} | \cdots | X_{B^{N-1}}]$, where X_{B^j} is a block of $j+1$ columns scaled by ρ^{-j}. Specifically, $X_{B^j} = \rho^{-j}[T_0(\mathbf{t})T_j(\mathbf{s}) \,|T_1(\mathbf{t})T_{j-1}(\mathbf{s})| \cdots |T_j(\mathbf{t})T_0(\mathbf{s})]$. Since $\tilde{\mathbf{c}}$ is constrained to lie in an n-sphere of radius $4M\sqrt{n}$, the manifold $\mathcal{P}$ is contained in a hyperellipsoid H_P with cross-sectional widths characterized by the singular values of X. One can show that the singular values of X must decay at, at least, a subgeometric rate. An argument similar to the one used in Theorem 2 shows that for $2 \leq j \leq n$,

$$\sigma_j(X) \leq \frac{3\sqrt{C_2}}{2} n \rho^{-\left\lfloor \sqrt{8(j-1)+1}/2 - 1/2 \right\rfloor}, \tag{3.33}$$

where $C_2 = (1 + \rho^{-2} + \rho^{-4})/(1 - \rho^{-2})^3$ and $\lfloor \cdot \rfloor$ represents the floor function. One can use H_P and Eq. (3.31) to explicitly construct a hyperellipsoid H_Y that must contain $\mathcal{Y}$. While our results are stated in terms of Chebyshev expansions, a similar argument can be made using 2D Taylor expansions, and all of these ideas extend naturally to the multidimensional case.

3.4 Summary

Our results explain a fundamental feature of the global geometry of sloppy models, and establish a rigorous framework that explains the role of model smoothness in the observation of sloppiness. An important implication of our results is that *any* model that satisfies the smoothness condition in Eq. (3.16) is *guaranteed* to be bounded in a manifold that exhibits this hierarchical structure. As such, it serves as a natural test of sloppiness. The implications of sharper bounds that depend on time-points are the focus of future work, as they open up far-ranging applications in optimizing the experimental design to focus data

collection at time-points that maximize information extraction by minimizing the decay rate in hyperribbon widths. Furthermore, sloppy features appear in probabilistic models, such as the Ising Model of atomic spins in statistical physics and the dark energy cold dark matter ΛCDM cosmological predictions of the cosmic microwave background discussed in Chapter 4. An extension of this approach could be used to constrain the predictions for general, probabilistic models (beyond the least-squares models considered in this chapter), and is the focus of future study.

CHAPTER 4

VISUALIZING PROBABILITY DISTRIBUTIONS AND THE INTENSIVE EMBEDDING

In this Chapter[1], we develop a new nonlinear manifold learning technique which achieves a compromise between preserving local and global structure. We accomplish this by developing an isometric embedding for general probabilistic models based on the replica trick [104]. By taking the number of replicas to zero, we reveal an intensive property – an information density characterizing the distinguishability of distributions – ameliorating the canonical orthogonality problem and 'curse of dimensionality.' We then describe a simple, deterministic algorithm that can be used for any such model, which we call Intensive Principal Component Analysis (InPCA). Importantly, our method quantitatively captures global structure while preserving local distances.

We study five probabilistic models: (1) a biased coin, (2) the canonical set of one-dimensional Gaussian distributions, (3) the Ising model of magnetism, which defines probabilities of spin configurations given interaction strengths, (4) the learning trajectory of a neural network, which predicts the probability of an image representing a single handwritten digit given weights and biases, and (5) ΛCDM, which predicts the distribution of CMB radiation given fundamental constants of nature.

[1]Work in this chapter was done in collaboration with Colin Clement, Francesco De Bernardis, Michael Niemack and Jim Sethna. A manuscript has been accepted for publication, and a preprint is available on the arXiv [118]. We thank Mark Transtrum for guidance on algorithms and for useful conversations. We thank Pankaj Mehta for pointing out the connection to MDS. KNQ was supported by a fellowship from the Natural Sciences and Engineering Research Council of Canada (NSERC), and JPS and KNQ were supported by the National Science Foundation through grant NSF DMR-1312160 and DMR-1719490. MDN was supported by NSF grant AST-1454881.

In addition, we use three of these models (as well as the MNIST dataset of handwritten digits) as a basis for comparison between InPCA and two other established manifold learning techniques, t-SNE and diffusion maps.

4.1 Model Manifolds of Probability Distributions

Any measurement obtained from an experiment with uncertainty can generally be understood as a probability distribution. For example, when some data x is observed with normally distributed noise ξ of variance σ^2, under experimental conditions $\boldsymbol{\theta}_j$, a model is expressed as

$$x = f(\boldsymbol{\theta}_j) + \xi \quad \text{where} \quad \mathcal{L}(\xi) \backsim \mathcal{N}(0, \sigma^2), \tag{4.1}$$

where $f(\boldsymbol{\theta}_i)$ is a prediction given the experimental conditions. This relationship is equivalent to saying that the probability of measuring data x given some conditions $\boldsymbol{\theta}$ is:

$$\mathcal{L}(x \mid \boldsymbol{\theta}) \backsim \mathcal{N}(f(\boldsymbol{\theta}), \sigma^2). \tag{4.2}$$

More complicated noise profiles with asymmetry or correlations can be accommodated with this picture. Measurements without an underlying model can also be seen as distributions, where a measurement x_i with uncertainty σ can induce a probability $\mathcal{L}(x \mid x_i, \sigma)$ of observing new data x.

We define a probabilistic model $\mathcal{L}(\boldsymbol{x} \mid \boldsymbol{\theta})$, the likelihood of observing data $\boldsymbol{x}$ given parameters $\boldsymbol{\theta}$. The *model manifold* is defined as the set of all possible predictions, $\{\mathcal{L}(\boldsymbol{x} \mid \boldsymbol{\theta}_i)\}$, which is a surface parameterized by the model parameters $\{\boldsymbol{\theta}_i\}$. The parameter directions related to the longest distances along the model manifold have been shown to predict emergent behavior (how microscopic parameters lead to macroscopic behavior) [99]. We will see that InPCA orders its

principal components by the length of the model manifold along their direction, highlighting global structure. The boundaries of the model manifold represent simplified models which retain predictive power [149], and the constraint of data lying near the model manifold has been used to optimize experimental design [146].

4.2 Hypersphere Embedding

We promised an embedding which is both isometric and preserves global structures. We satisfy the first promise by considering the hypersphere embedding:

$$\{z_{\boldsymbol{x}}(\boldsymbol{\theta_i})\} = \left\{2\sqrt{\mathcal{L}(\boldsymbol{x} \mid \boldsymbol{\theta_i})}\right\}, \tag{4.3}$$

where the normalization constraint of $\mathcal{L}(\boldsymbol{x} \mid \boldsymbol{\theta})$ forces $z_{\boldsymbol{x}}$ to lie on the positive orthant of a sphere. A natural measure of distance on the hypersphere is the Euclidean distance, in this case also known as the Hellinger divergence [70]

$$d^2(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) \quad = \quad \|z(\boldsymbol{\theta_1}) - z(\boldsymbol{\theta_2})\|^2 = 8\left(1 - \sqrt{\mathcal{L}(\boldsymbol{x} \mid \boldsymbol{\theta}_1)} \cdot \sqrt{\mathcal{L}(\boldsymbol{x} \mid \boldsymbol{\theta}_2)}\right)^2, \tag{4.4}$$

where $\cdot$ represents the inner product over x. Now we can see that the hypersphere embedding is isometric: the Euclidean metric of this embedding is equal to the Fisher Information metric $\mathcal{I}$ (from Eq. (2.10)) of the model manifold [60],

$$d^2(z_i, z_i + dz_i) = \sum_i dz_i dz_i = \sum_{\alpha\beta} \mathcal{I}_{\alpha\beta} d\theta^\alpha d\theta^\beta. \tag{4.5}$$

The Fisher Information Metric (FIM) is the natural metric of the model manifold [5], so the hypersphere embedding preserves the local structure of the manifold defined by $\mathcal{L}(\boldsymbol{x} \mid \boldsymbol{\theta})$.

As the dimension of the data increases, almost all features become orthogonal to each other, and most measures of distance lose their ability to discriminate between the smallest and largest distances [18]. For the hypersphere embedding, we see that as the dimension of x increases, the inner product in the Hellinger distance of Eq. 4.4 becomes smaller as the probability is distributed over more dimensions. In the limit of large dimension, all non-identical pairs of points become orthogonal and equidistant around the hypersphere (a constant distance $\sqrt{8}$ apart), frustrating effective dimensional reductions and visualization.

To illustrate this problem with the hypersphere embedding, consider the Ising Model, which predicts the likelihood of observing a particular configuration of binary random variables (spins) on a lattice. The probability of a spin configuration is determined by the Boltzmann distribution, and is a function of a local pairwise coupling and a global applied field. The dimension is determined by the number of spin configurations, 2^N where N is the number of spins. Holding temperature fixed at one, we vary h and J: external magnetic field ($h \in (-1.3, 1.3)$) and nearest neighbour coupling ($J \in (-0.4, 0.6)$), using a Monte Carlo method weighted by Jeffrey's Prior to sample 12,000 distinct points (see Section D.2 for plot of parameter ranges). From the resulting set of parameters, we compute $X_{ij} = \{z_i(\boldsymbol{\theta}_j)\}$ using the Boltzmann distribution, and visualize the model manifold in the N-sphere embedding of Eq. 4.3 by projecting the predictions onto the first three principal components of X. Figure 4.1(a) shows this projection of the model manifold of a 2×2 Ising model which is embedded in 2^4 dimensions. Figure 4.1(b) shows a larger, 4×4 Ising model, of dimension 2^{16}. As the dimension is increased from 2^4 to 2^{16}, we see the points starting to wrap around the hypersphere, becoming increasingly equidistant and less dis-

tinguishable.

A natural way to increase the dimensionality of a probabilistic model is by drawing multiple samples from the distribution. If D is the dimension of x, then N identical draws from the distribution will have dimension D^N. The more samples drawn, the easier it is to distinguish between distributions, mimicking the 'curse of dimensionality' for large systems. We see this demonstrated for our Ising model in Fig. 4.1(c), where we drew 4 replica samples from the same model. Notice that as compared to the original 2×2 model, the model manifold of the 4-replica 2×2 model 'wraps' more around the hypersphere, just like the larger, 4×4 Ising model. High dimensional systems have 'too much information,' in the same way that large numbers of samples have 'too much information'. In the next section, we consider the contraposition of the insight that a large number of replicas leads to the the curse of dimensionality, and discover an embedding which is not only isometric but also ameliorates the high-dimensional wrapping around the n-sphere.

4.3 Replica Theory

We saw in Fig. 4.1 that increasing the dimension of the data led to a saturation of the distance function Eq. 4.4. This problem is referred to as the loss of relative contrast or the concentration of distances [18], and to overcome it requires a non-Euclidean distance function, discussed below. In the last section we saw the same saturation of distance could be achieved by adding replicas, increasing the embedding dimension. Figure 4.2(a) shows this process taken to an extreme: the model manifold of the 2×2 Ising model with the number of

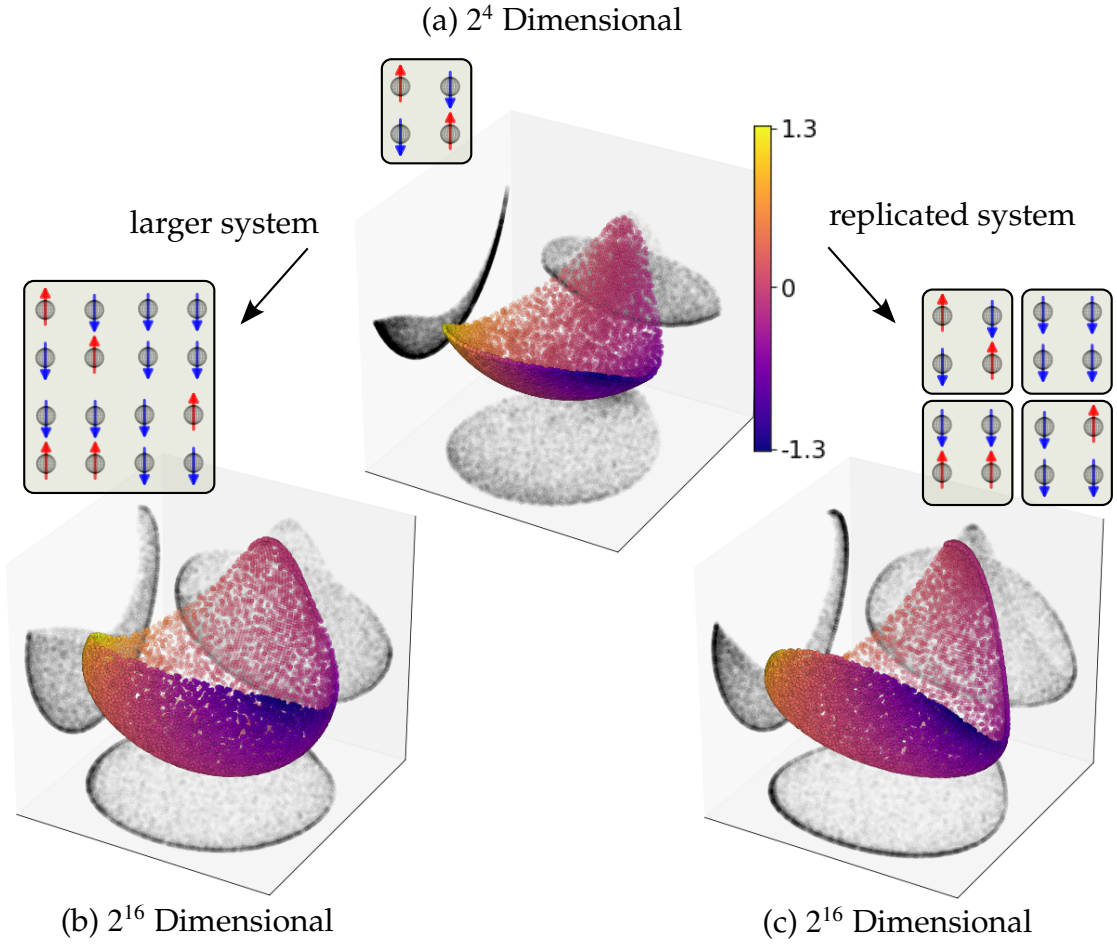


Figure 4.1: **Hypersphere embedding**, illustrating an embedding of the two dimensional Ising model. Points were generated through a Monte Carlo sampling and visualized by projecting the probability distributions onto the first 3 principal components. The points are colored by magnetic field strength. As the system size increases from 2×2 to 4×4, the orthogonality problem is demonstrated by an increase in 'wrapping' around the hypersphere. This effect can be also be produced by instead considering four replicas of the original system, motivating the replica trick which takes the embedding dimension or number of replicas to zero.

replicas taken to infinity. All the points cluster together, obscuring the fact that the underlying manifold is two-dimensional. In order to cure the abundance of information which makes all points on the hypersphere equidistant, we seek an intensive distance, such as the distance per number of replicas observed. Next, because the limit of many replicas artificially leads to the same symptoms of the curse of dimensionality, we will consider the limit of zero replicas, a procedure which is often used in the study of spin glasses and disordered systems [110]. Figure 4.2(b) shows the result of this analysis, the intensive embedding, where the distance concentration has been cured, and the inherent two-dimensional structure of the Ising model has been recovered.

To find the intensive embedding, we must first find the distance between replicated models. The likelihood for N replicas of a system is given by their product

$$\mathcal{L}\left(\{\boldsymbol{x_1}, \ldots, \boldsymbol{x}_N\} \mid \boldsymbol{\theta}\right))^{(N)} = \mathcal{L}\left(\boldsymbol{x_1} \mid \boldsymbol{\theta}\right) \cdots \mathcal{L}\left(\boldsymbol{x}_N \mid \boldsymbol{\theta}\right), \tag{4.6}$$

where the set $\{\boldsymbol{x_1}, \ldots, \boldsymbol{x}_N\}$ represents the observed data in the replicated systems. Writing the inner product or cosine angle between two distributions as

$$\langle \boldsymbol{\theta_1}; \boldsymbol{\theta_2} \rangle = \sqrt{\mathcal{L}\left(\boldsymbol{x} \mid \boldsymbol{\theta_1}\right)} \cdot \sqrt{\mathcal{L}\left(\boldsymbol{x} \mid \boldsymbol{\theta_2}\right)}, \tag{4.7}$$

where again $\cdot$ represents the inner product over x, and using Eq. (4.4), the distance per replica d_N^2 between two points on the model manifold is

$$d_N^2(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = \frac{d^2(\boldsymbol{\theta_1}, \boldsymbol{\theta_2})}{N} = -8\frac{\langle \boldsymbol{\theta_1}; \boldsymbol{\theta_2} \rangle^N - 1}{N}. \tag{4.8}$$

We are now poised to define the intensive distance by taking the number of replicas to zero

$$d_I^2(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = \lim_{N \to 0} d_N^2(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = -8 \log \langle \boldsymbol{\theta_1}; \boldsymbol{\theta_2} \rangle . \tag{4.9}$$

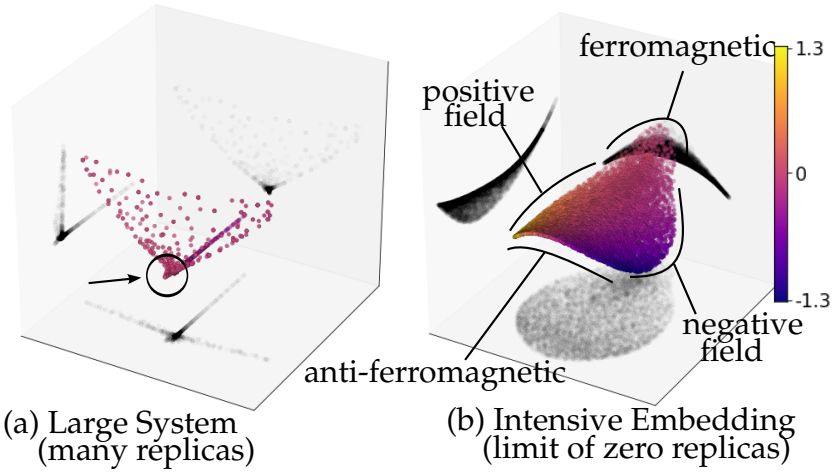


Figure 4.2: **Replicated Ising Model** illustrating the derivation of our intensive embedding. All points are coloured by magnetic field strength. (a) Large dimensions are characterized by large system sizes; here we mimic a 128×128 Ising model which is of dimension 2^{128^2} (very high dimensional). The orthogonality problem becomes manifest as all points are effectively orthogonal, producing a useless visualization with all points clustered in the cusp. (b) Using replica theory, we tune the dimensionality of the system and consider the limit as the number of replicas goes to zero. In this way, we derive our intensive embedding. Because only two parameters are varied, we know that the manifold is two-dimensional, a feature captured in the intensive embedding. Note that the z-axis reflects a negative-squared distance, a property which allows violations of the triangle inequality and is discussed in the text.

The last equality is achieved using the standard trick in replica theory, $(x^N - 1)/N \to \log x$ as $N \to \infty$, which is a basis trick used to solve challenging problems in statistical physics [110]. The trick is most evident using the identity $x^N = \exp(\log Nx) \approx 1 + N \log x$. We show in Section A.3 that the intensive distance is isometric:

$$d_I^2(\boldsymbol{\theta}, \boldsymbol{\theta} + \delta\boldsymbol{\theta}) = \delta\theta^\alpha \delta\theta^\beta g_{\alpha\beta} = \delta\theta^\alpha \delta\theta^\beta \mathcal{I}_{\alpha\beta}, \tag{4.10}$$

where again $\mathcal{I}$ is the Fisher Information Metric in Eq. (2.10), so that we can be confident the intensive embedding distance preserves local structures.

Importantly, the intensive distance does not satisfy the triangle inequality (and is thus not Euclidean, as will be discussed in Section 4.5): the distance between points on the hypersphere can go to infinity, rather than lie constrained to the finite radius of the hypersphere embedding. Because of this, the intensive embedding can overcome the loss of relative contrast [18] discussed at the beginning of this section. Distances in the intensive embedding maintain distinguishability in high dimensions, as illustrated in Fig. 4.2(b), wherein the two dimensional nature of the Ising model has been recovered. We hypothesize that this process, which cures the curse of dimensionality for models with too many samples, will also cure it for models with intrinsically high-dimensionality. The intensive distance obtained here is proportional to the Bhattacharyya distance [19]. By considering the zero replica limit of the Hellinger divergence, we discovered a new way to derive the Bhattacharyya distance. The importance of this will be discussed further in the following section.

4.3.1 Connection to Least Squares

Consider the concrete and canonical paradigm of models $y_{\boldsymbol{\theta}}(t_i)$ with data points x_i and additive white Gaussian noise, usually called a nonlinear least-squares model (and three examples of which are given in Section 3.2). The likelihood $\mathcal{L}(\boldsymbol{x} \mid \boldsymbol{\theta})$ is defined by

$$-\log \mathcal{L}(\boldsymbol{x} \mid \boldsymbol{\theta}) = \sum_i \frac{(y_{\boldsymbol{\theta}}(t_i) - x_i)^2}{2\sigma_i^2} + \log \mathcal{Z}(\boldsymbol{\theta}), \tag{4.11}$$

where $\mathcal{Z}$ sets the normalization. A straightforward evaluation of the intensive distance given by Eq. (4.9) and shown in Section A.6 finds for the case of non-

linear least squares

$$d_I^2(\boldsymbol{\theta_1}, \boldsymbol{\theta_2}) = \sum_i \frac{(y_{\theta_1}(t_i) - y_{\theta_2}(t_i))^2}{\sigma_i^2}, \tag{4.12}$$

so that the intensive distance is simply the variance-scaled Euclidean distance between model predictions.

4.4 Intensive Embedding

Classical Principal Component Analysis (PCA) takes a set of data examples and infers features which are linearly uncorrelated. [77]. The features to be analyzed with PCA are compared via their Euclidean distance. Can we generalize this comparison to utilize our intensive embedding distance? Given a matrix of data examples $X \in \mathbb{R}^{m \times p}$ (with features along the rows), PCA first requires the mean-shifted matrix $M_{ij} = X_{ij} - \bar{X}_i = PX$, where $P_{ij} = \delta_{ij} - 1/p$ is the mean-shift projection matrix and p is the number of sampled points. The covariance and its eigenvalue decomposition are then

$$\mathrm{Cov}(X, X) = \frac{1}{p} M^T M = X^T P P X = V \Sigma V^T, \tag{4.13}$$

where the orthogonal columns of the matrix V are the natural basis onto which the rows of M are projected:

$$MV = (UDV^T)V = UD = U\sqrt{\Sigma}, \tag{4.14}$$

where the columns of $U\sqrt{\Sigma}$ are called the principal components of the data X.

The principal components can also be obtained from the cross-covariance matrix, MM^T, since

$$MM^T = PXX^T P = (UDV^T)(UDV^T)^T = U\Sigma U^T. \tag{4.15}$$

The eigenbasis U of the cross-covariance is the natural basis for the components of the data, and the eigenbasis V of the covariance is the natural basis of the data points. For us this flexibility is invaluable, as the cross-covariance is more natural for expressing the distances between distributions of different parameters.

Writing our data matrix as $X_{ij} = z_i(\theta_j)$ using Eq. (4.3) for replicated systems, the cross-covariance is

$$\begin{aligned}(MM^T)^{(N)}_{ij} &= (PXX^TP)_{ij} \\ &= (z(\boldsymbol{\theta_i}) - \bar{z}) \cdot (z(\boldsymbol{\theta_j}) - \bar{z}) \\ &= 4\left\langle \boldsymbol{\theta}_i; \boldsymbol{\theta}_j \right\rangle^N + \frac{4}{p^2}\sum_{k,k'=1}^{p} \langle \boldsymbol{\theta}_k; \boldsymbol{\theta}_{k'} \rangle^N - \frac{4}{p}\sum_{k=1}^{p}\left(\langle \boldsymbol{\theta}_i; \boldsymbol{\theta}_k \rangle^N + \left\langle \boldsymbol{\theta}_j; \boldsymbol{\theta}_k \right\rangle^N\right), \end{aligned} \tag{4.16}$$

where $\bar{z}$ is the average over all sampled parameters, and we used the definition of z in Eq. (4.6) extended to the replicated likelihood function in Eq. (4.6). As with the intensive embedding, we can take the limit as the number of replicas goes to zero to find

$$W_{ij} = \lim_{N\to 0} \frac{1}{N}(MM^T)^{(N)}_{ij}. \tag{4.17}$$

Explicitly, the intensive cross-covariance matrix (derived in Section A.7) is

$$\begin{aligned} W_{ij} &= 4\log\left\langle \boldsymbol{\theta}_i; \boldsymbol{\theta}_j \right\rangle + \frac{4}{p^2}\sum_{k,k'=1}^{p} \log\langle \boldsymbol{\theta_{k'}}; \boldsymbol{\theta_k} \rangle - \frac{4}{p}\sum_{k=1}^{p}\left(\log\langle \boldsymbol{\theta_i}; \boldsymbol{\theta_k} \rangle + \log\left\langle \boldsymbol{\theta_j}; \boldsymbol{\theta_k} \right\rangle\right) \\ &= (PLP)_{ij} \end{aligned} \tag{4.18}$$

where $L_{i,j} = 4\log\left\langle \boldsymbol{\theta}_i; \boldsymbol{\theta}_j \right\rangle$ and P is the same projection matrix as defined above. In taking the limit of zero replicas, the structure of the cross-covariance has transformed

$$PXX^TP \xrightarrow[N\to 0]{} PLP, \tag{4.19}$$

and thus the symmetric Wishart structure is lost. It is therefore possible to obtain negative eigenvalues in this decomposition, which give rise to imaginary com-

ponents in the projections. Note the similarity between the form of this cross-covariance, and the double-centered distance matrix used in PCA and multidimensional scaling (MDS). This arises because both InPCA and PCA/MDS rely on mean-shifing the input data before finding an eigenbasis. Thus we view InPCA as a natural generalization of PCA to probability distributions, and MDS to non-Euclidean embeddings.

4.4.1 InPCA Algorithm

In summary, Intensive Principal Component Analysis (InPCA) is achieved by the following procedure:

1. *Compute the cross-covariance matrix from a set of probability samples:* Compute W_{ij} as derived in Eq. (4.18).
2. *Compute the eigenvalue decomposition* $W = U\Sigma U^T$.
3. *Compute the coordinate projections,* $T = U\sqrt{\Sigma}$.
4. *Plot the projections* using the columns of T. Order the components based on the *magnitude* of the corresponding eigenvalues, from largest to smallest (*i.e.* the first projected component corresponds with eigenvalue of largest magnitude)[2].

[2]Understanding the optimization process, order of eigenvalues, and the full effect of the negative squared directions is the focus of ongoing work with Han Kheng Teoh.

4.5 Properties of the Intensive Embedding and InPCA

The new space characterized by our intensive embedding has two weird properties: first it is formally one dimensional, yet there are multiple orthogonal directions upon which it can be projected, and second it is Minkowski-like, in that it has negative squared distances, violating the triangle inequality. We posit that, fundamentally, this second property is what allows InPCA to cure the orthogonality problem.

We begin with a discussion of the the one-dimensional nature of the embedding space. The embedding dimension is given by D^N where D is the original dimension of data x and N is the number of replicas. In the case of non-integer replicas the space becomes 'fractional' in dimension, and in the limit of zero replicas ultimately goes to one. However, it is still possible to obtain projections themselves along the dominant components of this space, by leveraging the cross-covariance instead of the covariance, summarized in step 2 of our algorithm. Visualizations produced by InPCA are cross-sections of a space of the dimension equal to the number of sampled points of the model manifold, instead of the dimension D or D^N.

In the limit of zero replicas in Eq. (4.18), the positive-definite, Wishart structure of the cross-covariance matrix is lost. It is therefore possible to have negative squared distances. The Minkowski-like nature of the embedding does not suffer from the concentration of distances which plagues Euclidean measures in high dimensions, thus allowing the model manifold to be 'unwound' from the N-sphere and for InPCA to produce useful, low-dimensional representations.

Finally, the eigenvalues of InPCA correspond to the cross-sectional widths of

the model manifold. We see this quite explicitly with the following example of a biased coin (specifically, in Fig. 4.3(b)) where the eigenvalues extracted from InPCA map directly to the manifold widths measured along the direction of the corresponding InPCA eigenvector. Therefore, we see that InCA produces a hierarchy of directions, ordered by the global widths of the model manifold. Note that, as with classical PCA, this correspondence depends on how faithfully the model manifold was originally sampled, *i.e.* InPCA can only tell you about the structure of the manifold from observed points.

4.6 Examples

We illustrate InPCA with five different probabilistic models. The first two, a biased coin in Section 4.6.1 and Gaussians in Section 4.6.2, are simple intuitive models with known properties of their respective manifolds. They reveal the importance of the negative-squared distances. We then apply InPCA to three more complicated models (Ising model in Section 4.6.3, neural network in Section 4.6.4, and the ΛCDM cosmological model in Section 4.6.5) to explore properties of the models themselves.

Finally, as an application to data, we apply InPCA to the MNIST dataset of handwritten digits in Section 4.6.6.

4.6.1 Coin Toss

To illustrate the Minkowski-like nature of InPCA, consider a biased coin. Given some bias θ for the coin (representing the likelihood of heads), the probability

state vector is given by

$$(\mathcal{L}(Heads \mid \theta), \mathcal{L}(Tails \mid \theta)) = (\theta, 1-\theta) \tag{4.20}$$

$$= (\cos^2\tilde{\theta}, \sin^2\tilde{\theta}), \tag{4.21}$$

defining a one-dimensional manifold, where we have re-parametrized the probability distributions to be in terms of $\tilde{\theta} \in [0, \pi/2]$. From Eq. (4.7), we compute the cosine-angle between two distributions to be:

$$\left\langle \tilde{\theta}_1; \tilde{\theta}_2 \right\rangle = \cos\tilde{\theta}_1 \cos\tilde{\theta}_2 + \sin\tilde{\theta}_1 \sin\tilde{\theta}_2 \tag{4.22}$$

In a useful embedding, one would wish 'all-heads' and 'all-tails' states to be far apart. Here, we have that 'all-heads' corresponds to $\tilde{\theta}_1 = 0$ and 'all-tails' to $\tilde{\theta}_2 = \pi/2$, and so the cosine-angle is zero. From Eq. (4.9) we see that the intensive distance between these two thus goes to infinity, making the two extremely biased coins infinitely far apart.

Figure 4.3 shows the top two InPCA components of the biased coin model manifold, which are related to the bias and variance of the coin. Curves of constant distance from a fair coin are hyperbolas (gray lines): two points can be finitely far from a fair coin but infinitely far from each other (demonstrating the violation of the triangle inequality).

To generate the manifold lengths in Fig. 2.5, and the InPCA embedding of the manifold in Fig. 4.3, 1,999 probabilites were sampled from 0 to 1 (excluding 0 and 1). Figure 4.4 shows the one-dimensional plot of the different probabilities considered.

Using $\mathcal{L}(Heads) = \cos^2\theta$ and $\mathcal{L}(Tails) = \sin^2\theta$, points were uniformly sam-

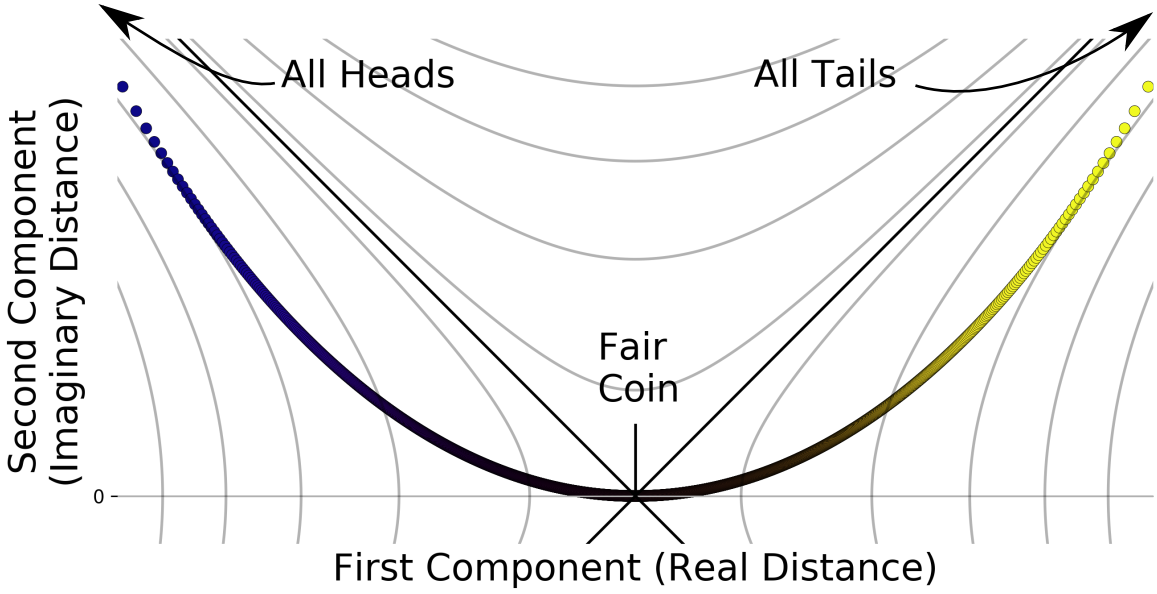


Figure 4.3: **InPCA visualization of biased coins**. The aspect ratio between axes is one. The first two InPCA components correspond to the coin bias and variance, yet the first is real and the second is imaginary. The contour lines represent constant distances from a fair coin and form hyperbolas: points can be a finite distance from a fair coin yet an infinite distance from each other.

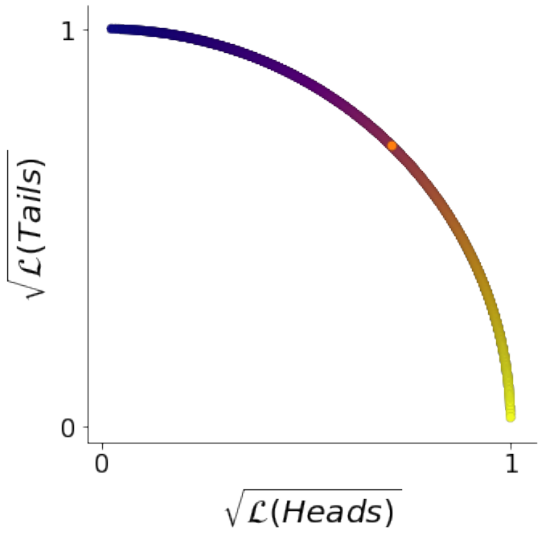


Figure 4.4: **Hypersphere of biased coin probabilities** showing the sampled ranged used in this thesis. Orange point in the middle represents a fair coin.

pled in θ to reflect a uniform Jeffreys Prior:

$$\mathcal{L}_{\text{Jeff}}(\theta) = \sqrt{\mathcal{I}_{\theta\theta}} = \sqrt{\left(\frac{\partial}{\partial\theta}\log(\cos^2\theta)\right)^2\cos^2\theta + \left(\frac{\partial}{\partial\theta}\log(\sin^2\theta)\right)^2\sin^2\theta} = 2 \tag{4.23}$$

4.6.2 Gaussians

A canonical probabilistic model is that of Gaussians[3] with varying means and standard deviations. The classic one-dimensional Gaussian is:

$$\mathcal{L}(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \tag{4.24}$$

and it is known that the set of Gaussians is embedded on a manifold of constant negative curvature and can be isometrically embedded on the Poincaré halfplane [33]. From Eq. (4.7), we compute the cosine-angle between two distributions to be:

$$\langle\{\mu_1, \sigma_2\}; \{\mu_2, \sigma_2\}\rangle = \sqrt{\frac{2\sigma_1\sigma_2}{\sigma_1^2 + \sigma_2^2}}\exp\left(-\frac{(\mu_1-\mu_2)^2}{4(\sigma_1^2+\sigma_2^2)}\right) \tag{4.25}$$

To visualize the model manifold, parameter ranges were sampled as shown in Fig. 4.5(a). The first two components extracted from InPCA are shown in Fig. 4.5(b). Gaussians that are easy to distinguish (narrow, with means that are far apart, *i.e.* $|\mu_1 - \mu_2| \gg 1$) are very far apart. As the widths increase and the resulting Gaussians have greater and greater overlap, they begin to cluster together in the lower peak of the figure.

[3]The model manifolds of Gaussians was explored in part with Qingyang Xu and Han Kheng Teoh.

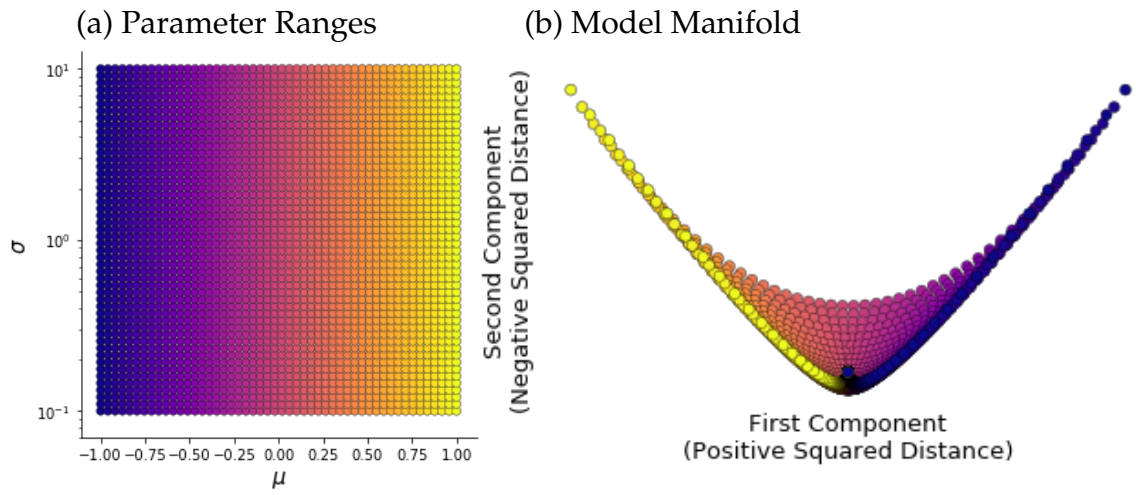


Figure 4.5: **Model manifold of Gaussians** visualized with InPCA, coloured by μ. (a) Parameter ranges considered to generate the manifold. (b) The longest direction corresponds to μ, and the second component is a reflection of σ. Points in the lower peak are Gaussians with large variance (very wide) and upper points are narrow Gaussians.

4.6.3 Ising Model

The canonical model from statistical physics considered in this chapter is the Ising model, described in Section 4.2. The likelihood of observing a particular spin configuration $\boldsymbol{S}$ is determined by the Boltzmann distribution:

$$\mathcal{L}(\boldsymbol{S} \mid \boldsymbol{\theta}) = \frac{1}{\mathcal{Z}(\boldsymbol{\theta})} \exp\left(-\mathcal{H}(\boldsymbol{S} \mid \boldsymbol{\theta})\right) \tag{4.26}$$

where $\mathcal{H} = \sum_\mu \theta^\mu \phi_\mu(\boldsymbol{S})$ is the Hamiltonian of the system, and $\mathcal{Z}(\boldsymbol{\theta})$ is the partition function which sets the normalization, given as

$$\mathcal{Z}(\boldsymbol{\theta}) = \sum_{\boldsymbol{S}} \exp\left(-\mathcal{H}(\boldsymbol{S} \mid \boldsymbol{\theta})\right). \tag{4.27}$$

Functions $\phi_\mu(\boldsymbol{S})$ are functions of the states used to represent the sum of spins, sum of nearest neighbour couplings, next-nearest neighbours, etc. Using

Eq. (4.9), the intensive distance between two Ising models is

$$d_I^2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = -8 \log\left(\sum_{\boldsymbol{s}} \sqrt{\mathcal{L}(\boldsymbol{s} \mid \boldsymbol{\theta}_1)}\sqrt{\mathcal{L}(\boldsymbol{s} \mid \boldsymbol{\theta}_2)}\right) \tag{4.28}$$

$$= -8 \log\left(\frac{1}{\sqrt{\mathcal{Z}(\boldsymbol{\theta}_1)}\sqrt{\mathcal{Z}(\boldsymbol{\theta}_2)}} \sum_{\boldsymbol{s}} \exp\left(-\frac{\mathcal{H}(\boldsymbol{s} \mid \boldsymbol{\theta}_1) + \mathcal{H}(\boldsymbol{s} \mid \boldsymbol{\theta}_2)}{2}\right)\right) \tag{4.29}$$

$$= 4\left(\log \mathcal{Z}(\boldsymbol{\theta}_1) + \log \mathcal{Z}(\boldsymbol{\theta}_2) - 2 \log \mathcal{Z}\left(\frac{\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2}{2}\right)\right) \tag{4.30}$$

$$= 4\left(\mathcal{F}(\boldsymbol{\theta_1}) + \mathcal{F}(\boldsymbol{\theta_2}) - 2\mathcal{F}\left(\frac{\boldsymbol{\theta}_1 + \boldsymbol{\theta}_2}{2}\right)\right) \tag{4.31}$$

where we make use of the linear properties of the Hamiltonian[4] to obtain the last line and reveal the relationship between the intensive distance and the concave difference in the free energy $\mathcal{F}(\boldsymbol{\theta}) = \frac{1}{T} \log \mathcal{Z}(\boldsymbol{\theta})$. This fundamental connection to the free energy makes the intensive distance uniquely suited to study the model properties of statistical systems, and is the focus of much ongoing work.

The first two components extracted from InPCA reveal important features in the model manifold: the high/low field regimes and the ferromagnetic/antiferromagnetic regimes, as shown in Fig. 4.2(b) and in the upper left of Fig. 4.6. If we also explore additional components, we obtain the hierarchy of widths shown in Fig. 2.5 as well as reveal a twist in the manifold around the critical point, shown in Fig. 4.6. The interpretable, hierarchical nature of the visualizations serve as a natural test of InPCA's utility[5]. It also forms the basis for future research, where manifold changes under coarse graining are used to better understand properties of the renormalization group and the geometry of the manifold near the critical point can be explored, following predictions in [124].

[4] This relationship was initially discovered by Archishman Raju.

[5] A comparison of different manifold learning methods on the manifold of Ising models is presented in Section E.1

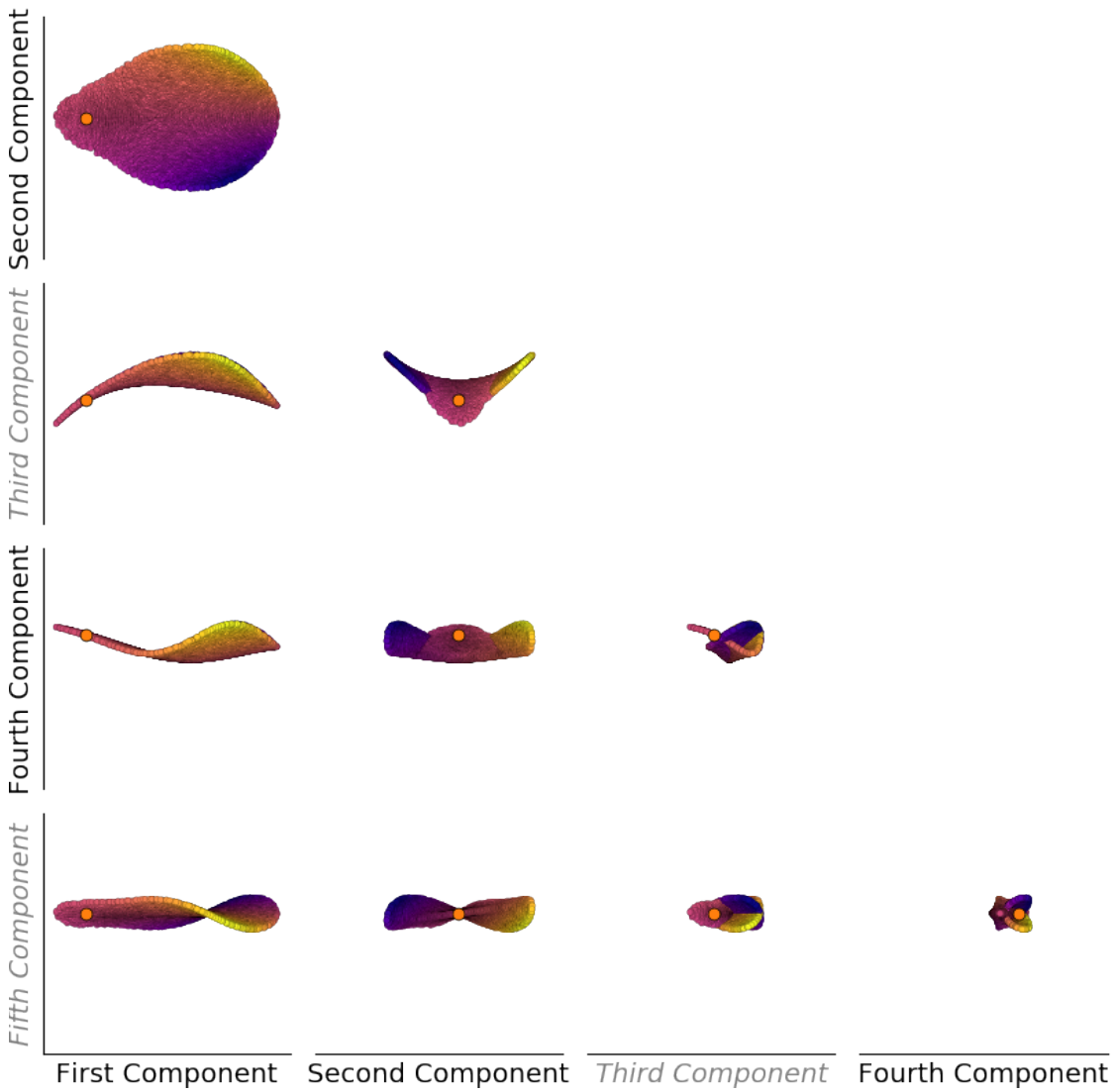


Figure 4.6: **Multiple manifold directions of the manifold of the Ising model** are visualized, and coloured by magnetic field. Orange point represents the critical point. The first and second components reveal the important high/low field and ferromagnetic/anti-ferromagnetic regimes. The second and fifth directions further reveal a twist in the manifold around the critical point. In the above figure, the third and fifths components are imaginary (negative squared distances).

4.6.4 Neural Network MNIST Digit Classifier

To demonstrate the utility of InPCA, we use it to visualize the training of a two layer convolution neural network (CNN), constructed using TensorFlow [1], and trained on the MNIST data set of hand-written digits[6]. A set of 55,000 images were used to train the network, which was then used to predict the likelihood that an additional set of 10,000 images are each classified as a specific digit between 0 and 9. We use softmax [20] to calculate the probabilities from the category estimates supplied by the network. The CNN defines the likelihood $\mathcal{L}(x \mid \theta)$ that some input image θ contains the image of a particular handwritten digit x. The InPCA projections of the CNN output in Fig. 4.7 visualizes the clustering learned by the CNN as a function of the number of learning epochs. The initialized network's model manifold shows no knowledge of the digits (colored dots), but as training commences, the manifold clearly separates digits into separate regions of its manifold. InPCA can therefore be used as a fast, interpretable, and deterministic method for qualitatively evaluating what a neural network has learned.

4.6.5 ΛCDM Predictions of the CMB

We compare our manifold learning technique to two standard methods, t-SNE and the diffusion maps, applied the six parameter ΛCDM cosmological model predictions of the cosmic microwave background (CMB). The ΛCDM predicts $\mathcal{L}(x \mid \theta)$ where x represents fluctuations in the CMB, and θ are the different cosmological parameters (*i.e.* it predicts the angular power spectrum of tempera-

[6] A comparison of different visualization methods for the CNN is presented in Section E.2.

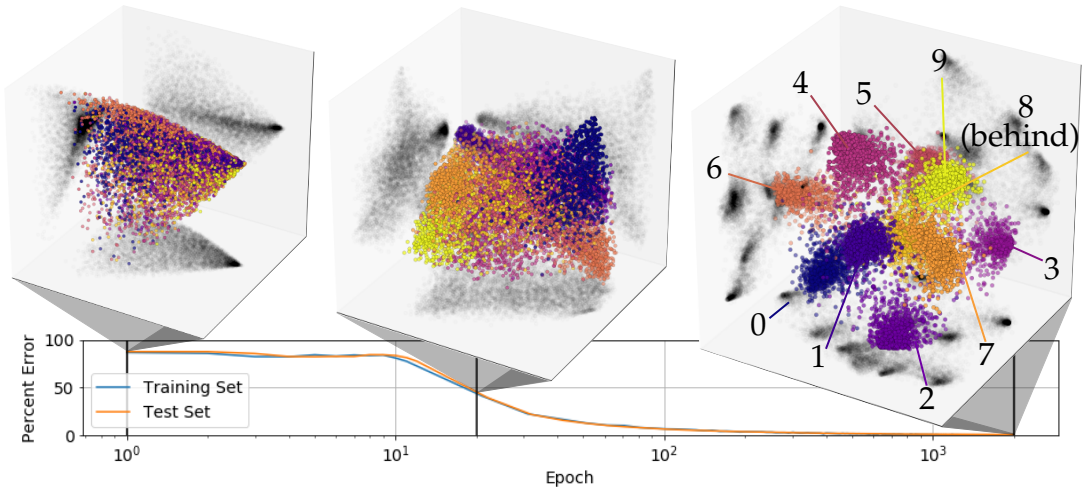


Figure 4.7: **Stages of training a convolutional neural network (CNN).** Each point in the above 3D projections represents one of 10,000 test image supplied to the CNN. At the first epoch, the neural network is untrained and so is unable to reliably classify images, with about a 90% error rate – an effect reflected in the cloud of points. As training progresses and error rate decreases, the cloud begins to cluster as shown by InPCA at the 20th epoch. Finally, when completely trained, the clustered regions are manifest at the 2000th epoch with ten clusters representing the ten digits.

ture and polarization anisotropies in sky maps of the CMB). Observations of the CMB from telescopes on satellites, balloons, and the ground provide thousands of independent measurements from large angular scales to a few arcminutes, that are use to fit model parameters. Here we only consider CMB observations from the 2015 Plank data release [114]. The ΛCDM model we consider has six parameters, the Hubble constant (H_0) which we sampled in a range of 20 to 100 $km\ s^{-1}\ Mpc^{-1}$, the physical baryon density ($\Omega_b h^2$) and the physical cold dark matter density ($\Omega_c h^2$) both sampled from 0.0009 to 0.8, the primordial fluctuation amplitude (A_s) sampled from 10^{-11} to 10^{-8}, the scalar spectral index (η) sampled from 0 to 0.98, and the optical depth at reionization (τ) sampled from 0.001 to 0.9.

To determine the likelihood functions, we use the CAMB software package to generate power spectra [94]. We perform a Monte Carlo sampling of 50,000 points around the best fit parameters provided by the 2015 Planck data release [114], with sample weights based on the intensive distance to the best fit. In Fig. 4.8(c), we see that the top two InPCA components correspond to A_s and the Hubble constant, parameters which control the two most dominant features in Planck data.

In Fig. 4.8 we show the first three components of the manifold embedding for InPCA, t-SNE, and diffusion maps. In order to apply t-SNE and the diffusion map to probabilistic data we must provide a distance. We therefore use our intensive distance, from Eq. (4.9), for consistency and ease of comparison. In all three cases, the first component from each method is directly related to the primordial fluctuation amplitude A_s, which reflects the amplitude of density fluctuations in the early universe, and is the dominant feature in real data [114]. The second InPCA component predicts the Hubble constant, whereas the diffusion map predicts the scalar spectral index (a reflection of the size variance of primordial density fluctuations). In all cases, the parameters are plotted against components values, and the Pearson coefficient of correlation (r) is calculated. The values of r range from -1 to 1, and a result of $|r| > 0.9$ indicates very high correlation [72].

The results from InPCA are shown in Fig. 4.9. The only parameters with very strong correlations to components are the primordial fluctuation amplitude (A_s) and Hubble constant (H_0), mapping to the first and second component respectively. Furthermore, there appears to be additional structure in the plots related to the physical baryon density ($\Omega_b h^2$) as well as the optical depth at reioniza-

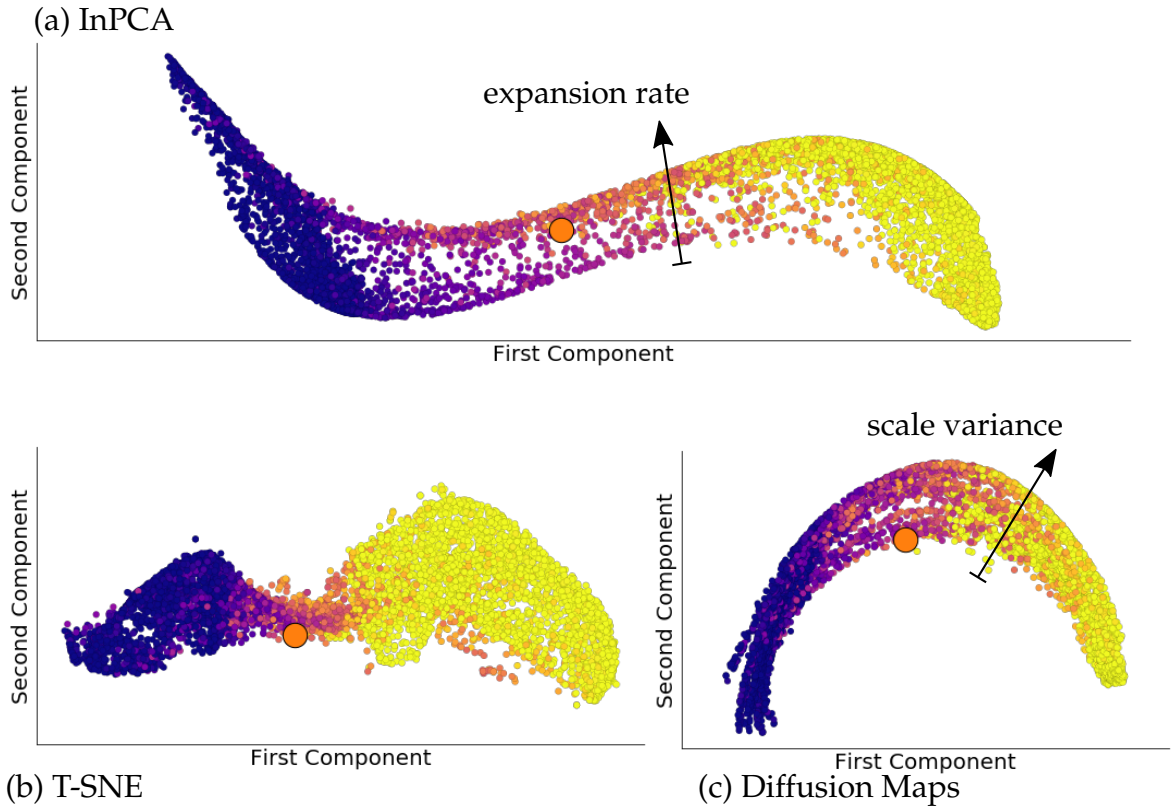


Figure 4.8: **Model manifold of the six parameter dark energy cold dark matter (ΛCDM) cosmological model predictions of temperature and polarization power spectra in the CMB** using InPCA, t-SNE and the diffusion map. Axes reflect true aspect ratio from extracted components in all cases. Here the model manifold is colored by the primordial fluctuation amplitude, the most prominent feature in CMB data. (a) InPCA extracts, as the first and second component, this amplitude term as well as the Hubble constant. These parameters control the two most dominant features in the Plank data, and so reflect a physically meaningful hierarchy of importance. In contrast, (b) t-SNE only extract the amplitude term and (c) the diffusion map extract the amplitude term and a different parameter, the scalar spectral index η, which reflects the scale variance of the density fluctuations in the early universe. In all plots, the orange point represents our universe, as represented by Planck 2015 data.

tion (τ), however the effect of parameter combinations and degeneracies on the more nuanced geometry of the manifold is the focus of future work (some preliminary results are shown in Appendix F).

The results from t-SNE are shown in Fig. 4.10. The only parameter that maps directly to a component is the primordial fluctuation amplitude. This effect is not entirely unsurprising, as the strength of t-SNE is in revealing local features (such as clustering) and so the global properties of the manifold are not often manifested with this technique.

The results from diffusion maps are shown in Fig. 4.11. The primordial fluctuation amplitude is very strongly correlated with the first extracted component. Note that the visualization from diffusion maps produces a crescent-shaped object. The scalar spectral index (η) is strongly correlated with the radial component of this visualization (calculated as the Euclidean distance from a point to the center of the projection), with a Pearson coefficient of $r = -0.93$. Because $\eta = 1$ corresponds to scale invariance in the primordial density fluctuations of the early universe, an increasing radial component of the diffusion maps corresponds to an increase in scale variance.

Such stark differences between manifold learning methods are surprising, as all techniques aim to extract important features in the data distribution, *i.e.* important geometric features in the manifolds. Given the ranges of sampled parameters, one would expect the variation in the Hubble constant to relate in some way to one of the dominant components, which InPCA satisfies.

To understand why, we discuss anisotropies in the CMB and how they relate to the different model parameters. The anisotropy in the CMB can be expressed

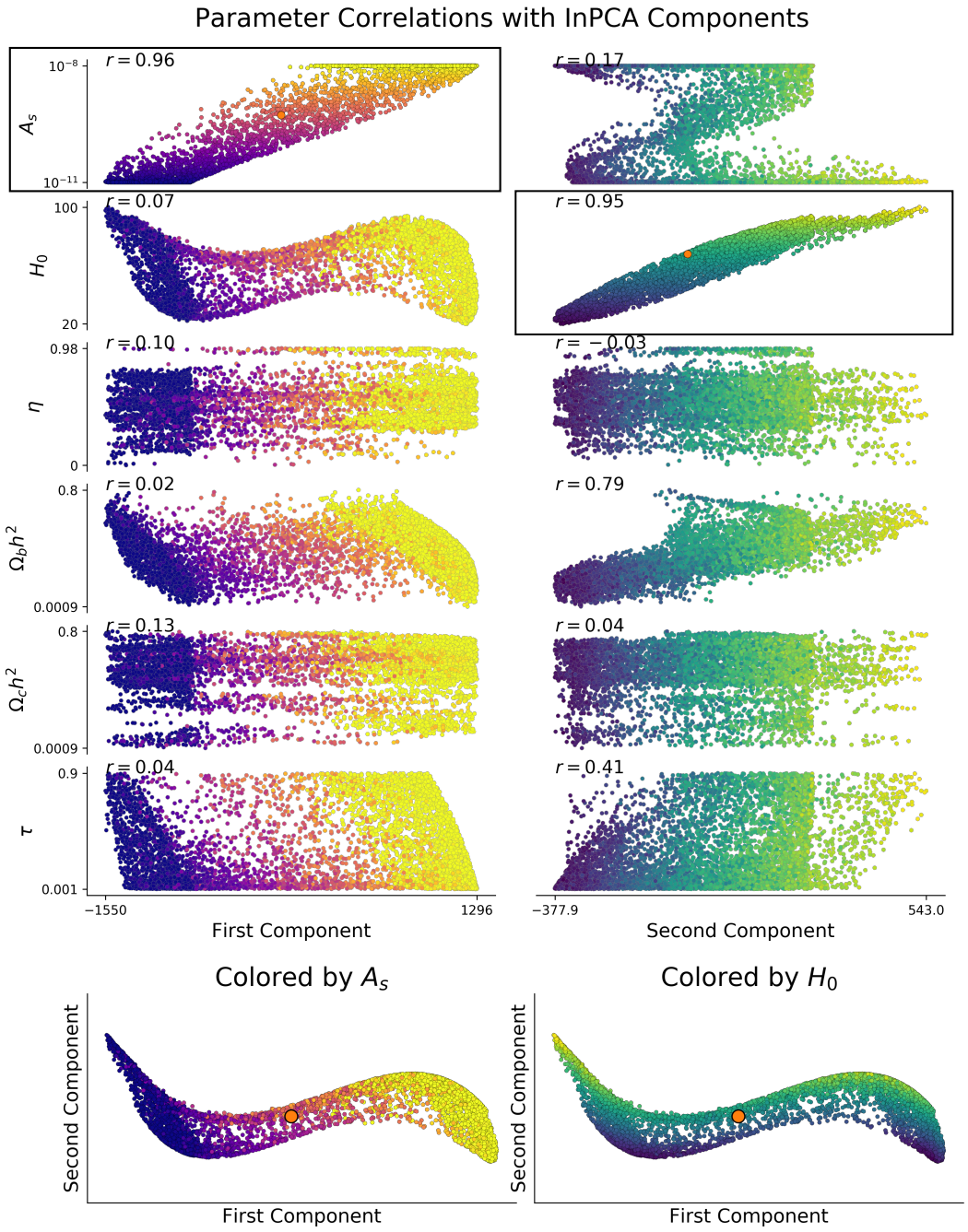


Figure 4.9: **The parameter correlations with InPCA components** for different CMB predictions of the cosmology model, with the Pearson coefficient (r in upper left of each plot) to determine the significance of correlations. The primordial fluctuation amplitude (A_s) and Hubble constant (H_0) are very strongly correlated with the first and second component, respectively. We show the different parameter regions on the manifold, with color maps that match the respective parameters.

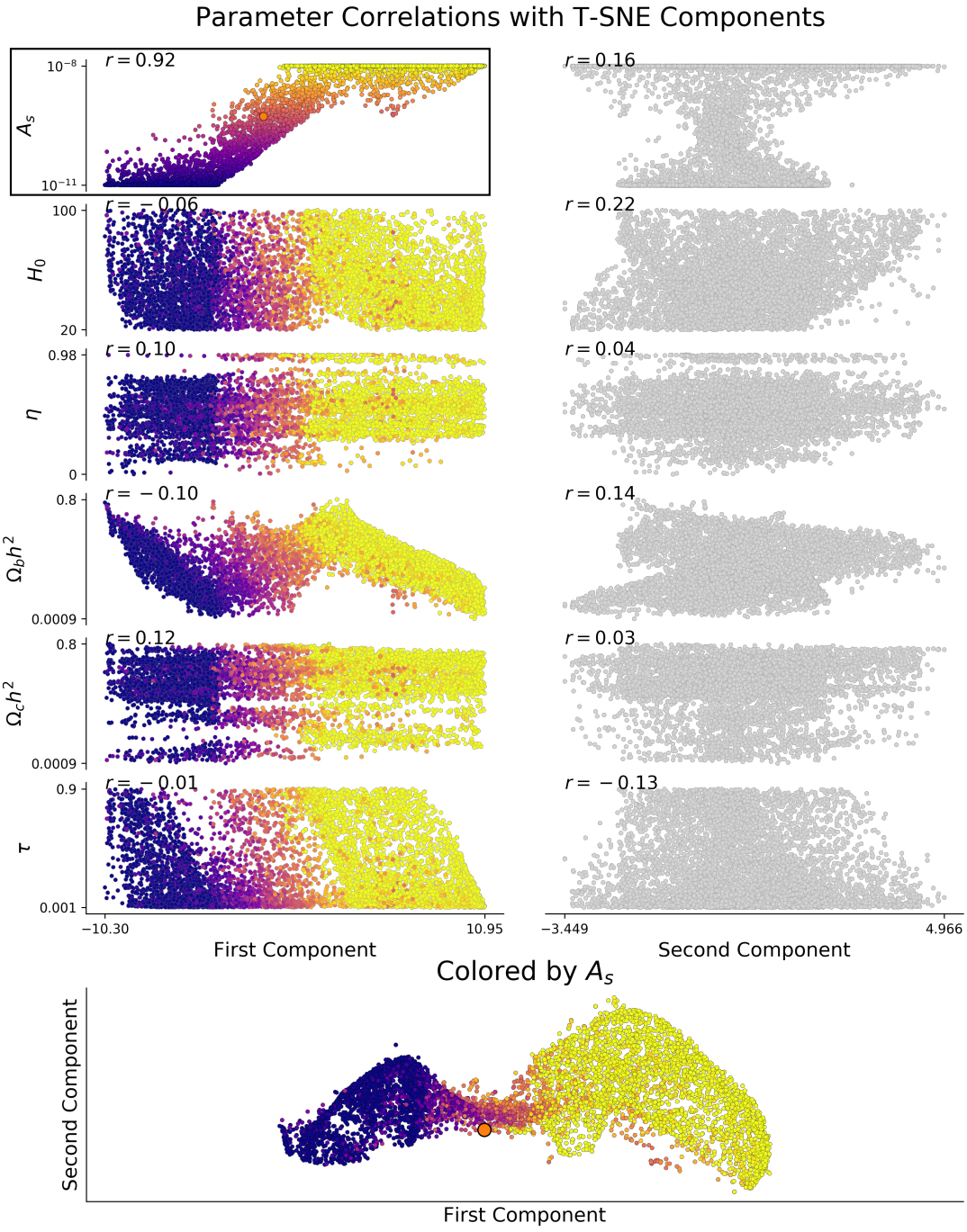


Figure 4.10: **The parameter correlations with t-SNE components** for different CMB predictions of the cosmology model considered in the main text. Superimposed on each plot the Pearson coefficient (r in upper left of each plot) to determine the significance of correlations. The primordial fluctuation amplitude (A_s) is very strongly correlated with the first component. We show the different parameter regions on the manifold in the bottom part of the figure.

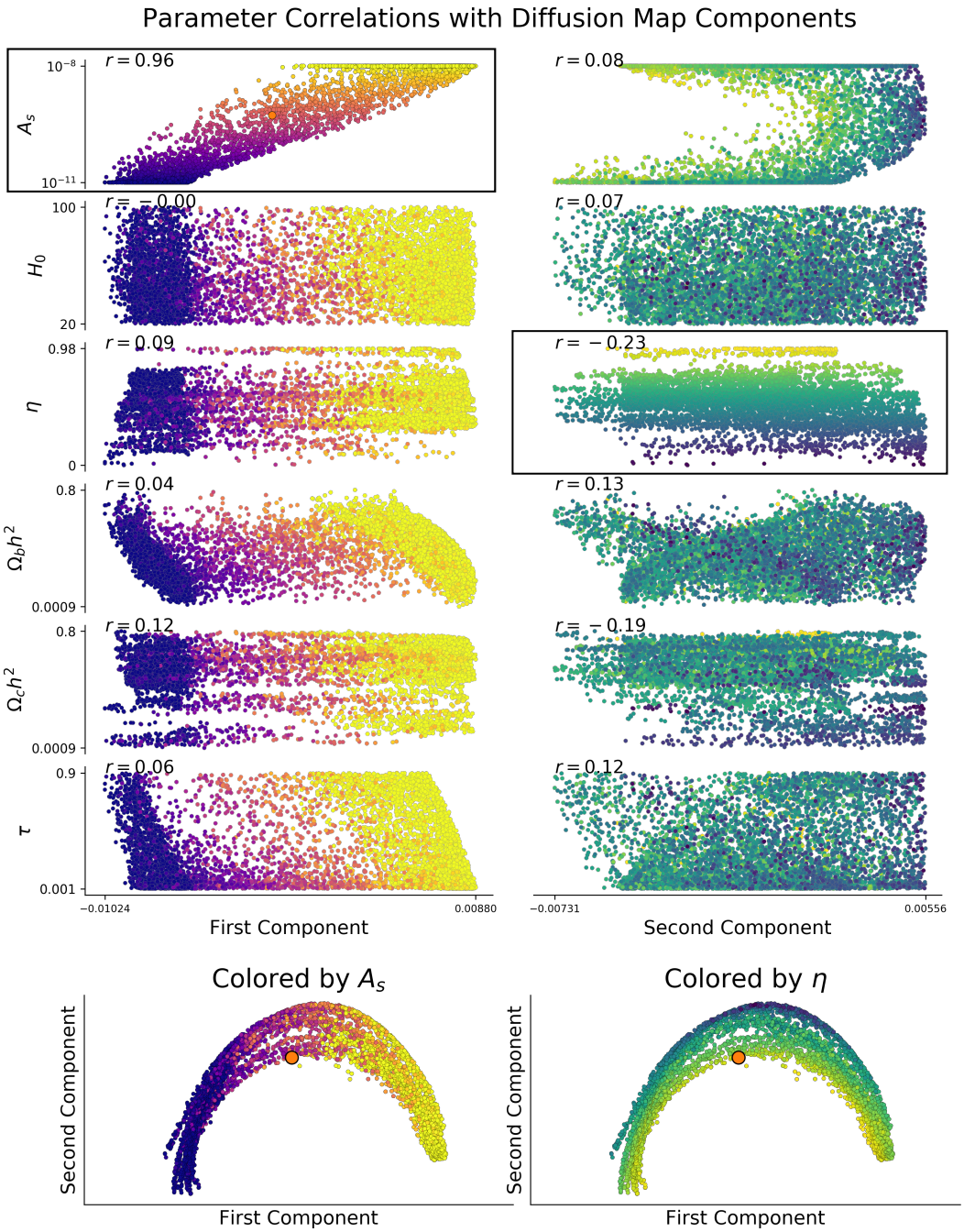


Figure 4.11: **The parameter correlations with diffusion map components** for different CMB predictions of the cosmology model, with the Pearson coefficient (r in upper left of each plot) to determine the significance of correlations. The primordial fluctuation amplitude (A_s) is very strongly correlated with the first component. The scalar spectral index (η) is very strongly correlated with the radial component of the crescent-shaped manifold visualization ($r = -0.93$).

as fluctuations in the spherical harmonics of a skymap of microwave radiation intensity, $\hat{a}_{\ell m} = \{\hat{a}^T_{\ell m}, \hat{a}^E_{\ell m}, \hat{a}^B_{\ell m}\}$ (where ℓ and m are coefficients in the multipole expansion)[7]. The correlation and cross-correlations of these fluctuations are expressed as a matrix:

$$C_\ell = \begin{pmatrix} C_\ell^{TT} & C_\ell^{TE} & 0 \\ C_\ell^{TE} & C_\ell^{EE} & 0 \\ 0 & 0 & C_\ell^{BB} \end{pmatrix}. \tag{4.32}$$

The likelihood functions of observing a particular set of fluctuations (expresed in terms of $\hat{a}_{\ell m}$) are high-dimensional Gaussians, expressed explicitly as

$$\mathcal{L}(\{\hat{a}_{\ell m}\} \mid \theta) = \prod_{\ell m} \frac{1}{\sqrt{(2\pi)^3 |C_\ell|}} \exp\left(-\frac{1}{2} \hat{a}^\dagger_{\ell m} C_\ell^{-1} \hat{a}_{\ell m}\right). \tag{4.33}$$

In this model, the C_ℓ vary with the six model parameters. The full nuance of how all parameters (and parameter combinations) impact the C_ℓ is beyond the scope of this chapter, however we highlight [156] as providing useful guide to understanding the impact of each parameter. As a crude illustration of the impact of the three main parameters extracted by the different manifold learning techniques, we show how the CMB spectra change as they are are varied in Fig. 4.12, following results from [79]. Increasing the primordial fluctuation amplitude increases the amplitude of the power spectra, as shown in Fig. 4.12(a). Increasing the Hubble constant shifts the power spectra, as shown in Fig. 4.12(b). In contrast, the scalar spectral index appears to primarily impact

[7] Expanded in greater detail in Section D.4

certain multipoles, while leaving others relatively unaffected. One therefore expects that parameters which strongly impact all multipoles) should be related to the dominant features in the manifold visualizations, which InPCA extracts via the primordial fluctuation amplitude and the Hubble constant.

4.6.6 MNIST Images

New manifold learning methods need to be applied to simple, standard learning tasks in order for their utility to be demonstrated and so that they can be contextualized within the larger set of established manifold learning methods. Unfortunately, because our method is specifically designed to address the problem of visualizing probabilistic data and models, few such tasks exist (*e.g.* the task of "unwinding the coil" or "visualize the sphere" would be inappropriate for our method, since these are not inherently probabilistic systems).

A standard test of manifold learning techniques is to visualize the set of images contained in the MNIST dataset [90]. A comparison of different methods on this task is shown in the python manifold learning package[8]. While this task is not inherently probabilistic, because the images are greyscale they can be interpreted as probability distributions by normalizing the pixels of an individual image (such that pure white pictures, uniform grey, and pure black would all be uniform distributions). To compare with these established methods, we show the outputs in Fig. 4.13. All three methods reveal underlying clusters, representing the different digits considered.

[8]The sklearn package for python provides a useful comparison with numerous manifold learning methods, available at https://scikit-learn.org/stable/auto_examples/manifold/plot_lle_digits.html

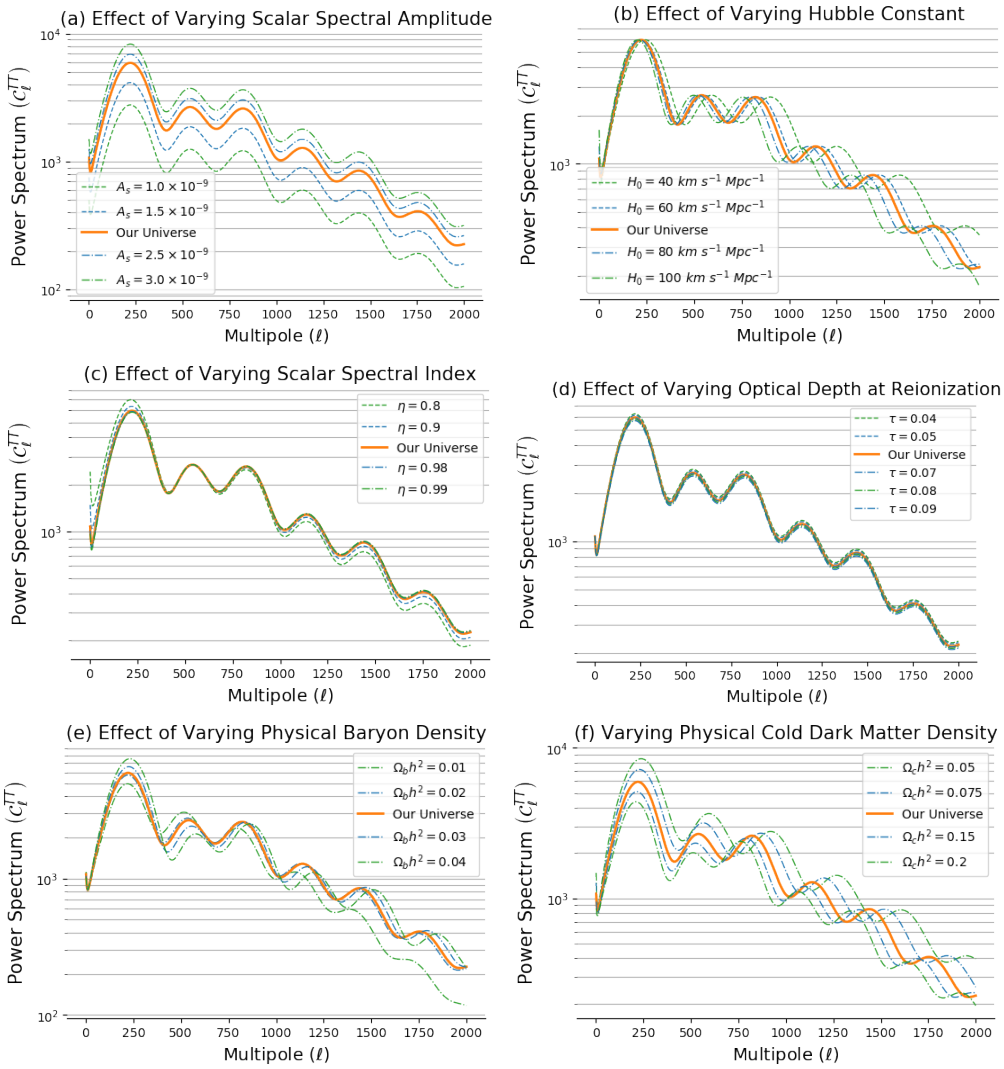


Figure 4.12: **Changes in CMB temperature (TT) spectra**, as different parameters are varied. (a) The impact of varying the primordial fluctuation amplitude (A_s) increases the amplitude of the power spectrum, impacting all mulitpoles. (b) Increasing the Hubble constant (H_0) shifts the power spectrum. (c) Diffusion maps extract the scalar spectral index (η) as a parameter which highly impact manifold features. However, η primarily impacts certain multipoles while having little effect on others. (d) Varying optical depth at reionization (τ), (e) varying physical baryon density, and (f) varying physical cold dark matter density. Spectra generated from CAMB software [94].

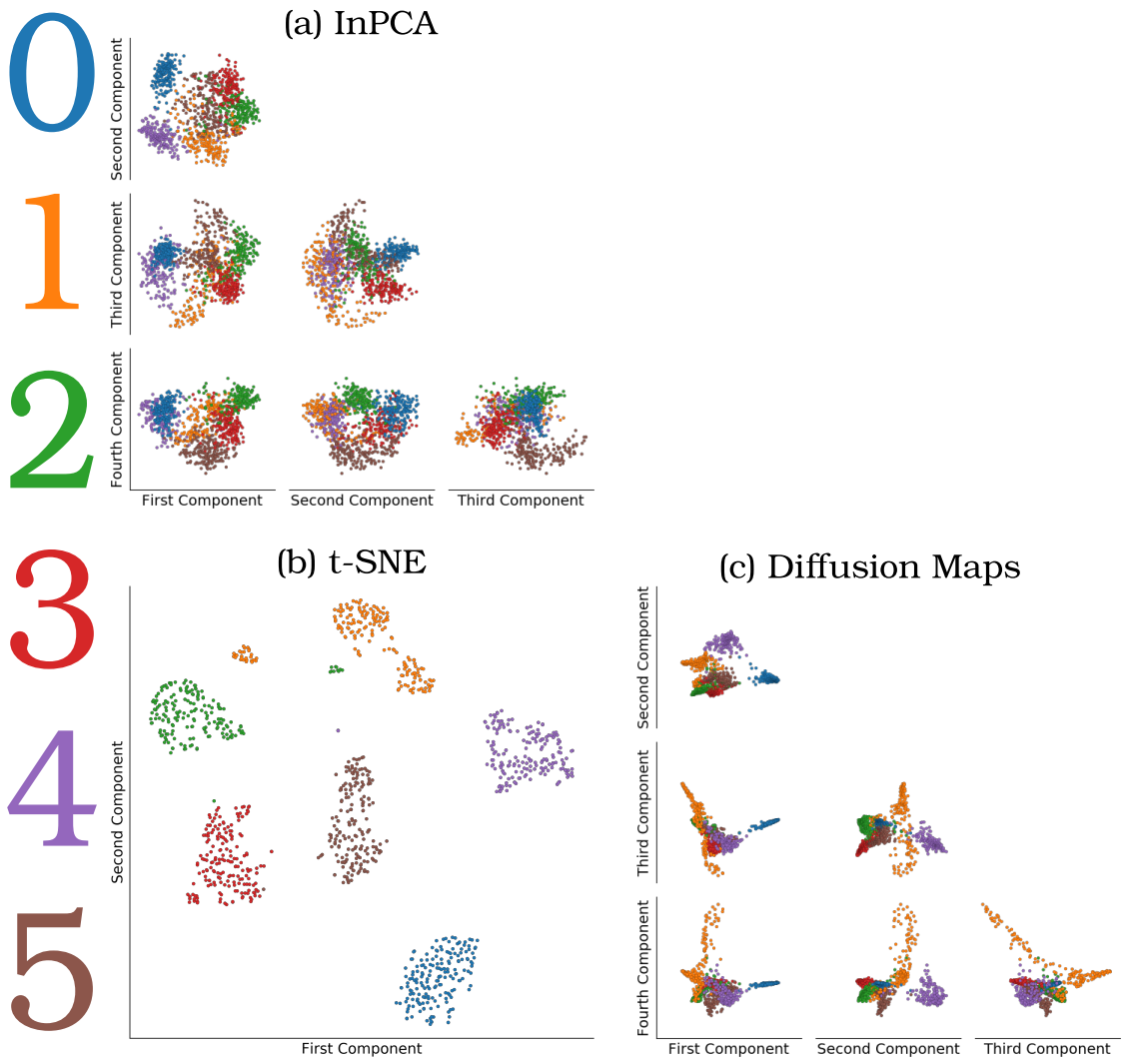


Figure 4.13: **Visualizing raw MNIST data.** All three manifold learning methods discussed in Section 4.6.5 (InPCA, t-SNE, diffusion maps) appear to cluster the raw images. We use the triangle plots to illustrate the number of components needed to cleanly visualize all 6 clusters. In this way, InPCA performs well on this more standard task. t-SNE, being specifically design to reveal clusters and local features, performs best on this task – however global properties (distance between clusters) are not meaningful.

4.7 Summary

There are two important differences between InPCA and other methods. First is that InPCA has no tunable parameters, and is a geometric object defined entirely by the model distribution. For example, t-SNE embeddings rely on parameters such as the perplexity, a learning rate, and a random seed (yielding non-deterministic results), and the diffusion maps rely on a diffusion parameter and choice of diffusion operator, all of which must be manually optimized to obtain good results. Second, t-SNE and diffusion maps embed manifolds in Euclidean spaces in a way which aims to preserve local features. However, InPCA seeks to preserve both global and local features, by embedding manifolds in a non-Euclidean (flat but Minkowski-like) space.

In this chapter, we introduced an unsupervised manifold learning technique, InPCA, which captures low-dimensional features of general, probabilistic models with wide-ranging applicability. We consider replicas of a probabilistic system to tune its dimensionality and consider the limit of zero replicas, deriving an intensive embedding that ameliorates the canonical orthogonality problem. Our intensive embedding provides a natural, meaningful way to characterize a symmetric distance between probabilistic data and produces a simple, deterministic algorithm to visualize the resulting manifold.

CHAPTER 5

QUANTIFYING GENDERED BEHAVIOR IN PHYSICS LABS

The demographic composition of physicists is not representative of the general population, with men over-represented not only in numbers but also in rank [113]. In exploring the underlying reasons for this, there has been a large focus on gaps in performance between men and women on concept inventories and course grades [132, 100]. While useful, this approach provides an incomplete picture [100, 8], and so there is a shift in physics education research towards investigating other metrics such as sociocultural factors [46, 130], self-efficacy [109], sense of belonging [95], and identity formation [81]. Moreover, participation in the physics community through the *roles* people take on can heavily shape one's identity as a physicist [80]. Understanding how these roles develop and how they are shaped through behaviours in physics courses is critical, as a gendered division of roles influences the modern practice of physics so greatly that it is laden with masculine connotations [57, 53].

In this chapter[1], we present the results of a study in which we explored gendered differences in behaviour in the context of physics labs. Labs provide an environment where students interact with peers and accumulate experience in ways that can influence their perception of physics and of themselves as physicists [101, 38]. We assume that the ways in which students behave provide information about their experience in the lab, and thus influence these percep-

[1]Work in this chapter was done in collaboration with Michelle Kelley, Kathryn McGill, Emily Smith, Zachary Whipps and Natasha Holmes. Part of this work has been published [119], and is available on the arXiv [120]. We also thank the teaching assistants and lab instructors for the course used in this study for their invaluable support and cooperation, as well as Chris Gosling for valuable conversations and insight. This study was supported by the President's Council for Cornell Women's Affinito-Stewart Grant and the Cornell's College of Arts and Sciences Active Learning Initiative.

tions. In what ways do gendered differences in behaviour manifest in physics labs? How do factors such as group composition and lab type impact these behaviours? To answer these questions, we analysed the quantified behaviour of students in an introductory physics lab to explore the gendered differences that manifest in these spaces and see how behavioural differences vary with pedagogical structure as well as group composition.

We explored student experience through the behaviours students take on in an introductory physics lab course, building on previous work regarding gendered action in first-year physics labs [40, 74, 119]. Previous work has shown mixed results with regards to gendered action in first-year physics labs such as men using desktop computers more than women [40], women using laptops more than men [119], and that management of equipment apparatus is heavily impacted by gender in mixed-group pairs [74]. To unify these seemingly disparate results, we generalized these studies within the framework of post-structural gender theory [25], described in Section 5.1.

Data for this study were collected at two levels of granularity. *Coarse behaviours* were captured for all students in multiple lab periods, and were determined based on what the students were handling (lab desktop computer, laptop or personal device, writing on paper, handling equipment, or other). *Detailed behaviours* were described for a subset of these students, by analysing videos that captured individual groups during entire lab periods. Specifically, coarse behaviours were used to determine differences in task division, and the video analysis was used to both describe behaviours within tasks and gain insight into how tasks were allocated.

We found that, in lab sections designed to foster collaborative work and

promote student agency, women used laptops or personal devices more than men, and men behaved differently depending on the gender composition of their group (specifically, men behaved differently depending on whether they were working with other men or with at least one woman). We found no such differences in our more traditional lab sections, in which students were guided through experiments and individually filled out paper worksheets. We conjecture that, due to the pedagogical differences in lab types, students in the inquiry labs were afforded the opportunity to divide tasks within their groups and therefore did so along gendered lines. We use these results to guide future research in task allocation and *positioning* (how one positions oneself and others into particular roles or stances trough verbal and nonverbal cues) [39, 16], as well as better characterize the gendered student roles that manifest in the context of physics labs.

5.1 Poststructural Gender Theory

While exploring gendered differences in performance on standardized tests and assessments can be useful, such research often does so in a manner that reflects a deficit model, where women are seen as deficient when compared to men [153]. Not only does this limit avenues of research, but such an approach can reinforce gendered inequalities of power as it reinforces the use of men's achievements as standard [54]. As Traxler expressed in [153],

> Is the goal to change women so that they can succeed in a culture where men are successful, or would it be better to change the culture so that the experience of men, particularly straight, white, married men, is not assumed to be the best standard?

To better understand gendered behaviour in physics labs, we use the framework described in poststructural gender theory [25]. In this framework, gender is not described as an unchanging aspect of a person but is instead performative. It is enacted through dress, speech, composure, and employment among other aspects [159] and is actively reinforced through daily social interactions [153, 25]. Gender is reflected through the *role* one assumes in a community. For example, through an anthropological study of an engineering department, researchers observed that students identified their peers as having particular roles such "leader", "jock" or "curve-breaker" [143]. Importantly, of the 36 roles documented in the study, only four were available to women (in the sense that the students only identified their female peers as being in one of those four roles). Furthermore, all four "female" roles were defined strictly in terms of social achievements (such as "sorority chick"), whereas the roles assumed by their male counterparts were defined in terms of both social and academic achievement (such as "frat boy" and "curve-breaker"). Because of the difference in roles available to men and women, there is a stark gender division with respect to available roles in undergraduate STEM courses [117, 58]. By assuming different roles in academic settings, men and women have very different academic experiences [38].

Furthermore, gender cannot be treated in isolation and must be considered in relation to multiple factors as part of identity formation, and viewed as a fluid, context-dependent state [25, 153]. We refer to *identity* as defined in [58]:

> the sum total of one's beliefs about oneself, one's actions, and how one's behavior is interpreted by others in a given context.

Identity formation is a complicated, multi-dimensional process that includes

gender, race, physical ability, socioeconomic status, sexual orientation, and religion, among many, many other factors. The formative process includes individual agency as well as broader cultural and societal factors [21]: the impact of the broader culture outside of the physics classroom strongly influences one's identity formation (such as a culturally-perceived notion of physics as a masculine field [10]). Importantly, how one develops a sense of identity impacts the set of available roles one may take on in a particular context, and strongly determines persistence in a particular field [27].

In Fig. 5.1, we show how the interplay between context and identity constrains the set of roles available for a student to take on. Here we consider the students' self-reported gender identity and the specific context of the lab type (lab sections of two pedagogically different structures) and the composition of their group (mixed-gender and single-gender groups).

As a way of probing the roles students assume in physics labs, we analyzed the quantified behaviour profiles (discussed further in Section 5.2.2) of students in multiple lab periods. We assume that students assuming very different roles will behave quite differently in labs, producing a measurable effect. Students in pedagogically different lab sections should behave quite differently: specifically, in labs with increased student agency, there should be a wider range of available roles, and so we expect a broader range of behaviours to be present. If men and women are assuming different roles in physics labs, then we expect behaviour differences along gender lines. Finally, if students assume different roles depending on the gender of their lab partners (*i.e.* whether they are in single-gender or mixed-gender groups) then we expect behaviours to vary with group composition.

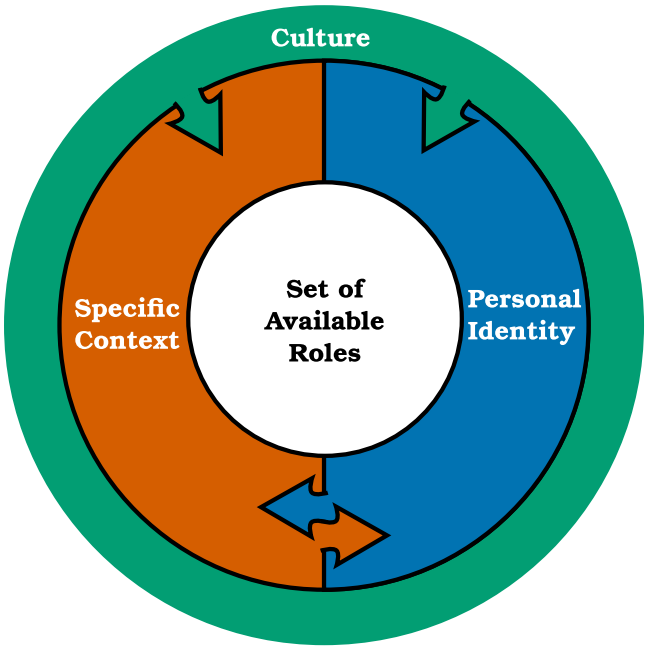


Figure 5.1: **Schematic of the theoretical framework** used in this chapter, illustrating how the broader culture, personal identity and specific context define the set of roles available to a student in a physics lab. Arrows are used to emphasize interactions between factors which influences the boundary.

5.2 Research Methods

All participants in this study were undergraduate students at a major research university. They were enrolled in the honours-level mechanics course of a calculus-based physics sequence (the first course in the sequence). The mechanics course was taught in both Fall 2017 and Spring 2018, and students from both semesters were included in this study.

During Fall 2017, all students attended the same lecture, mixed together in discussion sections, but were separated into two pedagogically different lab types discussed below (three *traditional lab* sections and two *inquiry lab* sections). During Spring 2018, the two lab sections under study were both inquiry labs.

Note that we observed students across multiple lab periods, and so while each student is in one lab section they appear in multiple lab periods. All participants were unaware of the differences between lab types: students in Fall 2017 self-selected into their lab sections prior to the start of the course, and only the inquiry lab sections were available to students in Spring 2018.

The *traditional labs* were designed to reinforce physics content knowledge presented in lecture. Students were provided with detailed paper worksheets to follow during lab, guiding them through experiments that provided them with hands-on experience. Importantly, while students worked in groups to collect data for the experiments, they were required to fill out worksheets individually and hand them in at the end of each lab period.

The *inquiry labs* were designed to emphasize the process of experimentation in physics. Students were provided with a specific goal, but were expected to design their own experiment in order to achieve that goal. Lab instruction and lab activities were focused on iterating, improving, and extending investigation. Importantly, in contrast to the traditional labs, students worked collaboratively to design and implement their experiments and submitted one electronic notebook per group.

5.2.1 Collecting Demographic Information

We used in-class surveys to obtain student demographic information. In all, 143 students across multiple lab sections were used in this study. While they had the option to disclose a gender other than *woman* or *man*, no student chose to do so, and only two students did not disclose their gender identity. As a result,

all students were included in the initial cluster analysis, however the gender analysis follows the traditional gender binary of *woman* or *man* (with the two undisclosed students omitted from the graphs in Fig. 5.6 and Fig. 5.8 due to insufficient statistics). Table 5.1 shows the demographic breakdown of student participants in this study.

Table 5.1: **Student demographics** of this study. Errors were computed using standard error for population fractions, shown in Eq. (G.1). In all, 143 students were considered in this study.

	Traditional Labs		**Inquiry Labs**	
	N	%	*N*	%
Women	11	19 ± 5	21	25 ± 5
Men	46	79 ± 5	63	74 ± 5
Undisclosed	1	2 ± 2	1	1 ± 1

In both the traditional and inquiry labs in Fall 2017, observers were present to document student behaviour and single groups were video and audio recorded. In Spring 2018, the whole class was video taped to document student behaviour and single groups were video and audio recoreded. These were the two means of data collection used in this study, and we elaborate on them in the following subsections.

5.2.2 Quantifying Coarse Student Behaviours

In all lab sections, observers documented student behaviours following the observation protocol used in [40]. Every five minutes, an observer noted each student's actions in the lab using one of five codes: Desktop, Equipment, Lap-

top, Paper, and Other. In Fall 2017 an observer was present in the lab, and in Spring 2018 the same procedure was applied to video recordings. One code was applied to each student in the class at each five-minute interval, except in cases where students could not be observed (*e.g.* because they were late or left early). The codes are described in Table 5.2, and were based on what a student could be handling in the lab. The Other code captured all other action such as engaging in whole-class discussions, writing on whiteboards, discussing with the TA or UTA, or off-task behaviours, and is used to ensure all in-lab time is coded. The Desktop code was separated from the Laptop code for three reasons. First, the desktop was property of the lab (with one provided at every lab bench) whereas the laptops belonged to individual students. Second, the Desktop was often required for data collection because it was directly connected to a detector or piece of equipment. Finally, while desktops were present in both lab types, only students in the inquiry labs actively used laptops to analyze data, document their lab procedures, and submit their electronic notebooks. Further descriptions of the codes were obtained through video analysis and are outlined in the following section.

To validate this method, two observers coded student actions in the same lab period using the described protocol but at different five-minute intervals. Observers were specifically not coding the same student at the same time. This was done to address two issues: (1) the reliability of the codes, and (2) the validity of the five-minute interval at capturing coarse student behaviours in a two-hour lab period. Note that because observers were explicitly not observing the same student at the same, percent agreement or calculating Cohen's Kappa would not provide the necessary information to validate the method. Instead, a standard chi-squared analysis was performed on the contingency table con-

Table 5.2: **Action codes used in observations**. The Laptop code is used for both handling a laptop or personal device (students used laptops, phones, and tablets for the purpose of notetaking, writeup, data analysis and reading instructions in the inquiry labs).

Code	Description
Desktop	Using the desktop computer at the lab bench.
Equipment	Handling equipment.
Laptop	Using a laptop or personal device.
Paper	Writing on paper or in a notebook.
Other	Other action or behaviour.

structed from the accumulated codes (the frequency each observer noted each code, summed over all students). We provide an example of observer comparisons in Section G.2 for illustrative purposes. We used the criteria that if two sets of observations are statistically indistinguishable from each other, then the observers captured the same distributions. In all cases observers' distributions were statistically indistinguishable, and so single observers coded subsequent lab periods.

Because students were observed during multiple lab periods over a full semester, we were able to document individual students more than once. As a result, we obtained 522 unique *student profiles*, each quantifying the actions of one student in one lab period through the frequency of associated codes. We show a schematic of the lab breakdowns in Fig. 5.2 to illustrate the connection between students and student profiles. Table 5.3 shows a demographic breakdown of the student profiles used in this study.

While a natural analysis on such data could be a comparison of mean fre-

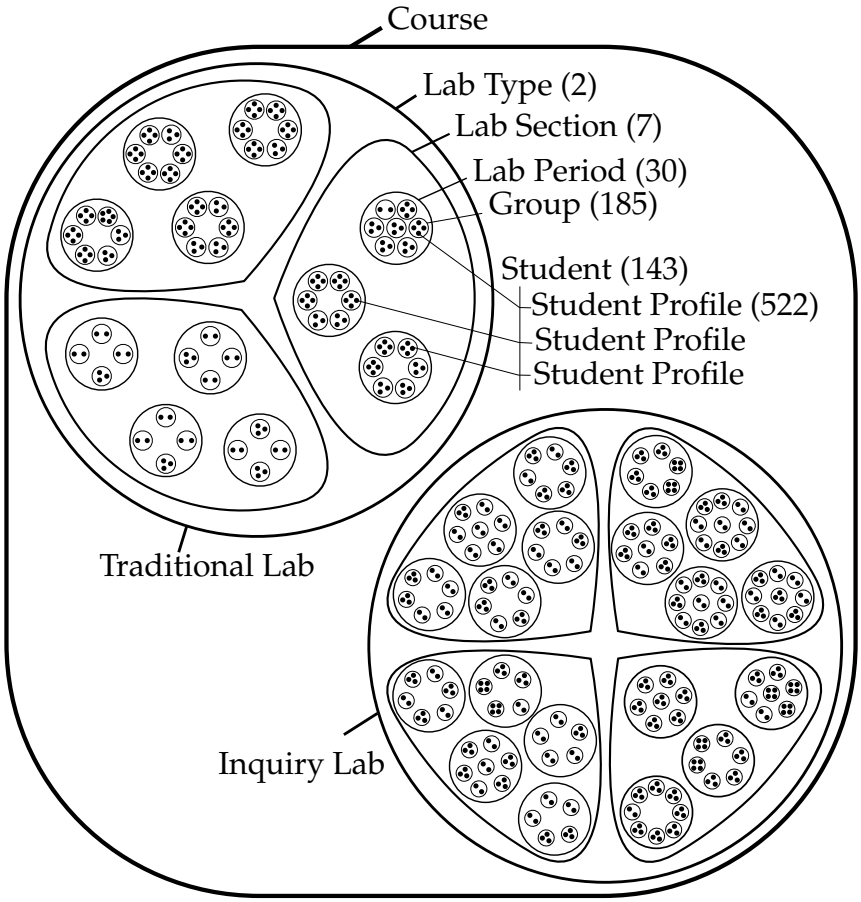


Figure 5.2: **Set diagram of the lab breakdown** for this study. We observed 143 students in multiple lab periods, generating 522 student profiles. All students were in the same physics course, but were in one of two lab types (*traditional* or *inquiry*). Students registered for a particular lab section (three sections were of the traditional lab type, and four were of the inquiry lab type), and worked in different groups during the semester. Each student generated multiple profiles, occurring in different lab periods with different groups (but for fixed section and lab type).

Table 5.3: **Demographic breakdown of student profiles** measured in this study. Errors were computed using standard error for population fractions, shown in Eq. (G.1). In all, 143 students were observed across multiple lab periods, resulting in 522 unique student profiles.

	Traditional Labs		**Inquiry Labs**	
	N	%	*N*	%
Women	34	18 ± 3	87	26 ± 2
Men	152	81 ± 3	246	74 ± 2
Undisclosed	2	1 ± 1	1	0.3 ± 0.3

quencies (broken down by gender or lab type) or a regression (linear modeling in some way), such standard methods rely on the assumption of Gaussian distributions for the underlying data. In this study, the distribution of code frequencies are highly skewed, with most students engaging in a particular activity infrequently or not at all and some students engaging in an activity a lot. Figure 5.3 shows box plots of the raw data, illustrating the non-Gaussian features of the data. For this reason, we instead perform a cluster analysis. Such an analysis can be used to characterize *behaviour types* instead of *average behaviour*. Clustering can account for non-linearities missed in common regression analyses (such as capturing *dominant* behaviour as opposed to *average* behaviour) and has been used in similar studies of this type to provide fruitful results [32]. By performing a demographic analysis on the student groupings (*i.e.* clusters) we can quantitatively characterize coarse gendered behaviour. A full description of the clustering method is described in Section 5.2.5.

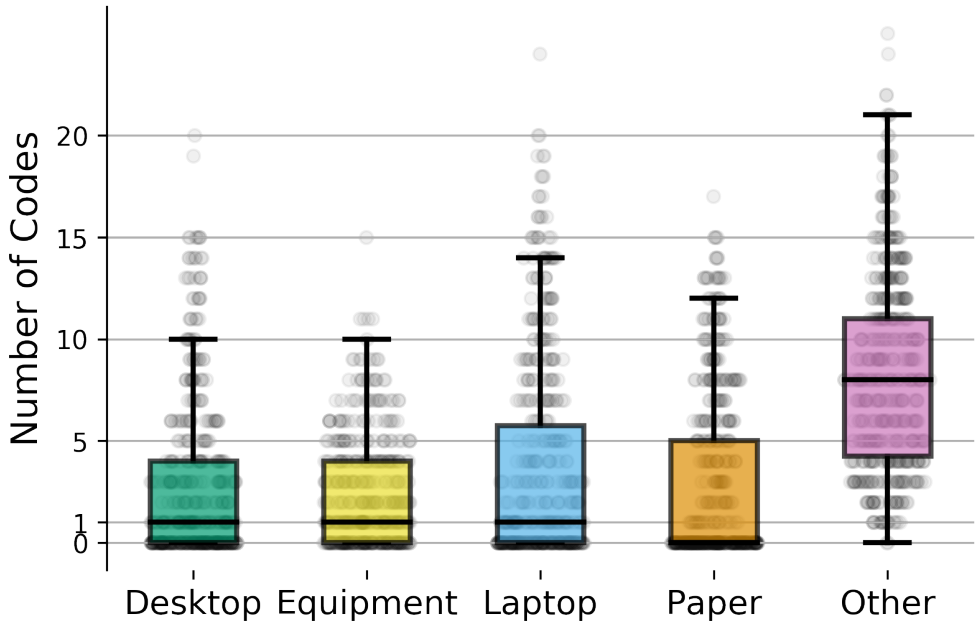


Figure 5.3: **Box plots of raw data** revealing the highly non-Gaussian nature of the code distributions. Each faded point is the accumulated codes for a student in a lab period, and so darker regions represent more total codes of that value (with the darkest regions near zero). Note that the median for all codes except Other is less than or equal to one, reflecting the fact that over half of students were observed engaging in that behavior once or less than once. This, combined with the fact that there are a large number of outliers, is an indication that students either engage in a particular activity a lot or not at all.

5.2.3 Describing Detailed Student Behaviour

We used video recording of single-groups during full lab periods to better describe more detailed student behaviour. Because the striking genderred differences in behaviour occur in the inquiry lab, and we observed no measurable difference in the traditional labs, videos focussed on groups in the inquiry labs, and looked mostly at mixed-gender groups. In all, ten videos were coded, decomposing 23 profiles from 17 students (five students appeared in more than one video). All 23 decomposed profiles from all 10 videos with demographic

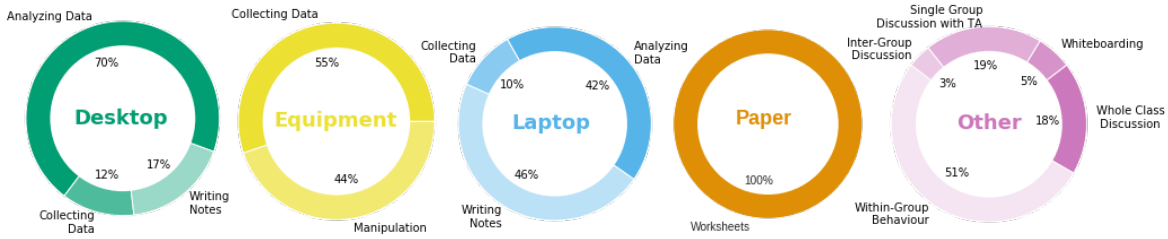
information are shown in Appendix J.

BORIS software was used to code videos [55], specifically the fraction of time students engaged in different behaviours. The five codes in Table 5.2 were further broken down by what a student was doing (*e.g.* analysing data) while engaged in coarse behaviour (*e.g.* using the Desktop). A breakdown of the codes are shown in Fig. 5.4.

The Paper code was used to predominantly describe students filling out paper worksheets in the traditional labs, and so it was not further decomposed. Both the Desktop and Laptop codes were used to describe students analysing data, collecting data, or writing lab notes, and so both of these codes were broken down in this way. However, when collecting data, the Desktop was often connected directly to equipment whereas gathering data on a laptop was purely represented by students manually entering data into their electronic notebook or analysis software. Students handling equipment were primarily doing so to either collect data or manipulate the setup in some way (setup, cleanup, calibration, playing) and so the Equipment code can be further decomposed into these two tasks. In this way, the Desktop, Equipment, Laptop, and Paper codes were explicitly decomposed.

To better describe student behaviour while coded as Other, we introduced four new state codes. These were used to describe significant events in lab, and are elaborated in Table 5.4. By overlapping the event codes with Other, we broke down the Other code and provide a more qualitative picture of classroom activities, such as engaging in whole-class discussions, using whiteboards to sketch out ideas and concepts, single group discussions with the TA or UTA, or engaging in inter-group discussions with neighbouring groups.

(a) Breakdown of Codes



(b) Sample Profile from Time-Coded Video

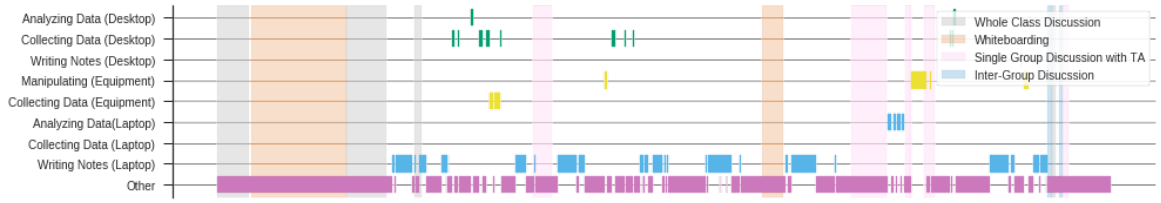


Figure 5.4: **Breakdown of codes** by decomposing coarse behaviour (*e.g.* "handling laptop") into more fine-grained behaviour (*e.g.* "analyzing data"). Ten videos were coded, resulting in 23 decomposed profiles from 17 different students (five students appeared in more than one video). (a) A breakdown of each code, showing the fraction of time students engaged in a particular task while coded as a particular behaviour. Three of the five codes (Desktop, Equipment and Laptop) were directly decomposed into sub-codes while analyzing videos, as shown in (b) illustrating a sample coded time-series. Four additional *group states* were coded in the videos, representing large group behaviour (discussing with a TA or UTA, conversing with other groups, whole class discussions and announcements, and using a whiteboard). We decomposed the Other code by overlapping it with these larger group states. The Paper code was purely represented by students filling out paper worksheets in the traditional labs.

Table 5.4: **Event codes used in video observations**. These codes described significant events in the lab, and were used to decompose the more coarse-grained Other code. A sample time series illustrating a coded video is shown in Fig. 5.4(b)

Code	**Description**
Whole Class Discussion	The TA or UTA makes an announcement to the class, or holds a whole class discussion.
Whiteboarding	Students perform invention activities in the lab, and use a white board to sketch out ideas and concepts.
Single Group Discussion with the TA	TA or UTA engages in a discussion with the group (but not as part of a whole class discussion).
Inter-Group Discussion	Groups compare results or discuss among each other (not as part of a whole class discussion).

To validate this method, two observers coded the same video as a means of testing the inter-rater reliability. Cohen's kappa was used as a measure of agreement between two observers, with a value of 0.61-0.80 representing substantial agreement. Two observers coded the same video, and obtained a Cohen's kappa value of 0.79, indicating substantial agreement between the two. As a result, only one researcher coded the subsequent videos.

Video analysis was also used to better understand task allocation. Point-events were identified when one student explicitly instructed another to perform a task. We breakdown the criteria for inclusion and exclusion in the following way:

- *Criteria for Inclusion:* A student needs to be addressing another, and explicitly directing them in some way, such as by saying "you should do X".
- *Criteria for Exclusion:* Suggesting a task should be done that a student assumes without being asked is not included. Examples of such events are characterized by statements such as "We should do X.", "I think we should focus on X.", "Does someone want to work on X?". Additionally, a student asking another for help performing a task is excluded (such as asking another student how to sum a row in a spreadsheet, and the student telling them how).

5.2.4 Rescaling Student Profiles

To perform a cluster analysis on multidimensional data, the scales for each measure must be the same. In this study, there were two effects present which caused differences in scales (that were accounted for).

First, the amount of coded time for each student was highly variable, ranging from less than 45 minutes to over 175 minutes (a full histogram of students' time in lab is shown in Fig. G.3). To account for this effect, we normalized each student profile by the total number of observed codes for that student. In this way, each measure represents the fraction of time spend on a particular task.

Second, there is the inherent differences in the five measures. For instance, from Fig. 5.3, we can see that the distribution for Other is more spread out than for Equipment. To account for this, each measure was grand mean scaled so that, averaged over all students, each measure had mean 0 and standard deviation of 1. In doing so, each measure becomes a Z-score [32, 133]. Thus, each

student's z-score tells us whether the time they spent on a particular activity was above or below average as compared to other students. Moreover, the Euclidean distance between two profiles has a statistical interpretation in this z-score format: it measures the dissimilarity of two student profiles in units of standard deviations [32].

To illustrate this two-step rescaling, a sample profile is shown in Appendix G.3. Importantly, we rescale all student profiles and turn them into z-scores in order to cluster them simultaneously, rather than consider sub-groups (such as lab type or gender). We do this for ease of comparison, so that we can contrast the distribution of student profiles in the inquiry and traditional labs and men's and women's profiles after clustering (rather than imposing a divide before clustering). To investigate the impact of sub-dividing groups prior to rescaling and clustering, we present the results of such an analysis in Appendix H, which shows minimal impact on the distribution of profiles and no impact on the number or description of the clusters themselves.

5.2.5 Cluster Analysis

We performed a standard k-means clustering on the student profiles. K-means is an iterative algorithm, where the researcher specifies the number of clusters. The algorithm clusters and then re-clusters the data in an iterative manner until the sum of the square of the distances from all points to their respective cluster's center is minimized and no point changes cluster between iterations [67].

Note that not all data can be meaningfully clustered. For example, even if all data form a structure-less blob, a researcher can still input two or more

clusters and the algorithm will converge to a solution. Therefore, in order to determine (1) if the data are clusterable, and (2) if so, what the optimal number of clusters is, we used the elbow method [141]. We plotted the average squared distance from each point to the center of its assigned cluster, as a function of the number of clusters, and compared the results to 10,000 randomly generated student profiles. We used enough random data to generate a smooth function and ensure that the comparison is not hindered by statistical fluctuations. The results of the elbow plot are shown in Fig. 5.5. The plot for our collected data was substantially below random, indicating that the data are clusterable. There is a distinct kink in the plot for five clusters, indicating that the optimal number of clusters is five.

From the elbow plot in Fig. 5.5, we can see that the five optimal clusters account for 70% of the variance in the data (73% of Desktop use, 60% of Equipment use, 78% of Laptop use, and 59% of Other activities), well above the 50% threshold used for a study of this type [32, 133]. We provide a 2D visualization[2] of the set of student profiles using t-SNE [98], with profiles coloured by assigned cluster, in Fig. 5.6.

Because each student had multiple profiles, arising from several lab periods over the course of a semester, we investigated whether or not it is possible to further collapse the profiles to determine "semester-long" behaviours. We did this by analysing whether or not individual students' profiles appear in multiple clusters over the course of a semester. In the traditional labs, $87 \pm 4\%$ of students have profiles appearing in more than one cluster. Similarly, $86 \pm 4\%$ of students in the inquiry lab appear in more than one cluster. This effect is

[2]Data and analysis for this work was done prior to the publishing of work done in Chapter 4, and so we used the established visualization methods to visualize the data. For a quantitative visualization of the data using InPCA, see Appendix I.

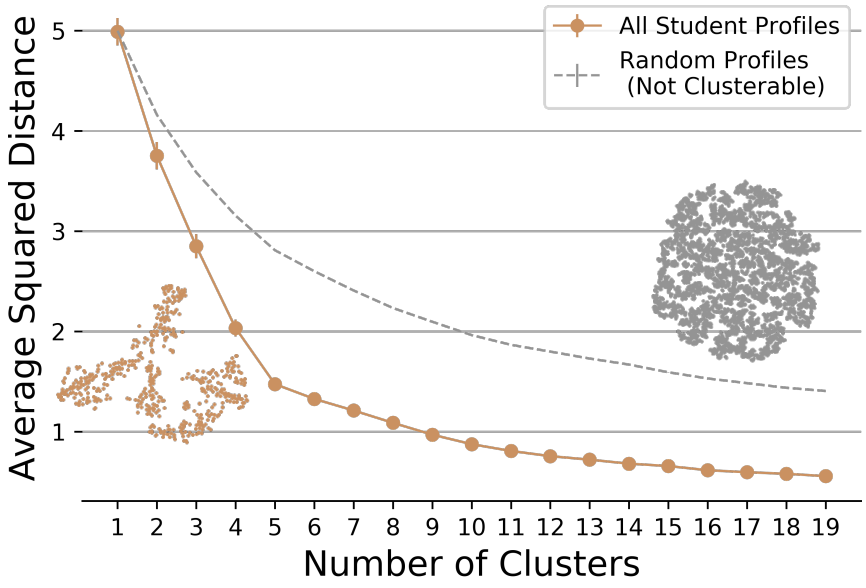


Figure 5.5: **Elbow plot** used to determine the optimal number of clusters for the data. The average squared distance from each point to the center of its assigned cluster is plotted as a function of the number of clusters. There is a kink at five, indicating that the optimal number of clusters for the data is five. Our results were compared against 10,000 randomly generated student profiles. Note that the elbow is well below random, a sign that the data can be clustered. Superimposed on the graph is a two-dimensional visualization of the data and random points for qualitative comparison. The data show structure (brown points in lower left), whereas the random points form a blob (grey points in center right).

highlighted in Fig. 5.6 by connecting profiles from individual students, with grey lines representing between-cluster matching and colored lines indicating within-cluster matching. Because so many students have profiles appearing in multiple clusters, the weekly variation in an individual's profile is too great to further collapse (for numerous reasons, such as variability in lab content and students changing lab partners). Figure 5.6 is a two-dimensional representation of a five-dimensional space, and so is used primarily for qualitative illustration: cluster composition is quantitatively analysed in Section 5.3.

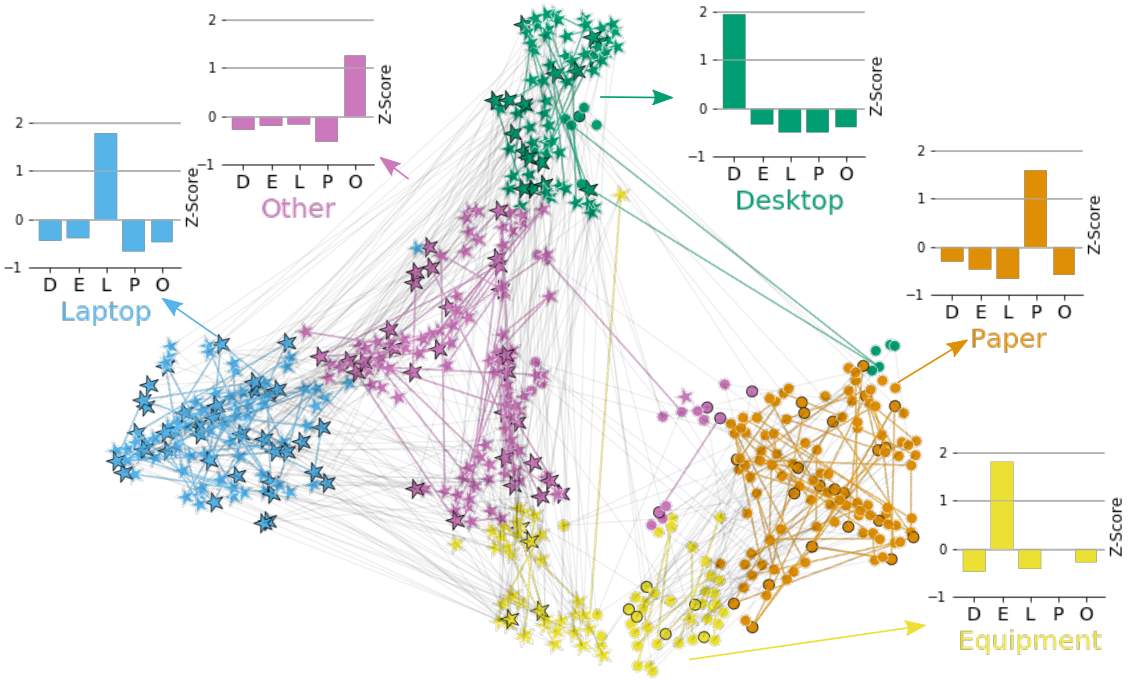


Figure 5.6: **Two-dimensional visualization of clusters** and their centers. Each point represents a unique student profile, with profiles from the same student connected by a faded line (grey representing between-cluster matching and colored lines representing within-cluster matching). Circles represent students in the traditional labs and stars in the inquiry labs, and black edges indicate women's profiles. All points in the Laptop cluster are stars, whereas all points in the Paper cluster are circles, a reflection of the pedagogical differences in the labs (students in the traditional labs were filling out paper worksheets, whereas in the inquiry labs were filling out electronic notebooks). Note that there are far more black-edged points in the Laptop cluster than in the Equipment cluster, hinting at gendered differences in cluster compositions. This effect is explored in detail in Fig. 5.8. Clusters are characterized by their centers, and here the centers of the five clusters are given by large Z-scores for each of our codes.

Clusters from k-means are characterized by their centers. Here, the centers of the five clusters (shown in Fig. 5.6) matched the five codes used in this study and so we labelled the clusters accordingly. Therefore, the clusters characterize "high users" of a particular measure, *i.e.* a student in the Equipment cluster spends a larger fraction of their time on the equipment than the average student. Note that this description fits with the raw data, shown in Fig. 5.3, which illustrates that the majority of students engage in a particular task either frequently or very rarely. Note that this is not just a feature of the data (see blob of random data in Fig. 5.5) but is a feature of the students' behaviour and the validity of the coding scheme (*i.e.* at approximating student behaviours).

5.3 Results

Based on the pedagogical differences between the two lab types, one can use our theoretical framework to predict that students in the inquiry labs (who worked collaboratively within a group) would divide tasks among group members far more than students in the traditional labs. To address this prediction, we analyzed the cluster assignment of group members to see if members predominantly fell into the same or different clusters. In the traditional labs, $43 \pm 6\%$ of groups had all members in the same cluster (predominantly the paper cluster) whereas only $14 \pm 2\%$ of groups in the inquiry lab had all members in the same cluster. This is illustrated in greater detail in Fig. 5.7.

We performed a quantitative analysis of the cluster compositions by considering the cluster distributions over lab type, gender, and group composition. In all cases, when comparing cluster compositions, we used a chi-squared test

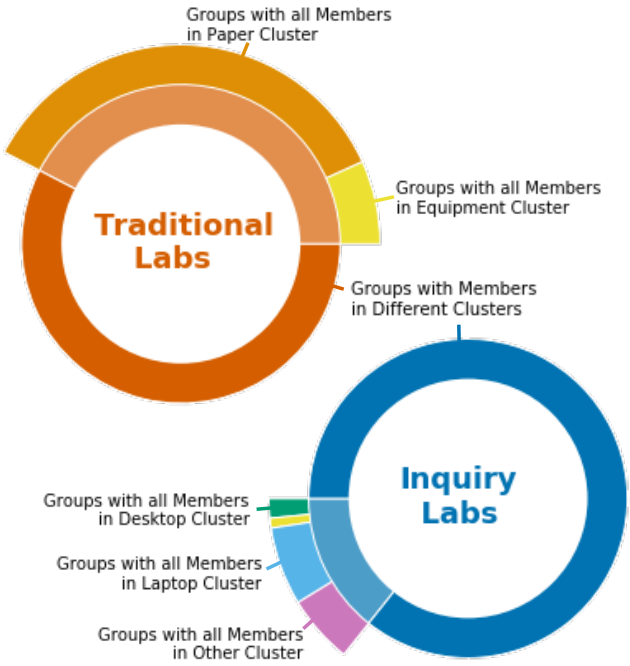


Figure 5.7: **Fraction of groups** with members in identical clusters (dark ring) and different clusters (light ring) illustrating task division in the different labs. Almost half of groups in the traditional labs had all members in the same cluster (primarily Paper cluster), whereas the majority of groups in the inquiry labs had members in multiple clusters indicating an increase in task division.

of frequencies on the contingency tables of the raw counts (description and examples provided in Section G.2). We first looked at the cluster compositions as broken down by lab type, shown in Fig. 5.8(a). Students in the traditional labs spent a large portion of their time filling out paper worksheets, as revealed by the fact that 60% of their profiles were in the Paper cluster. In contrast, students in the inquiry labs engaged in a wider range of activities, as reflected in a more uniform distribution across clusters. This further supports the argument from the theoretical framework that more roles were available to students in the inquiry labs.

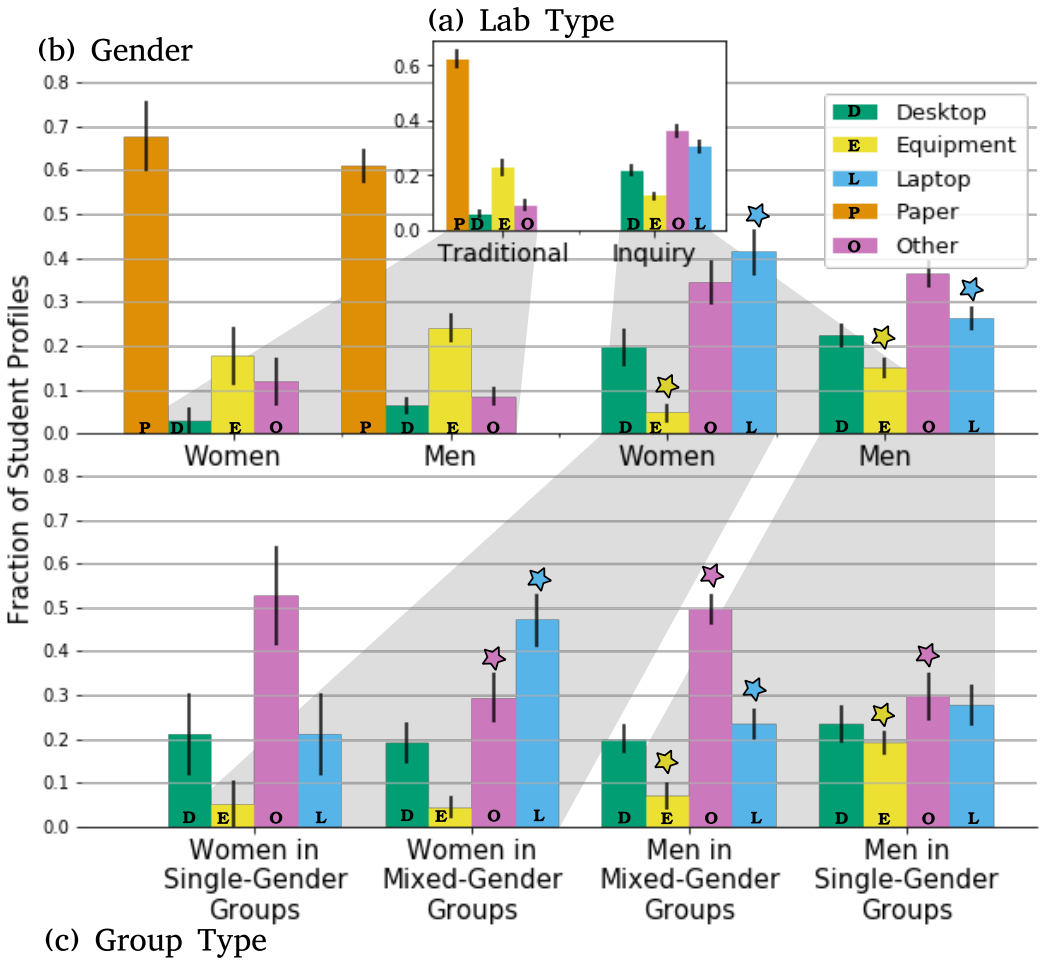


Figure 5.8: **Cluster compositions** for each of the five clusters, broken down both by lab type, gender and group composition. In all plots, y-axis represents fraction of student profiles and errors are calculated using the standard error on the fraction of a population shown in Eq. (G.1). (a) Cluster distributions broken down by lab type. (b) Clusters were further broken down by gender. Upon inspection, we see that there are disproportionately more women in the Laptop cluster than men (blue stars), and disproportionately more men than women in the Equipment cluster (yellow stars). (c) Cluster distributions were further broken down in the inquiry lab by group type (men and women in mixed-gender groups and single-gender groups). Upon inspection, we see that the Laptop difference remained (blue stars), while a difference emerged in Other (purple starts). Furthermore, far more men are high-equipment users when in single-gender groups (yellow stars). Due to insufficient statistics, no comparison can be made with women in single-gender groups ($p>0.17$ in all cases).

We then decomposed the clusters by gender and lab type, as shown in Fig. 5.8(b). Through a chi-squared test of frequencies, we determined that there is no statistically significant difference in the distributions for men and women in the traditional laps (p=0.65) but that there is one for the inquiry labs (p=0.01). Upon inspection of the resulting distribution, we noted that there are disproportionately more women in the Laptop cluster than men and disproportionately more men in the Equipment cluster than women.

To investigate the cause of the equipment and laptop differences in the inquiry labs, and to consider the impact of group demographics (whether or not a student working only with same-gendered peers affects their behaviour), we further decomposed the clusters by group type in the inquiry labs (men and women in mixed-gender or single-gender groups) in Fig. 5.8(c). A statistically significant difference persists in the inquiry labs between men and women in mixed-gender groups (p=0.02), with women being high-laptop users far more than men but now with men engaging in Other activities far more than women. Furthermore, the initial difference in equipment usage in the inquiry labs appears to be a result of a difference in men's behaviour when in mixed- versus single-gender groups (p=0.007), namely that men are far more likely to be high equipment users when working with other men and that all group members in mixed-gender groups are unlikely to be high equipment users. Unfortunately, due to insufficient statistics, no comparison can be made with women in single-gender groups (p>0.17 in all cases). Note that the fraction of students in the inquiry labs in the Desktop cluster remained effectively constant ($\sim$ 20%) for all cluster breakdowns shown in Fig. 5.8.

From the cluster analysis, we noticed a significant difference in behavior

between men and women in mixed-gender groups, and for men who worked solely with other men as compared to when they worked with women. Specifically, women were high-laptop users more often than men. Men were substantially more likely to be high equipment users when working only with other men, and were more often engaged in other activities when working with at least one woman. To better characterize these behavioural differences, we used single-group video analysis of students in mixed-gender groups in the inquiry labs, with the data shown in Appendix J. The difference in laptop usage between men and women in mixed-gender groups appeared to be the results of students engaging in data analysis, with women having spent about twice as much time as men analyzing data on their laptops ($14 \pm 7\%$ of their time for women as compared to $6 \pm 3\%$ for men). The difference in the Other cluster was solely due to differences in within-group behaviours such as talking, observing, or interacting with group members. Men in mixed-gender groups spent a third of their time ($30 \pm 4\%$) engaging in these activities, as compared to women in mixed-gender groups who spent about a quarter of their time ($26 \pm 5\%$).

To better understand the source of these behavioural differences, we used video analysis of single-groups to identify instances of explicit task allocation. We identified very few such events occurring in a lab (two or three times per lab period per group). Furthermore, all such events were best described as specific direction of a student already engaged in a task. A typical example of such an occurrence is the following: a group of students were calculating the period of a simple pendulum, and while one student was analysing the data another instructed them to calculate the period by dividing the time of multiple oscillations by the number of oscillations by saying "you should probably divide all those by five."

From the single-group video analysis, we conclude that the significant difference in behaviour is not the result of overt, explicit allocation. Rather, it is the result of the accumulation of subtle interactions at the individual level that become manifest at the classroom level.

5.4 Discussion

Do men and women behave differently in physics labs? How are behaviours related to participation and the *roles* students take on in their group of peers, and in what ways do these roles contribute to student experiences and persistence in the field? To address these questions, we quantified the behaviour of 143 students in multiple lab periods in two pedagogically different sections of the same lab course. To probe the set of available roles for students in these sections (elaborated in our theoretical framework in Fig. 5.1), we quantified their behaviour through five coarse metrics and performed a cluster analysis, using the resulting clusters as a proxy for these roles. The five resulting clusters represent students engaging in a particular task substantially more than average, and so characterize high equipment users, high desktop users, high laptop users, high paper users, or engaging in other activities far more than average (such as talking or observing their peers).

The traditional labs, designed to reinforce concepts introduced in lecture, were highly guided and structured. Students had very little room for active decision making about the experiment, as they primarily followed detailed instructions. Furthermore, although they worked in groups, each student was responsible for completing their own individual worksheet. As a result, the set

of available roles is both highly constrained and manifestly similar for all students.

In contrast, the inquiry labs were designed to emphasize the process of experimentation and thus students were given more agency for active decision making about the experiment. In particular, Fig. 5.8(a) illustrates the more even distribution across the different measures for the inquiry labs (whereas the majority of students in the traditional labs are high paper users) and Fig. 5.7 provides strong evidence for within-group task allocation as over 85% of groups have members spread across multiple clusters. As a result, the set of available roles is much greater in the inquiry labs and students are able to assume very different roles from one another.

Women in the inquiry labs were more likely to be heavy laptop users than men when working in mixed-gender groups. Specifically, women spent about twice as much time as their male group members analysing data on their laptops. While we obtained insufficient statistics to make definitive claims about women who worked in single-gender groups, we can see qualitative differences between women who worked with other women as compared to when they worked with men in Fig. 5.8(c) (women were probably less likely to be heavy laptop users and more likely to engage in other within-group activities when working with other women).

Men in the inquiry labs behaved very differently when working only with other men as compared to when they worked with women. Specifically, when men were in single-gender groups, they were far more likely to be high equipment users (than either their female peers or male peers in mixed-gender groups). This difference is an indication of the different roles men assume de-

pending on social context. When in groups of men, there were different social dynamics as compared to when in groups with women, thereby changing the set of available roles (and thus observed behaviours). We speculate that this is the increased use of equipment in men-only groups was the results of "playfulness" [69], with men more likely to play around with the equipment, and is the focus of further study. Furthermore, when men were in groups with women, they were considerably more likely to engage in other within group activities (such as talking or observing) than both their female group members and male peers in single-gender groups.

To explore the cause of the gendered division of roles, we used single-group video analysis to determine task allocation. These allocations were not overt, *i.e.* students were not directly assigning each other tasks through explicit instruction. The only instruction from one student to another was in the form of quick, directed statements about an existing task. Students must have been either predominantly self-assigning tasks within groups, "falling into" roles, or directing each other through *positioning* (subtle verbal and non-verbal social cues [39, 16]). Exploring these other mechanisms of task allocations is the focus of future study, to better understand how roles become gendered.

Substantial gendered behaviours occurred in the inquiry labs with regards to equipment manipulation, laptop usage, data analysis, and within-group interactions, however we did not measure these same features in our traditional labs. Such vastly different results in the same study better contextualize the conflicting results from previous studies, which has shown mixed results with regards to gendered action in first-year physics labs [38, 82] such as men using desktop computers more than women [40], and how management of equipment appara-

tus is heavily impacted by gender in mixed-group pairs [74]. By understanding how the lab structure impacts the set of roles available to students as well the ways in which these are roles gendered, researchers can better unify seemly conflicting results (*i.e.* what appears as "masculine" or "feminine" behaviour in one context can change in another). For example, the conflicting results with computer usage between this study and [40] can be better understood in relation to data collection versus analysis, and how the corresponding roles of "data collector" and "data analyst" are viewed as "masculine" or "feminine" roles.

A more nuanced understanding of behaviours and roles and how they are assumed by students can better inform instructors and physics departments wishing to implement institutional changes. The vastly different roles students take on when in the same physics program greatly influences their experience, identity formation as physicists and future prospects, ultimately impacting persistence and representation in the field. As the pedagogical structure of labs are changed and improved, we argue there is an equal need to structure group functions (equity in task and role allocation) as there is to design the lab procedure itself. If not, we risk inadvertently reinforcing gendered roles in the labs.

CHAPTER 6

CONCLUSIONS

In the era of big data and emergent phenomena in complex systems, researchers need advanced statistical and analytic tools. How can we faithfully reduce complex, high-dimensional data sets to reveal underlying structure within them? How can we systematically reduce the complexity of nonlinear models so as to preserve predictive power? What properties of models and data make this possible (and, conversely, what properties would make this task impossible)?

Information geometry, described in Chapter 2, can be such a tool. We used it to generate a new manifold learning method called InPCA in Chapter 4, which can faithfully reveal underlying structure in complex, nonlinear models with very high dimensional, probabilistic predictions. We combined information geometry with approximation theory to describe and quantify seemingly universal patterns in model behaviors, namely their "sloppy" properties (*i.e.* a hierarchical dependence on certain parameter combinations), by deriving bounds on model predictions through underlying model smoothness in Chapter 3. *Smoothness* in model predictions necessarily lends itself to a hierarchical structure in model predictions, and therefore makes it possible to reduce certain models.

There is a disparity in representation in physics, with men dominating in rank and number. Motivated by poststructural gender theory that extends beyond the classic "deficit model" of performance gaps, in Chapter 5, we quantified the complex behavior patterns of individuals in an introductory physics labs, and revealed the gendered division of roles that occurred in them. Our results indicate a pressing need to include structured equity in reformed labs

which aim to emphasis the process of experimentation and student agency. Numerous resources are currently used to redesign labs, in particular to align them with the new AAPT lab guidelines [86]. However, if we don't pay equal attention to the group dynamics and roles students assume within these spaces, we risk inadvertently reinforcing gendered roles and division of labour.

APPENDIX A

TEDIOUS CALCULATIONS AND DERIVATIONS

In this appendix, we provide detailed calculations for the FIM connections in Chapter 2, the proofs for bounds in Chapter 3 (bounds on non-analytic models and the hyperellipsoid lengths for monomial expansions), and derive the intensive-distance for least-squares models and the intensive cross-covariance in Chapter 4.

A.1 Fisher Information Matrix for Least-Squares Models

Here, we show that the FIM for least squares models matches the metric on the model manifold from Eq. (2.9). We do so by plugging in the likelihood function from Eq. (2.13) into the definition of the FIM from Eq. (2.10). To do so, we first find derivatives of the log-likelihood:

$$\partial_\alpha \log \mathcal{L}(\boldsymbol{x} \mid \theta) = -\sum_i \frac{y_{\boldsymbol{\theta}}(t_i) - x_i}{\sigma_i^2} \partial_\alpha y_{\boldsymbol{\theta}}(t_i) \tag{A.1}$$

where we use the convention that $\partial_\alpha := \frac{\partial}{\partial \theta^\alpha}$. The FIM is now given as

$$\mathcal{I}_{\alpha\beta} = \int d\boldsymbol{x} \left(\sum_i \frac{y_{\boldsymbol{\theta}}(t_i) - x_i}{\sigma_i^2} \partial_\alpha y_{\boldsymbol{\theta}}(t_i) \right) \left(\sum_j \frac{y_{\boldsymbol{\theta}}(t_j) - x_j}{\sigma_j^2} \partial_\beta y_{\boldsymbol{\theta}}(t_j) \right) \prod_k \frac{1}{\sqrt{2\pi\sigma_k^2}} e^{\left(-\frac{(y_{\boldsymbol{\theta}}(t_k) - x_k)^2}{2\sigma_k^2} \right)} \tag{A.2}$$

When $i \neq j$, the integral over x_i and x_j will both yield zero because of the exponential term. When $i = j$, the integral over all x_k for $k \neq i, j$ will be one because of the normalization of the likelihood function. The above expression can therefore

be simplified to

$$\mathcal{I}_{\alpha\beta} = \sum_i \partial_\alpha y_{\boldsymbol{\theta}}(t_i)\partial_\beta y_{\boldsymbol{\theta}}(t_i) \int dx_i \frac{(y_{\boldsymbol{\theta}}(t_i)-x_i)^2}{\sigma_i^4} \frac{1}{\sqrt{2\pi\sigma_i^2}} \exp\left(-\frac{(y_{\boldsymbol{\theta}}(t_i)-x_i)^2}{2\sigma_i^2}\right) \tag{A.3}$$

$$= \sum_i \frac{1}{\sigma_i^2}\partial_\alpha y_{\boldsymbol{\theta}}(t_i)\partial_\beta y_{\boldsymbol{\theta}}(t_i). \tag{A.4}$$

The FIM for least squares models can therefore be expressed using the Jacobian of the model, and matches the metric on the model manifold from Eq. (2.9).

A.2 Equivalent Representations of the Fisher Information Matrix

We show how the two forms of the FIM in Eq. (2.10) are equivalent. We do so by turning Eq. (2.10)(ii) into Eq. (2.10)(i):

$$\mathcal{I}_{\alpha\beta} = -\sum_{\boldsymbol{x}} \frac{\partial^2 \log \mathcal{L}(\boldsymbol{x}\mid\boldsymbol{\theta})}{\partial\theta^\alpha\partial\theta^\beta}\mathcal{L}(\boldsymbol{x}\mid\boldsymbol{\theta}) \tag{A.5}$$

$$= -\sum_{\boldsymbol{s}} \frac{\partial}{\partial\theta^\alpha}\left[\frac{1}{\mathcal{L}(\boldsymbol{x}\mid\boldsymbol{\theta})}\frac{\partial\mathcal{L}(\boldsymbol{x}\mid\boldsymbol{\theta})}{\partial\theta^\beta}\right]\mathcal{L}(\boldsymbol{x}\mid\boldsymbol{\theta}) \tag{A.6}$$

$$= -\sum_{\boldsymbol{x}} \frac{-1}{\mathcal{L}(\boldsymbol{x}\mid\boldsymbol{\theta})^2}\frac{\partial\mathcal{L}(\boldsymbol{x}\mid\boldsymbol{\theta})}{\partial\theta^\alpha}\frac{\partial\mathcal{L}(\boldsymbol{x}\mid\boldsymbol{\theta})}{\partial\theta^\beta}\mathcal{L}(\boldsymbol{x}\mid\boldsymbol{\theta}) - \sum_{\boldsymbol{x}}\frac{\partial^2\mathcal{L}(\boldsymbol{x}\mid\boldsymbol{\theta})}{\partial\theta^\alpha\partial\theta^\beta} \tag{A.7}$$

$$= \sum_{\boldsymbol{x}} \frac{\partial\log\mathcal{L}(\boldsymbol{x}\mid\boldsymbol{\theta})}{\partial\theta^\alpha}\frac{\partial\log\mathcal{L}(\boldsymbol{x}\mid\boldsymbol{\theta})}{\partial\theta^\beta}\mathcal{L}(\boldsymbol{x}\mid\boldsymbol{\theta}) - \frac{\partial^2}{\partial\theta^\alpha\partial\theta^\beta}\cancelto{0}{\sum_{\boldsymbol{x}}\mathcal{L}(\boldsymbol{x}\mid\boldsymbol{\theta})} \tag{A.8}$$

The last term cancels due to the normalization of $\mathcal{L}$ (the distribution, if normalized, integrates to one regardless of model parameters). We therefore obtain the form of the FIM from Eq. (2.10)(i).

A.3 Fisher Information Matrix and the Bhattacharyya Distance

Here, we show that the metric for the Bhattacharyya distance from Eq. (2.19) is proportional to the FIM in Eq. (2.10). We consider parameter-dependent distributions $\mathcal{L}(\boldsymbol{x} \mid \boldsymbol{\theta})$ that we express as $\mathcal{L}_{\boldsymbol{\theta}}$ to simplify the expressions, and compute the distance between parameters $\boldsymbol{\theta}$ and $\boldsymbol{\theta} + \delta\boldsymbol{\theta}$:

$$D_B(\boldsymbol{\theta} \,||\, \boldsymbol{\theta} + \delta\boldsymbol{\theta}) = D_B(\boldsymbol{\theta} \,||\, \boldsymbol{\theta}) + \partial_\alpha D_B(\boldsymbol{\theta} \,||\, \boldsymbol{\theta})\delta\theta^\alpha + \partial_\alpha\partial_\beta D_B(\boldsymbol{\theta} \,||\, \boldsymbol{\theta})\delta\theta^\alpha\delta\theta^\beta + O(\delta\boldsymbol{\theta}^3). \quad \text{(A.9)}$$

We compute the different terms in this expansion. The zeroth order vanishes,

$$D_B(\boldsymbol{\theta} \,||\, \boldsymbol{\theta}) = -\log\left(\int d\boldsymbol{x}\, \sqrt{\mathcal{L}_{\boldsymbol{\theta}}}\, \sqrt{\mathcal{L}_{\boldsymbol{\theta}}}\right) = 0. \quad \text{(A.10)}$$

We see that the first order term also vanishes:

$$\partial_\alpha D_B(\boldsymbol{\theta} \,||\, \boldsymbol{\theta}) = -\frac{1}{\int d\boldsymbol{x}\, \sqrt{\mathcal{L}_{\boldsymbol{\theta}}}\, \sqrt{\mathcal{L}_{\boldsymbol{\theta}}}} \int dx \frac{\sqrt{\mathcal{L}_{\boldsymbol{\theta}}}}{2\sqrt{\mathcal{L}_{\boldsymbol{\theta}}}} \frac{\partial \mathcal{L}_{\boldsymbol{\theta}}}{\partial\theta^\alpha} = -\frac{\partial 1}{\partial\theta^\alpha} = 0 \quad \text{(A.11)}$$

The second order term is given by:

$$\partial_\alpha\partial_\beta D_B(\boldsymbol{\theta} \,||\, \boldsymbol{\theta}) = \frac{1}{\left(\int d\boldsymbol{x}\, \sqrt{\mathcal{L}_{\boldsymbol{\theta}}}\, \sqrt{\mathcal{L}_{\boldsymbol{\theta}}}\right)^2} \left(\int dx \frac{\sqrt{\mathcal{L}_{\boldsymbol{\theta}}}}{2\sqrt{\mathcal{L}_{\boldsymbol{\theta}}}} \frac{\partial \mathcal{L}_{\boldsymbol{\theta}}}{\partial\theta^\alpha}\right)\left(\int dx \frac{\sqrt{\mathcal{L}_{\boldsymbol{\theta}}}}{2\sqrt{\mathcal{L}_{\boldsymbol{\theta}}}} \frac{\partial \mathcal{L}_{\boldsymbol{\theta}}}{\partial\theta^\beta}\right) \quad \text{(A.12)}$$

$$+ \frac{1}{\int d\boldsymbol{x}\, \sqrt{\mathcal{L}_{\boldsymbol{\theta}}}\, \sqrt{\mathcal{L}_{\boldsymbol{\theta}}}} \int dx \frac{\sqrt{\mathcal{L}_{\boldsymbol{\theta}}}}{4\mathcal{L}_{\boldsymbol{\theta}}^{3/2}} \frac{\partial \mathcal{L}_{\boldsymbol{\theta}}}{\partial\theta^\alpha} \frac{\partial \mathcal{L}_{\boldsymbol{\theta}}}{\partial\theta^\beta} \quad \text{(A.13)}$$

$$- \frac{1}{\int d\boldsymbol{x}\, \sqrt{\mathcal{L}_{\boldsymbol{\theta}}}\, \sqrt{\mathcal{L}_{\boldsymbol{\theta}}}} \int dx \frac{\sqrt{\mathcal{L}_{\boldsymbol{\theta}}}}{2\sqrt{\mathcal{L}_{\boldsymbol{\theta}}}} \frac{\partial^2 \mathcal{L}_{\boldsymbol{\theta}}}{\partial\theta^\alpha\partial\theta^\beta}. \quad \text{(A.14)}$$

Because the distributions are normalized, the only surviving term is the one on Line (A.13). Using the fact that $\partial_\alpha \log \mathcal{L} = \frac{1}{\mathcal{L}}\partial_\alpha \mathcal{L}$, we can therefore express Eq. (A.9) as

$$D_B(\boldsymbol{\theta} \,||\, \boldsymbol{\theta} + \delta\boldsymbol{\theta}) = \frac{1}{4} \underbrace{\int d\boldsymbol{x} \frac{\partial \log \mathcal{L}_{\boldsymbol{\theta}}}{\partial\theta^\alpha} \frac{\partial \log \mathcal{L}_{\boldsymbol{\theta}}}{\partial\theta^\beta} \mathcal{L}_{\boldsymbol{\theta}}}_{\mathcal{I}_{\alpha\beta}} \delta\theta^\alpha\delta\theta^\beta + O(\delta\boldsymbol{\theta}^3), \quad \text{(A.15)}$$

and so the metric for the Bhattacharyya distance is directly proportional to the FIM.

A.4 Bounding Non-Analytic Models

In Section 3.1.1, we considered models $y_\theta(t)$, $t \in [-1, 1]$, that are continuously dependent on parameters $\theta = (\theta_1, \ldots, \theta_K)$ and analytic in an open neighborhood of $[-1, 1]$. We bounded the model manifold $\mathcal{Y}$ of model predictions by considering the truncated Chebyshev approximation from Eq. (3.3). When y_θ is not analytic on $[-1, 1]$, the convergence of Eq. (3.3) to y_θ as $N \to \infty$ is still controlled by the smoothness of y_θ. A standard result supplied in [154, Ch. 7] states that if y_θ has $\nu - 1 \geq 0$ derivatives that are absolutely continuous on $[-1, 1]$, with the νth derivative of total bounded variation $V < \infty$, then

$$(i)\ \|y_\theta - p_{N-1}\|_\infty \leq \frac{2V}{\pi\nu}(N - 1 - \nu)^{-\nu}, \quad N > \nu + 1, \tag{A.16}$$

$$(ii)\ |c_j| \leq \frac{2V}{\pi}(j - \nu)^{-(\nu+1)}, \qquad j \geq \nu + 1. \tag{A.17}$$

To bound $\mathcal{P}$, the model manifold of $p_{N-1}(\mathbf{t})$, we note that $p_{N-1}(\mathbf{t}) = X\tilde{\mathbf{c}}$ for $\mathbf{t} = (t_0, \ldots, t_{N-1})^T$, where $X = JD$, with $J_{ij} = T_{j-1}(t_{i-1})$, $D_{jj} = (j - 1 - \nu)^{-(\nu+1)}$ for $j \geq \nu + 2$, with $D_{jj} = 1$ otherwise. Likewise, we set $\tilde{\mathbf{c}} = (\tilde{c}_0, \ldots, \tilde{c}_{N-1})^T$, where $\tilde{c}_j = (j - \nu)^{(\nu+1)}c_j$ for $j \geq \nu + 1$, and $\tilde{c}_j = c_j$ otherwise. The singular values of X decay at, at least, an algebraic rate that increases with ν (see Fig. A.1). As in the analytic case, one can use X as a linear map and construct a hyperellipsoid H_Y that bounds the model manifold associated with $y_\theta(\mathbf{t})$. Its cross sections are controlled by the singular values of X and typically shrink algebraically fast.

As a question of nomenclature, we suggest that an object with an algebraic decay of widths should also be described as a hyperribbon. Although our mathematical bounds control the the asymptotic decay of widths, the decay of the first few, longest axes is usually of most interest in model predictions.

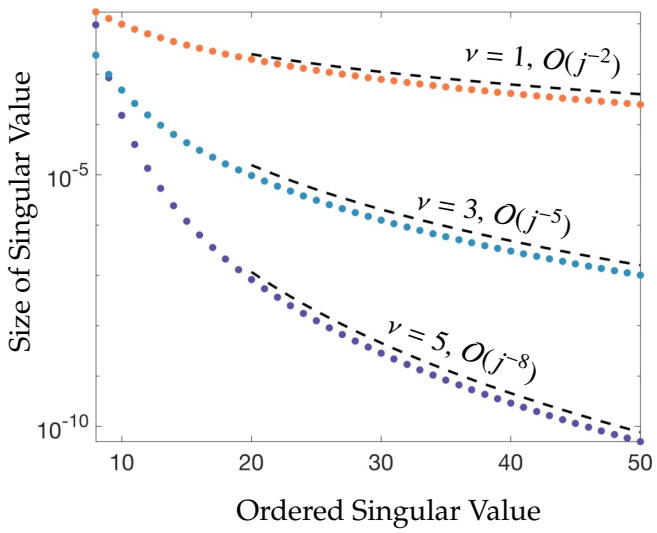


Figure A.1: **The singular values** $\sigma_j(X)$, where X is described in Appendix A.4, are plotted on a log scale against the index j for three models of the form $y_\theta(t) = f(\theta)|t|^\nu$: $\nu = 1$ (orange), $\nu = 3$ (blue), and $\nu = 5$ (purple). For simplicity, we assume f is smooth and independent of t. In each case, the model y_θ is ν-times differentiable on $[-1, 1]$. The asymptotic decay of the singular values (dotted black lines) is algebraic, with stronger decay rates as ν becomes larger. This suggests that continuously differentiable models have manifolds with (fat) hyperribbon structures, since a ν-times differentiable model y_θ has a manifold enclosed in H_Y, with $\ell_j(H_Y) \approx 2r\sigma_j(X)$ for some constant $r > 0$.

A.5 Deriving Manifold Bounds from Monomial Basis

In the Section 3.1.2, we bounded model predictions $y_\theta(\mathbf{t})$ evaluated at N points $\mathbf{t} = (t_0, \ldots, t_{N-1})^T$ by approximating y_θ with its degree $\leq N-1$ truncated Taylor expansion, which we denote by $p_{N-1}(t;\theta)$. The manifold associated with p_{N-1} is bounded within a hyperellipsoid H_P. The cross-sectional diameters of H_P are defined in terms of the singular values of the column-scaled Vandermonde matrix $X = VD$, where $(VD)_{ij} = t_{i-1}^{j-1} R^{-(j-1)}$. Here, we show how the bound on the hyperellipsoid was obtained, and provide numerical observations for high dimensional manifolds comparing the Chebyshev and monomial (Taylor expansion)

bases.

In deriving the bound for the monomial basis, we use the bound from Eq. (3.16). The Taylor series for y_θ expanded about any point $t \in [-1, 1]$ has a radius of convergence of at least R. One can use the Cauchy integral formula to show that the assumptions in Theorem 1 from the Chapter 3 implies that the derivative bound in Eq. (3.16) holds for some C and R dependent on M and ρ. If $t_0 = 0$ and $R > 1$, then we find by simple estimates that

$$\|y - p_{N-1}\|_\infty \leq \frac{C(NR - N + R)}{(1 - R)^2} R^{-N+1}. \tag{A.18}$$

As with the Chebyshev coefficients, we define $\tilde{a}_k = R^k a_k$, and express the polynomial predictions as $P(\theta) = VD\tilde{\mathbf{a}}$, where $V_{ij} = t_{i-1}^{j-1}$ and $D = \text{diag}(R^0, \ldots, R^{-(N-1)})$. While explicit bounds on the singular values of VD can be derived using its displacement structure [13], we require bounds that are characterized by the analyticity of y_θ. For this reason, we instead apply Theorem 2 to $DV^T VD$, so that $\sigma_j(VD)$ is bounded in terms of R. By applying the constraint from Eq. (3.16) to p_{N-1}, we see that $\|\tilde{\mathbf{a}}\|_2 < C\sqrt{N}$. It follows that the polynomial manifold $\mathcal{P}$ is bounded in a hyperellipsoid H_P from Eq. (3.17).

One can conclude the manifold associated with $y_\theta(\mathbf{t})$, is bounded in a hyperellipsoid H_Y with cross-sectional widths obeying

$$\ell_j(H_Y) \leq \ell_j(H_P) + 2\|y_\theta - p_{N-1}\|_\infty.$$

A.6 Connection Between Intensive Distance and Least-Squares

Here, provide the proof for Section 4.3.1, by showing that the intensive distance derived in Eq. (4.9) between two least-squares models is exactly the variance-

scaled Euclidean distance between predictions. Using the likelihood function shown in Eq. (4.11), we write out the intensive distance between two models as

$$d_I^2(\boldsymbol{\theta}_1, \boldsymbol{\theta}_2) = -8 \log \left(\int d\boldsymbol{x} \prod_i \frac{1}{\sqrt{2\pi\sigma_i^2}} \sqrt{\exp\left(-\frac{(x_i - y_{\boldsymbol{\theta}_1}(t_i))^2}{2\sigma_i^2} \right) \exp\left(-\frac{(x_i - y_{\boldsymbol{\theta}_2}(t_i))^2}{2\sigma_i^2} \right)} \right) \tag{A.19}$$

$$= -8 \sum_i \log \left(\frac{1}{\sqrt{2\pi\sigma_i^2}} \int dx_i \exp\left(-\frac{(x_i - y_{\boldsymbol{\theta}_1}(t_i))^2 + (x_i - y_{\boldsymbol{\theta}_2}(t_i))^2}{4\sigma_i^2} \right) \right) \tag{A.20}$$

$$= \sum_i \frac{(y_{\boldsymbol{\theta}_1}(t_i) - y_{\boldsymbol{\theta}_2}(t_i))^2}{\sigma_i^2}, \tag{A.21}$$

and so we get the distance shown in Eq. (4.12).

A.7 Deriving the Intensive Cross-Covariance Matrix

In this section, we use the replica trick to derive the intensive cross-covariance matrix in Eq. (4.18). Using the relation $x^N = 1 + N \log x + O(N)$, we write out the cross-covariance per replica as

$$\begin{aligned}
\frac{(MM^T)_{ij}}{N} &= \frac{4 \left\langle \boldsymbol{\theta}_i; \boldsymbol{\theta}_j \right\rangle^N}{N} - \frac{4}{Np} \sum_{k=1}^{p} \left(\langle \boldsymbol{\theta}_i; \boldsymbol{\theta}_k \rangle^N + \left\langle \boldsymbol{\theta}_j; \boldsymbol{\theta}_k \right\rangle^N \right) + \frac{4}{Np^2} \sum_{k,k'=1}^{p} \langle \boldsymbol{\theta}_k; \boldsymbol{\theta}_{k'} \rangle^N \\
&= \frac{4 + 4N \log \left\langle \boldsymbol{\theta}_i; \boldsymbol{\theta}_j \right\rangle - \frac{4}{p} \sum_{k=1}^{p} \left(2 + N \log \langle \boldsymbol{\theta}_i; \boldsymbol{\theta}_k \rangle + N \log \left\langle \boldsymbol{\theta}_j; \boldsymbol{\theta}_k \right\rangle \right)}{N} \\
&+ \frac{\frac{4}{p^2} \sum_{k,k'=1}^{p} 1 + N \log \langle \boldsymbol{\theta}_k; \boldsymbol{\theta}_{k'} \rangle}{N} + O(N) \\
&= 4 \log \left\langle \boldsymbol{\theta}_i; \boldsymbol{\theta}_j \right\rangle - \frac{4}{p} \sum_{k=1}^{p} \left(\log \langle \boldsymbol{\theta}_i; \boldsymbol{\theta}_k \rangle + \log \left\langle \boldsymbol{\theta}_j; \boldsymbol{\theta}_k \right\rangle \right) + \frac{4}{p^2} \sum_{k,k'=1}^{p} \log \langle \boldsymbol{\theta}_k; \boldsymbol{\theta}_{k'} \rangle + O(N).
\end{aligned} \tag{A.22}$$

When the above expression is considered in the limit as $N \to 0$, we obtain the form of the cross-covariance expressed in Eq. (4.18).

APPENDIX B

RANDOM MATRIX THEORY AND FISHER INFORMATION

This appendix shows preliminary results, speculating on a possible connection between the distribution of eigenvalues for the FIM in sloppy models and general ensembles of random matrices[1], through the construction of generalized Wishart ensembles with the Vandermone matrix.

Sloppy models are characterized by the eigenvalues of their FIM. The eigenvalues follow a geometric decay, *i.e.* they are roughly log-evenly distributed (successive eigenvalues are related by a constant factor). We would therefore expect that, for every point on the model manifold (*i.e.* for all fixed parameters $\boldsymbol{\theta}$), the eigenvalues of the metric follow a geometric decay. The ordered eigenvalues of the metric are therefore related by:

$$\log \lambda_i - \log \lambda_{i+1} \approx C, \tag{B.1}$$

for some constant C. This describes the probability distribution over the range of possible eigenvalues:

$$\mathcal{L}(\lambda) d \log \lambda = C d\mathcal{L}. \tag{B.2}$$

The probability density of the eigenvalues therefore goes like

$$\mathcal{L}(\lambda) \propto \lambda^{-C} \tag{B.3}$$

for some power C. Figure B.1 shows the numerically generated distributions for three such models (truncated by numerical precision), and all appear to follow a similar power-law decay.

Inspired by fruitful results in random matrix theory, such as the semi-circle law for the distribution of eigenvalues for Wigner matrices that has provided

[1] Motivation for this work was provided in an A-exam question by Liam McAllister.

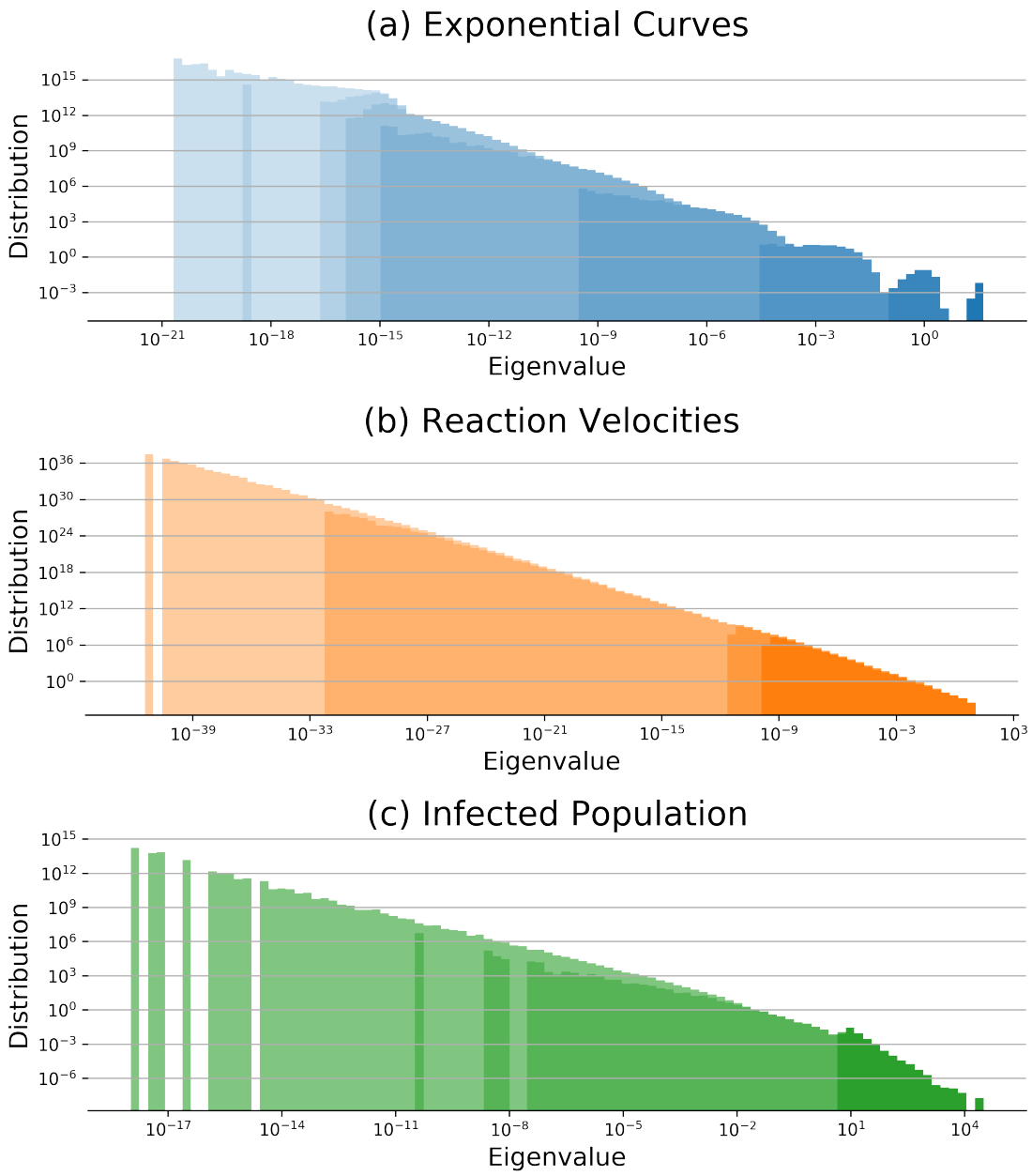


Figure B.1: **Eigenvalue distributions of local metrics** for three nonlinear least-squares models discussed in Chapter 3. (a) Exponential curves from Section 3.2.1, (b) reaction velocities of an enzyme-catalyzed reaction from Section 3.2.2, and (c) infected fraction of a population in an SIR epidemiology model from Section 3.2.3. Opacity represent order from largest to smallest (largest eigenvalue is darkest color). Distributions constructed from stacked histograms of numerically computed eigenvalues from sampled manifolds, and all appear to follow a power-law decay (note the huge range in values for the horizontal and vertical axes).

fruitful results for understanding energy levels in atomic nuclei [160, 138], we consider sets of random matrices to explain features in the eigenvalue distributions of sloppy models. Because of the FIM decomposition in Eq. (B.6), and the connection between the FIM and the Jacobian of the model (where $\mathcal{I} = J^T J$ for least-squares models, discussed in Section 2.2), we consider generalizations of the Wishart ensemble of random matrices.

A classic least squares model has predictions given by $y_{\boldsymbol{\theta}}(t)$. If we assume analyticity of the underlying model, then this model can be seen as a point in the space of all possible model predictions, as described by some complete polynomial basis $\{\phi_j\}_{j=0}^{\infty}$. We can perform a Taylor expansion of the model to obtain:

$$y_{\boldsymbol{\theta}}(t) = \sum_n \frac{1}{n!} \frac{\partial^n y_{\boldsymbol{\theta}}(\bar{t})}{\partial t^n} (t - \bar{t})^n, \tag{B.4}$$

where we are describing the model in terms of its location in the space of monomials. The Jacobian of the model is given by

$$J_{i\alpha} = \frac{1}{\sigma_i} \frac{\partial y_{\boldsymbol{\theta}}(t_i)}{\partial \theta^\alpha} = \sum_n \underbrace{\frac{1}{\sigma_i n!} \frac{\partial}{\partial \theta^\alpha} \frac{\partial^n y_{\boldsymbol{\theta}}(\bar{t})}{\partial t^n}}_{M_{n\alpha}} \underbrace{(t_i - \bar{t})^n}_{V_{in}}. \tag{B.5}$$

We can therefore decompose the Jacobian into a product of a Vandermonde matrix (that varies with sampled points) and a matrix of derivatives (that varies with parameters). The FIM can be expressed as

$$\mathcal{I} = M^T V^T V M = M^T U^T \Sigma U M, \tag{B.6}$$

where Σ is a diagonal matrix whose entries are the squares of the singular values of the Vandermonde matrix, and U is an orthonormal matrix. Without loss of generality, we shift and rescale all points t_i such that $|t_i - \bar{t}| < 1$. For each fixed parameters $\boldsymbol{\theta}$, we find a maximum characteristic length $R(\boldsymbol{\theta})$ such that for all t on the interval containing points $\{t_i\}$

$$\frac{1}{\sigma n!} \frac{\partial}{\partial \theta^\alpha} \frac{\partial^n y_{\boldsymbol{\theta}}(t)}{\partial t^n} < R(\boldsymbol{\theta})^{-n}, \tag{B.7}$$

where $\sigma = \min_i \sigma_i$. If we consider the parametrization where $\theta^\alpha = \frac{\partial^n y(\bar{t})}{\partial t^n}$ (*i.e.* where the terms in the Taylor series expansion are themselves the parameters), then using Theorem 2 (from Chapter 3), we have that the diagonal entries of the diagonal entries are bounded by

$$\Sigma_{nn}(\boldsymbol{\theta}) < O\left(\frac{R(\boldsymbol{\theta})^{-2n}}{|R(\boldsymbol{\theta})^2 - 1|}\right). \tag{B.8}$$

The spectra in Fig. B.1 are from three disparate models, all subject to the same constraint in R. They all follow the same power-law decay, (one can see this either by comparing the plots in Fig. B.1 or from Fig. 2.4 which superimposed the distributions) however there appears to be additional features that distinguish them from each other. For instance, while the eigenvalue spectra for reaction velocities in Fig B.1(a) appears to follow exactly a power-law decay, the other two show more structure (bumps and local peaks). There are noticeable differences in the location of the ordered eigenvalues (*i.e.* the distribution of the largest as compared to the smallest), as well as eigenvalues that span noticeably different overall ranges. To account for these "second order" effects in the distribution, we turn to random matrices. Specifically, we consider different ways to construct matrix M, the parameter-dependent component in Eq. (B.6).

B.1 Correlated Random Matrices

As a preliminary investigation of a possible connection between the FIM and random matrices, we consider the sets of matrices shown in Fig. B.1. We use the decomposition of the FIM shown in Eq. (B.6), truncating the series at finite values, to numerically approximate the exact the eigenvalues. If there is an ensemble of random matrices that describes the resulting eigenvalue distribution, then we can consider the set of matrices M as elements drawn by this

distribution (without needing to know its exact functional form or expression). By using the set of matrices M to recreate different ensembles (say, for all elements, for each matrix index separately, for the rows and columns) we control how correlated the elements are, to see what features are important for exactly recreating the original eigenvalue distributions.

Figure B.2(a,c,e) shows the distribution of matrix entries $M_{n\alpha}$ for the three least-squares models considered here, and Fig. B.2(b,d,f) confirms that the approximate eigenvalues constructed from the truncated series expansions in Eq. (B.6) match the exact values up to numerical precision.

We now use the entries of $M_{n\alpha}$ to form ensembles from which to draw elements to generate eigenvalue distributions. By varying (1) individual elements, $M_{n\alpha}$, (2) rows, $M[n,:]$, and (3) columns, $M[:,\alpha]$, we consider the effect of correlations on the resulting eigenvalue distributions. We generate random matrices of the form

$$F = X^T V^T V X \tag{B.9}$$

where V is the Vandermonde matrix from Eq. (B.6), and X is a random matrix constructed by drawing elements from the set of numerically generated matrices, $\{M\}$. X is an $N_e \times N_p$ matrix, where N_e is the order of the expansion ($N_e = 160$ for exponential curves, $N_e = 20$ for reaction velocities, and $N_e = 20$ for model of infected fraction of a population) and N_p is the number of parameters ($N_p = 8$ for exponential curves, $N_p = 4$ for reaction velocities, and $N_p = 3$ for model of infected fraction of a population).

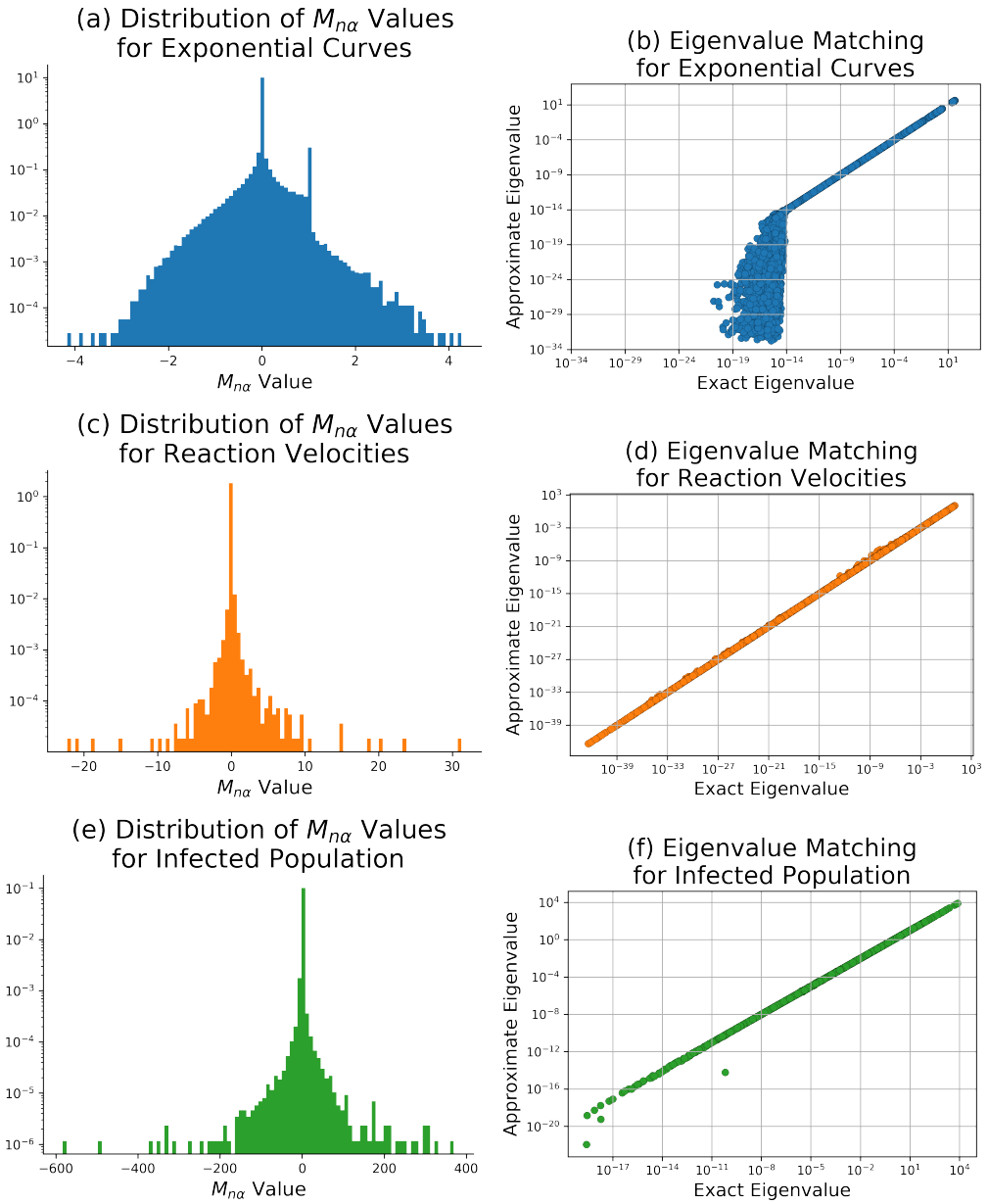


Figure B.2: **Approximating FIM** through truncated Taylor expansion, following the decomposition in Eq. (B.6), for the nonlinear model of exponential curves discussed in Section 3.2.1 and model of reaction velocities in an enzyme-catalyzed chemical reaction discussed in Section 3.2.2. (a,c,e) Distribution of matrix entries, $M_{n\alpha}$. (b,d,f) Matching FIM with truncated Taylor series. The horizontal axis corresponds to the exact eigenvalue of an FIM from the distribution of eigenvalues shown in Fig. B.1, and vertical axis represents the corresponding eigenvalue from the FIM of a truncated Taylor expansion. Eigenvalues match exactly, up to python's default numerical precision (note that the horizontal and vertical axes are log-scaled).

B.1.1 Uncorrelated Entries

In considering uncorrelated, independently drawn entries of X (for the random matrices $F = X^T V^T V X$ in Eq. B.9), we generate two different ensembles from the collection of matrices M. The first is shown in Fig. B.2(a). The collection of all entries from all M from the metrics sampled on the manifold for each model forms an ensemble $\mathcal{L}_{\text{all}}$ for each of these models, from which all elements of X are drawn ($X_{n\alpha} \sim \mathcal{L}_{\text{all}}$). The resulting eigenvalue distributions are shown in Fig. B.3(c), which preserves the same decay rate as the original distributions but which fails to capture the fine features at large eigenvalues: the bumps in the original distribution appear smoothed over.

As a second consideration, each index (n, α) of M (from the explicit FIM decomposition in Eq. (B.6)) is used to index disparate ensembles, which we label as $\mathcal{L}_{n,\alpha}$. The elements of X (for the random matrices F in Eq. B.9) are drawn from these ensembles, $X_{n\alpha} \sim \mathcal{L}_{n\alpha}$, and the resulting distribution of eigenvalues is shown in Fig. B.3(d). While the overall decay rate of the distribution is preserved, the structure of the distribution at large eigenvalues still isn't perfectly captured.

For uncorrelated entries, the overall eigenvalue distribution captures the correct decay rates, however detailed features for large eigenvalues are not preserved. Specifically, the first, second and third "bumps" are not as pronounced as they should be, and the location of the local beaks is shifted. The whole spectrum is also "shifted" to the right, with far larger maximum eigenvalues than the original distribution.

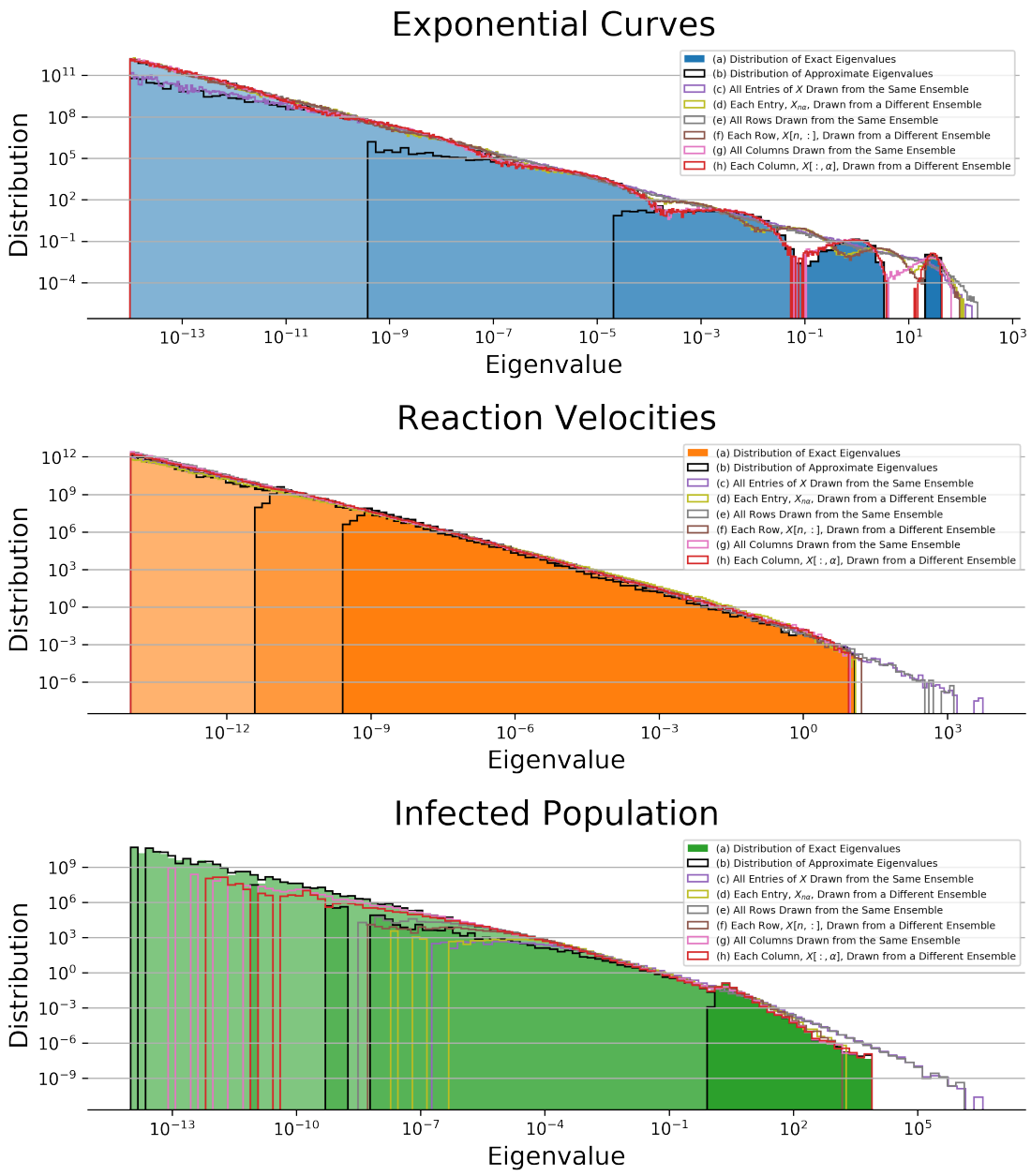


Figure B.3: **Eigenvalue distributions** for different nonlinear models. (a) Exact distribution compared to the eigenvalue distribution from (b) truncated series expansion. Opacity reflects the eigenvalue order. Random matrices are of the form in Eq. (B.9). Uncorrelated entries: each element of X is drawn from (c) the same ensemble (shown in Fig. B.2) and (d) different ensembles. Correlated matrices: we draw (e) all rows of X from the same ensemble and (f) each row $X[n,:]$ from its own ensemble, and (g) all columns of X from the same ensemble and (h) each column $X[:,\alpha]$ from its own ensemble. All ensembles follow the same geometric decay, but have different structure for large eigenvalues.

B.1.2 Correlated Rows and Columns

We consider random matrix ensembles $F = X^T V^T V X$ (from Eq. B.9) where the rows and columns of X are correlated. First, we generate an ensemble of all rows, $\mathcal{L}_{\text{row}}$, and for rows of fixed index n, $\mathcal{L}_n$, from the rows of M, $M[n, :]$. This is equivalent to considering correlations between parameters, for fixed point derivatives n. While the eigenvalue distributions for both $X[n, :] \sim \mathcal{L}_{\text{row}}$ and $X[n, :] \sim \mathcal{L}_n$ both capture the correct overall geometric decay, they both fail to capture features at large eigenvalues, as shown in Fig. B.3(e) and Fig. B.3(f) respectively.

Finally, we generate an ensemble of all columns, $\mathcal{L}_{\text{col}}$, and for columns of fixed index α, $\mathcal{L}_\alpha$, from the columns of M, $M[:, \alpha]$. This is equivalent to considering correlations between derivatives for fixed model parameter, θ^α. In other words, considering correlations between $\frac{\partial}{\partial \theta^\alpha} \frac{\partial^n y_\theta(t)}{\partial t^n}$ for fixed θ^α and varying n. Again, the overall geometric decay for the two resulting distributions captures the original decay, but now the ensemble appears to preserve much of the important features for large eigenvalues, as shown in Fig. B.3(g) and Fig. B.3(h) for $\mathcal{L}_{\text{col}}$ and $\mathcal{L}_\alpha$ respectively.

These results show the importance of correlations in the original matrices M in the FIM decomposition from Eq. (B.6). By allowing for correlated rows, randomly generated matrices $F = X^T V^T V X$ (from Eq. (B.9)) have eigenvalue distributions that capture the important features of the original eigenvalue distribution. Future work aims to determine exactly what these correlation functions look like, to see if they can be effectively approximated with two-point

correlation functions over all elements,

$$C_{n\alpha m\beta} = \left\langle M_{n\alpha} M_{m\beta} \right\rangle - \langle M_{n\alpha} \rangle \left\langle M_{m\beta} \right\rangle . \tag{B.10}$$

The Wishart ensemble considers random matrices of the form $X^T X$. By introducing a Vandermonde matrix into the heart of this multiplication, we are introducing correlations in the columns of $X \to VX$ (since all entries in a column of V are exponentiated by the same power). By considering correlated entries in X, we can explore properties of a more general Wishart ensemble that considers correlated rows and columns. This is the focus of future work.

B.2 Sequential Random Matrices

A great advantage of neural networks in machine learning is that they can be used as universal function approximators for non-linear models [68, 29]. We consider a multi-layer, deep neural network (say, of the type described in Section D.3). Each layer of a neural network is a matrix, whose entries are optimized during the training process. By considering a many layers of a network, we can consider the effect of sequentially multiplying many random matrices.

Note that a re-parametrization of the model changes the FIM in the following way[2]:

$$\mathcal{I}_{\tilde{\alpha}\tilde{\beta}} = \frac{\partial \theta^{\alpha}}{\partial \theta^{\tilde{\alpha}}} \frac{\partial \theta^{\beta}}{\partial \theta^{\tilde{\beta}}} \mathcal{I}_{\alpha\beta}. \tag{B.11}$$

Because the nonlinear model parameters $\boldsymbol{\theta}$ determine the coefficients in the expansion of Eq. B.4, we can view them as a re-parametrization of the coefficients.

[2]Motivation for considering nested random functions was the result of multiple conversations at the 2018 ICAM Workshop at CUNY, on Machine Learning and Physics.

Furthermore, because there are fewer model parameters than coefficients in the polynomial expansion, this is a natural way of reducing the intrinsic dimensionality of the system.

Given the form of the FIM shown in Eq. (B.6), we look at Wishart ensembles with matrices of the form $X^T X$ for rectangular matrices X, where the random matrix X is analogous to the Jacobian of the transformation. We decompose X in the following way:

$$X_{i\alpha} = \sum_{n=0}^{N-1} V_{in} R^{-n} M_{n\alpha} \tag{B.12}$$

where V is the Vandermonde matrix, $R \geq 1$ is a random number (reflecting the fact that the exact smoothness of the function varies with location on the manifold), $M_{n\alpha}$ is a matrix whose entries are correlated. V reflects the properties of the sampled points t_i (experimental conditions), M and R reflects the wide range of random functions available to describe the model (*i.e.* random entries in the Taylor series, with each coefficient it's own parameter, whose smoothness are characterized by R) and is used to reduce the intrinsic dimensionality of the system to the number of parameters in the nonlinear model and reflects the parameter dependence (*i.e.* from the dimension of the embedding space to that of the parameter space). The characterstic length R can be fixed, or allowed to vary slightly, reflecting the fact that $R(\boldsymbol{\theta})$ in Eq. (B.7) changes with parameters. We consider lengths such that $R_{\min} = \min_{\boldsymbol{\theta}} R(\boldsymbol{\theta}) \geq 1$.

To account for the effect of a many-layered neural network, we include it in the construction of the parameter matrix M. We let $M^{(0)}$ be an $N \times K$ matrix of random entries (where N is the number of points t_i and K is the number of parameters). We then let $M^{(j)}$ for $j \geq 1$ be a set of $K \times K$ matrices of random

entries, and construct the parameter matrix as

$$M = M^{(0)} M^{(1)} \cdots \tag{B.13}$$

where the number of $M^{(j)}$ in the above product reflects how correlated the parameters are. The FIM decomposition of Eq. B.6 can therefore be expressed as

$$\mathcal{I} = \cdots M^{(1)^T} M^{(0)^T} R \underbrace{V^T V}_{\text{(a)}} \underbrace{R}_{\text{(b)}} \underbrace{M^{(0)}}_{\text{(c)}} \underbrace{M^{(1)} \cdots}_{\text{(d)}} \tag{B.14}$$

where the term in (a) is a reflection of experimental conditions, (b) is a reflection of the underlying smoothness, (c) reflects the dimensionality reduction to the number of model parameters, and (d) determines the correlation between parameters.

The geometric decay in the eigenspectra is ultimately due to the Vandermonde matrix at the heart of the FIM. Finer details of the distribution (overall range of eigenvalues, number of local peaks, spread/isolation of the ordered eigenvalues) are affected by the variability in characteristic lengths R, dimensionality reduction from $M^{(0)}$ and the parameter correlations from $M^{(j)}$. To better understand these finer details, we vary characterstics of the underlying distribution and see what effect they have.

The eigenspectra can be decomposed into peaks, related to the order of the eigenvalues. For instance, if the model has 3 parameters, then the spectra has 3 local peaks, related to the largest, middle, and smallest eigenvalue. The size and spacing between these peaks relates directly to the eigenvalue spacing in the FIM because each ordered eigenvalue is drawn from its corresponding peak. Figure B.4 illustrate now the number of peaks changes with varying number of sampled points $\{t_i\}$, which affects the intrinsic dimensionality of the system (dimension of the embedding space for the model manifold) as well as by decreasing the number of model parameters.

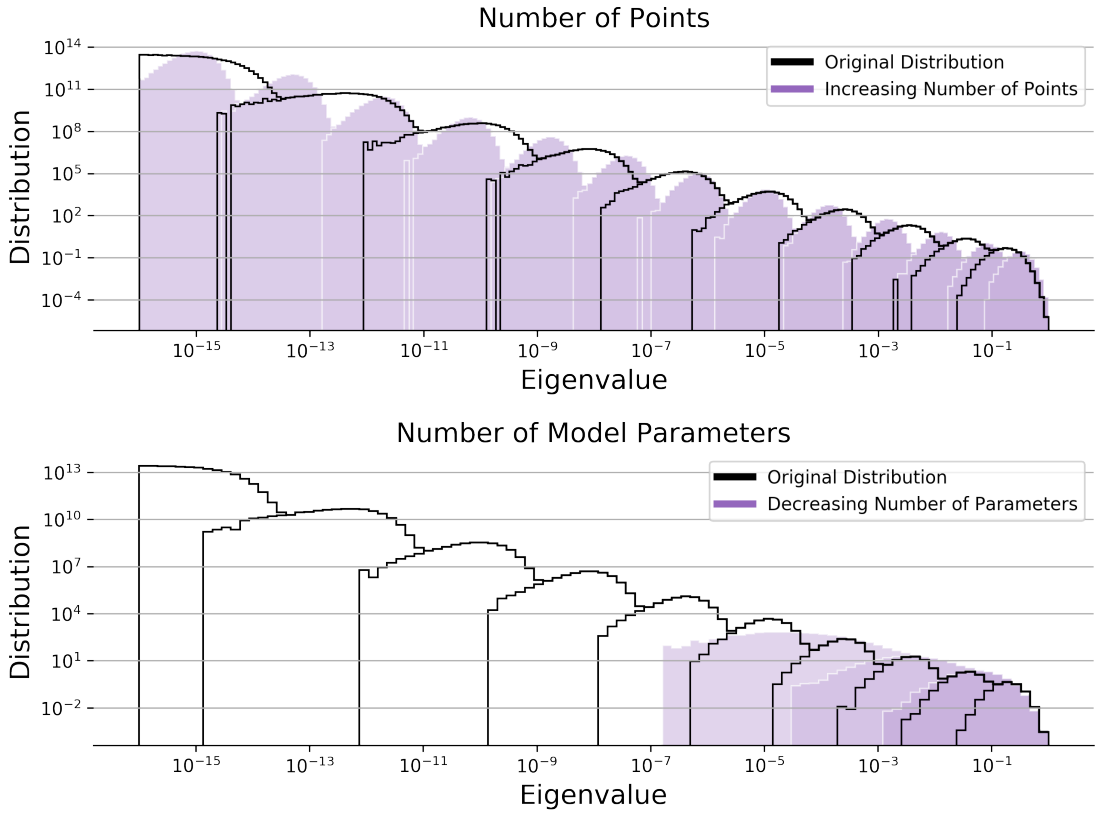


Figure B.4: **The number of peaks in the eigenvalue distributions of sequential random matrices** are characterized by the number of points in the Vandermonde matrix (number of sampled points) as well as the number of model parameters. By increasing the number of points from 11 to 21, the number of peaks increases from 11 to 21. By decreasing the number of model parameters from 11 to 3, the number of peaks also decreases to 3. In both figures, black line represents distribution from the original distribution, with each peak corresponding to the ordered eigenvalues. Purple distribution reflects the new eigenvalue distribution after transformation, with opacity reflecting eigenvalue order (dark purple being the largest eigenvalue).

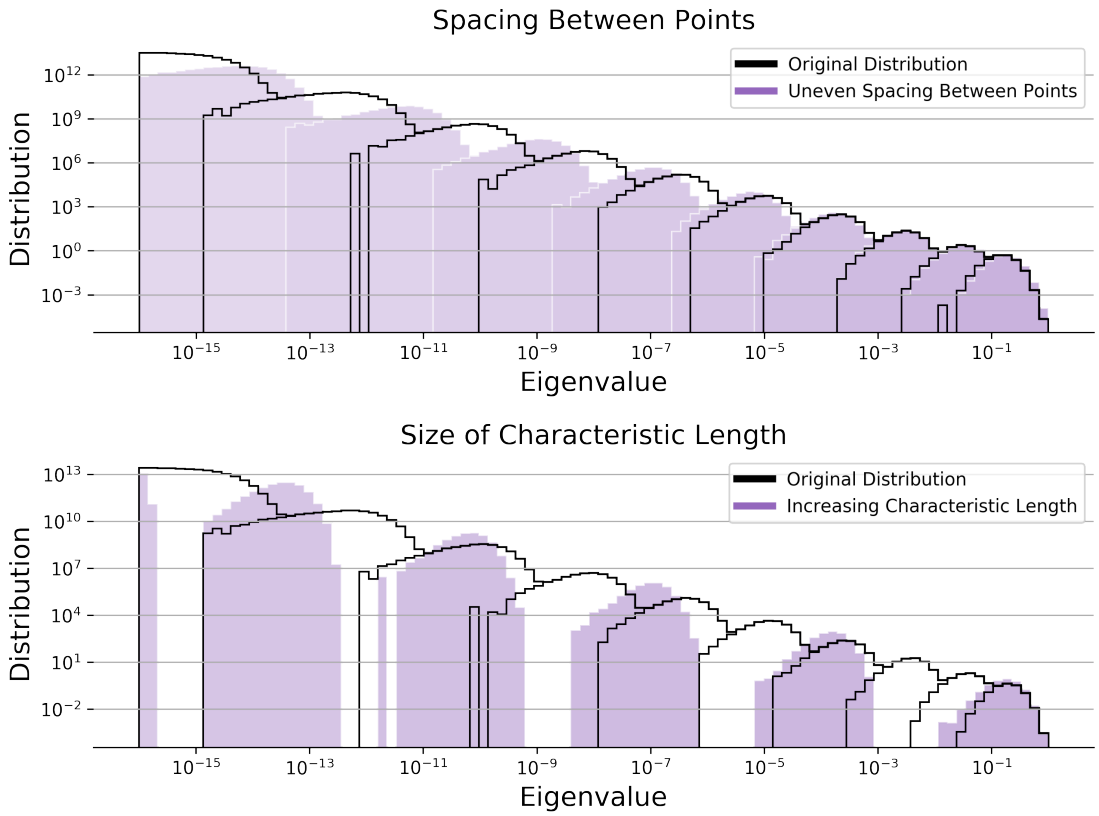


Figure B.5: **The spacing between peaks** are related to the spacing between points (even vs. uneven) as well as the characteristic length. Peaks become more pronounced by making the spacing between points more uneven, and by increasing the size of the characteristic length the peaks become more spread out. In both figures, black line represents distribution from the original distribution, with each peak corresponding to the ordered eigenvalues. Purple distribution reflects the new eigenvalue distribution after transformation, with opacity reflecting eigenvalue order (dark purple being the largest eigenvalue).

The spacing between peaks depends on the spacing between sampled points, $\{t_i\}$, as well as the characteristic length R, as shown in Fig. B.5. The more unevenly distributed the points, the more pronounced the peaks are, and the greater the characteristic length, the more separated the peaks are.

Finally, the spread in peaks is determined by how correlated the parameters

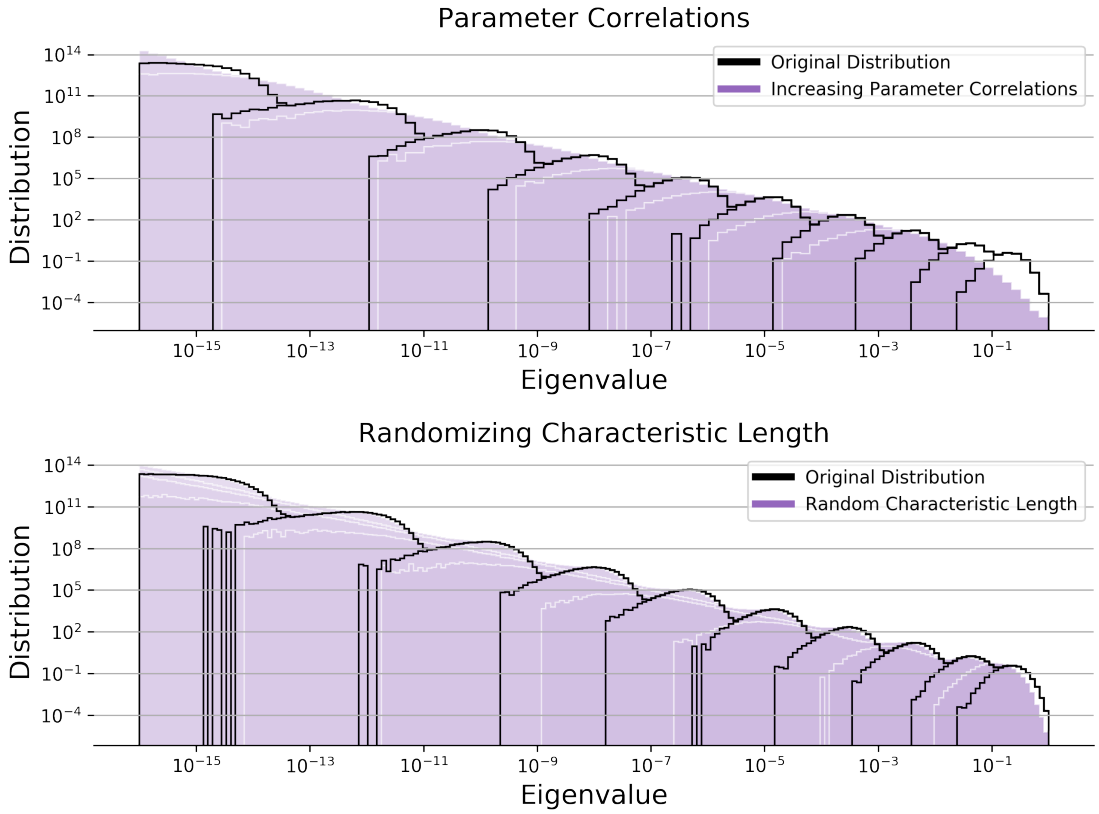


Figure B.6: **The spread in the peaks** are determined by the correlations in the FIM, as characterized by the correlations in model parameters as well as allowing the characteristic length to vary. The spread of the peaks increases by increasing the correlations and by allowing the characteristic length to vary. In both figures, black line represents distribution from the original distribution, with each peak corresponding to the ordered eigenvalues. Purple distribution reflects the new eigenvalue distribution after transformation, with opacity reflecting eigenvalue order (dark purple being the largest eigenvalue).

are and by the spread in the characteristic lengths, R, as shown in Fig. B.6. The more correlated the parameters (*i.e.* the more $M^{(j)}$ in the construction of the FIM), the more spread out the peaks are.

By varying these different features in the ensemble, we try to fit the eigenspectra from Fig. B.1. Here, elements of $M^{(j)}$ are drawn from a Gaussian dis-

tribution, $M_{ij}^{(j)} \sim \mathcal{N}(0,1)$. By varying the number of sequential matrices, and allowing R to vary slightly between drawn samples, the distributions can be more easily fit. However, this picture of sequential random matrices doesn't appear to quite fit the distributions perfectly, and a better approach may be explicitly incorporating correlated elements, as described in Section B.1.

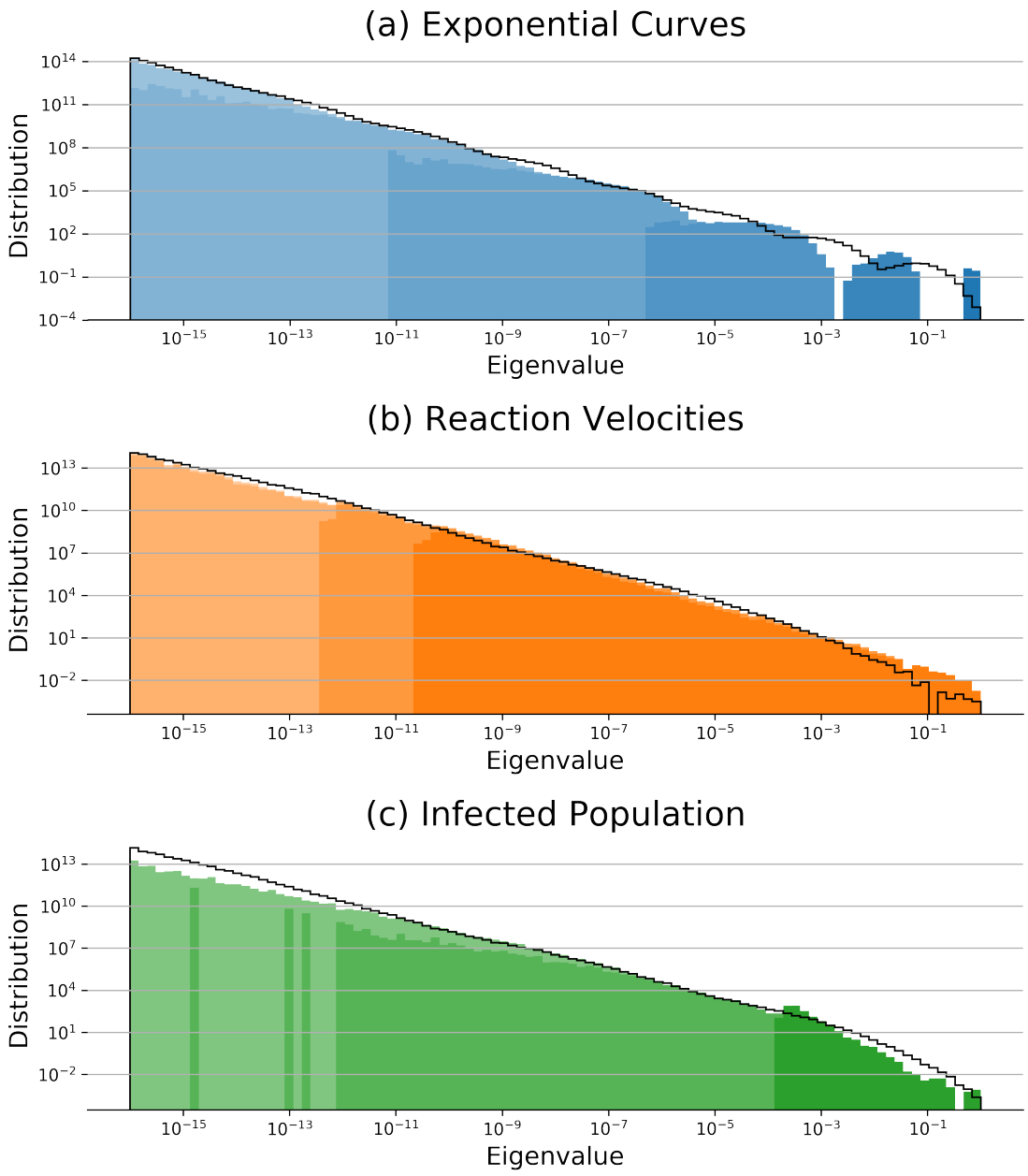


Figure B.7: **Eigenvalue distributions of local metrics** for three nonlinear least-squares models, with fits from RMT ensembles generated using Eq. (B.14). (a) Exponential curves from Section 3.2.1, (b) reaction velocities of an enzyme-catalyzed reaction from Section 3.2.2, and (c) infected fraction of a population in an SIR epidemiology model from Section 3.2.3. Opacity represent order from largest to smallest (largest eigenvalue is darkest color). Superimposed on each plot is a black line, representing the distribution constructed from *sequential* random matrices.

APPENDIX C

NUMERICAL OBSERVATIONS FOR HIGH DIMENSIONAL MANIFOLDS

In Chapter 3, we derive bounds on model predictions using two different basis in a polynomial approximation: (1) Chebyshev expansion in Section 3.1.1 and (2) Taylor expansions in Section 3.1.2.

The constraint in Eq. (3.16) implies that y_θ is analytic in the region $\mathcal{R}$ of the complex plane of distance $< R$ from $[-1, 1]$. It can be shown that y_θ must also be analytic and bounded by a function $M(\zeta)$ on any Bernstein ellipse $E_{\rho(\zeta)}$ in $\mathcal{R}$, with $\rho(\zeta) = \zeta + \sqrt{\zeta^2 + 1}$ [42]. The largest such ellipse is given by $\rho_{\max} = R + \sqrt{R^2 + 1}$, suggesting that Chebyshev-based bounds can improve the bounds from Eq. (3.17) by nearly a factor of 2^j. However, $M(\zeta)$ is unbounded as $\zeta \to R$, so one must select $0 < \zeta < R$ to minimize the Chebyshev bound. Even when ζ is selected carefully, the conversion from Eq. (3.16) to a constraint involving $E_{\rho(\zeta)}$ may introduce an unphysically large constant into the bound.

One expects that the decay rate $O(R^{-j})$ in Eq. (3.17) is weak as an upper bound on the ordered widths of the underlying hyperribbon $\mathcal{Y}$. This is related to the fact that unlike truncated Chebyshev expansions, truncated Taylor polynomials do not converge to y_θ at a rate that is asymptotically optimal for polynomial approximants (see [154, Ch. 12–16]).

However, we find that the singular values $\sigma_j(VD)$ behave in a surprising way: For small to moderate j, the magnitude of $\sigma_j(VD)$ decays at a rate close to the limit predicted by Chebyshev approximation: $O(\rho_{\max}^{-j})$, where $\rho_{\max} = R + \sqrt{R^2 + 1}$. It is only when j is sufficiently large that $\sigma_j(VD)$ appears to decay at the predicted rate $O(R^{-j})$. We do not yet fully understand why the

singular values of VD decay at two distinct rates, but speculate that it may be related to the kink observed in error plots for Clenshaw–Curtis quadrature on analytic functions [158].

Due to this phenomenon, we find that using $\sigma_j(VD)$ directly results in good bounds on model prediction spaces for low dimensions (the larger axes of the hyperellipsoid H_Y). At higher dimensions (shorter hyperellipsoid axes), the Taylor-based bounds become suboptimal, and it is beneficial to instead convert the constraint in Eq. (3.17) to one involving Bernstein ellipses. The conversion of the constraint can result in bounds that are inflated by a large unphysical constant, but the decay rate in the new bound, close to $O(\rho_{\max}^{-j})$, is nearly double the rate $O(R^{-j})$. When viewed together, the Chebyshev-based bounds and numerical Taylor-based bounds describe the successive lengths of the model manifold across two regimes (low vs. high dimension). We illustrate this observation using a high-dimensional manifold ($N = 100$) in Fig. C.1.

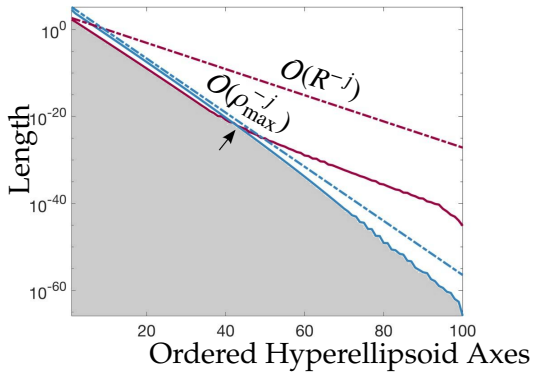


Figure C.1: **Bounds on the hyperellipsoid lengths** $\ell_j(H_P)$ using truncated Taylor (dotted purple) and truncated Chebyshev (dotted blue) expansions are plotted on a log scale against the dimension index j. These form a universal bound on the ordered manifold widths of the prediction space for models y_θ that satisfy Eq. (3.17). In this example, $C = 1$, $R = 2$, $N = 100$, and $\rho_{\max} \approx 4.2$. The solid lines show the actual computed hyperellipsoid cross-sectional lengths (on a log scale) $\ell_j(H_P) = 2r\sigma_j(X)$, where $X = VD$ for the Taylor-based bounds, and $X_{ij} = T_{j-1}(t_{i-1})\rho_{\max}^{-(j-1)}$ for the Chebyshev-based bounds. The largest 40 Taylor-based hyperellipsoid lengths decay at the rate predicted by the Chebyshev-based bounds. Then, a kink occurs (indicated by a black arrow) and the lengths decay at the rate predicted by the bound in Eq. (3.17). For the smaller dimensions, the Chebyshev-based results produce tighter bounds. Model manifold lengths outside of the shaded region cannot occur.

APPENDIX D

GENERATING AND VISUALIZING MODEL MANIFOLDS

In this appendix, we go through the details of how the different model manifolds were generated in Chapter 3 and Chapter 4. In Section D.1, we go through the least-squares models discussed in Chapter 3 and explain how the manifolds were projected along the natural axes of the bounding hyperellipsoids. In Section D.2, D.3, and D.4 we discuss the probabilistic models outlined in Chapter 4 and detail how their manifolds were sampled and visualized with InPCA.

D.1 Least Squares Models

Here, we provide a detailed description of how data for the 1D models used in Section 3.2 were generated: physics (exponential curves), chemistry (reaction velocities) and biology (SIR epidemiology model). Data for the 2D extension of all three models (shown in Section 3.3) were computed in a similar way.

In order to generate the model manifolds, a Monte Carlo sampling was performed on the parameter space of all three models. The model predictions for the randomly selected parameters were accepted or rejected based on whether or not they satisfied the constraint on the derivative from Eq. (3.17), where we set $C = 1$ and $R = 2$. Since we consider eleven equally spaced points for all three models in Section 3.2, in all example models the derivative constraint was applied up to the eleventh derivative.

1. For *exponential curves*, the model is of the form

$$y_\theta(t) = \sum_{\alpha=0}^{10} A_\alpha \exp\left(-\lambda_\alpha t\right), \tag{D.1}$$

and the derivative constraint from Eq. (3.17) can be expressed as

$$\sum_{k=0}^{N-1}\left(\sum_{\alpha=0}^{10}\frac{R^k A_\alpha}{k!}(-\lambda_\alpha)^k \exp\left(-\lambda_\alpha t\right)\right)^2 < C^2 N \tag{D.2}$$

for all $-1 \leq t \leq 1$. From a Monte Carlo sampling, 42,000 valid samples were randomly generated. A histogram of parameters used to generate the model manifold is shown in Fig. D.1(a).

2. The model of *reaction velocities* is given by

$$y_\theta(t) = \frac{\theta_1 t^2 + \theta_2 t}{t^2 + \theta_3 t + \theta_4}, \tag{D.3}$$

where t is the substrate concentration. The derivative constraint can be expressed as

$$\sum_{k=1}^{N}\left(\frac{R^k}{k!}\frac{d^k}{dt^k}\left(\frac{\theta_1 t^2 + \theta_2 t}{t^2 + \theta_3 t + \theta_4}\right)\right)^2 < C^2 N, \tag{D.4}$$

for all $-1 < t < 1$. We generated 24,000 valid parameter combinations, and a histogram of the different parameter values is shown in Fig. D.1(b).

3. Finally, for the *infected population* in an SIR model, the number of people susceptible (S), infected (I), and recovered (R) are determined through three coupled differential equations:

$$\begin{aligned}
&(i)\ \dot{S} = -\beta\frac{IS}{N_{tot}},\\
&(ii)\ \dot{I} = \beta\frac{IS}{N_{tot}} - \gamma I,\\
&(iii)\ \dot{R} = \gamma I,
\end{aligned}$$

where β is the infection rate, γ is the recovery rate, and N_{tot} is the total size of the population. If we let the model predictions be the infected population, then we have $y_\theta(t) = I(t)$. To find the kth derivative of such a model, we note that $\dot{S} = f_1(S, I)$ and $\dot{I} = g_1(S, I)$. The subsequent derivatives can

therefore be found recursively, by $\ddot{y}_\theta = \ddot{I} = \frac{dg_1}{dS}\dot{S} + \frac{dg_1}{dI}\dot{I} = g_2(S, I)$ and so on. From a Monte Carlo sampling, we obtained 20,000 valid parameter combinations. A histogram of parameter values used to generate the model manifold is shown in Fig. D.1(c).

In all three models, the smallest physically meaningful prediction is $y_\theta(t) = 0$. For exponentials and the SIR model, the largest physically meaningful prediction allowed by Eq. (3.17) is $y_\theta(t) = C\sqrt{N}$, and so the longest manifold distance possible is CN. With this sampling method, we obtained manifold lengths that are within 1.5% of this maximally allowed distance, and so while more refined sampling methods could be used to resolve the manifold boundaries, they are unnecessary for our purposes.

Once a sampling of the possible parameter combinations is obtained for a model, we visualize it. Each parameter combination is evaluated at eleven equally spaced points. The space spanned by the model predictions at these points forms the model manifold $\mathcal{Y}$.

To visualize $\mathcal{Y}$, it is rotated into the basis given by the hyperellpsoid axes constructed from the space of allowed polynomials predictions, $\mathcal{P}$. Let $\{\phi_j\}_{j=0}^{\infty}$ be a complete polynomial basis, and let $P(\mathbf{b}) = (P_0, \ldots, P_{N-1})$ define the model manifold $\mathcal{P}$ of $p_{N-1}(t) = \sum_{j=0}^{N-1} b_j \phi_j(t)$. Polynomial predictions are given by $P_k = p_{N-1}(t_k)$. By definition, $P(\mathbf{b}) = X\mathbf{b}$, where $X_{ij} = \phi_{j-1}(t_{i-1})$ and $\mathbf{b} = (b_0, \ldots, b_{N-1})^T$. To find the rotation matrix used to visualize the model manifold $\mathcal{Y}$, we perform a singular value decomposition on X,

$$X = U\Sigma V^T, \tag{D.5}$$

to extract the rotation matrix U. The data points on the model manifold are then

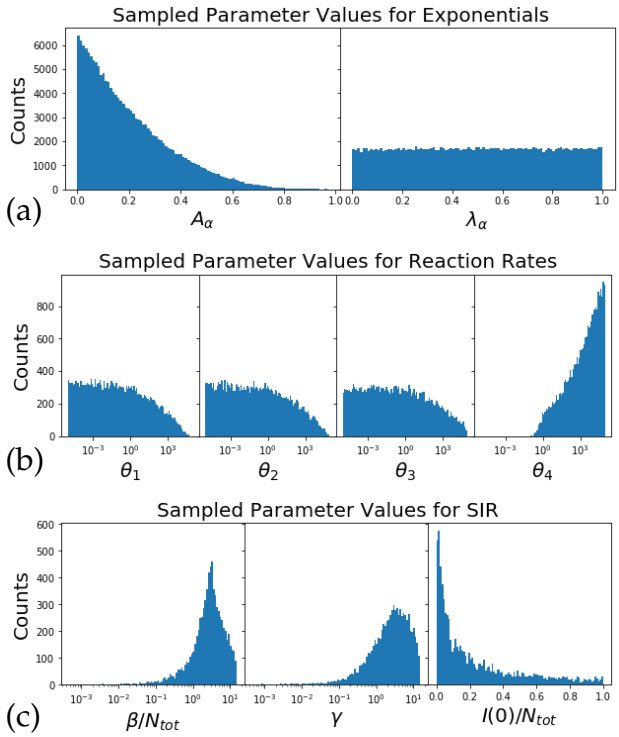


Figure D.1: **Histograms of valid parameter values** used to generate the model manifolds. In all the models, a Monte Carlo sampling was performed, with parameters accepted or rejected based on whether or not they satisfied the derivative condition from Eq. (3.17). (a) Parameter values for exponentials, showing the distributions for the amplitudes A_α and decay rates λ_α. (b) Parameter values for the reaction velocities, for each $\theta_1, \theta_2, \theta_3$ and θ_4. (c) Parameter values for the SIR epidemiology model, showing the distribution of infection rates β/N_{tot}, recovery rates γ and initial infected population.

rotated using this matrix, and visualized in Fig. 3.1(b) and Fig. 3.2(b) where we set $X = VD$ to be the column-scaled Vandermonde matrix.

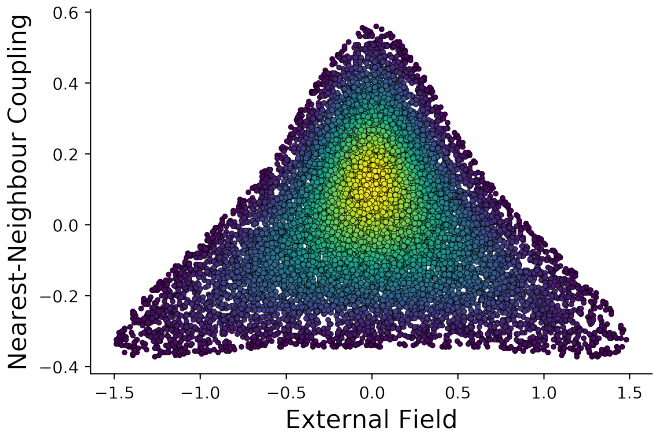


Figure D.2: **Ising parameter ranges** sampled in this thesis, coloured by Jeffrey's prior (yellow indicates regions of high probability, purple areas of low probabilities).

D.2 2D Ising Model

The Ising model manifolds generated in Chapter 4 were sampled around the critical point using an MC with probability of accepting/rejecting a step given by Jeffrey's prior, equivalent to the determinant of the FIM:

$$\mathcal{L}_{\text{Jeff}}(\boldsymbol{\theta}) = \sqrt{|\mathcal{I}(\boldsymbol{\theta})|}. \tag{D.6}$$

We calculate the FIM using Eq. (2.10):

$$\mathcal{I}_{\mu\nu} = -\sum_{\boldsymbol{S}} \left(\partial_\mu \partial_\nu \log \mathcal{L}(\boldsymbol{S} \mid \boldsymbol{\theta})\right) \mathcal{L}(\boldsymbol{S} \mid \boldsymbol{\theta}) \tag{D.7}$$

$$= \frac{1}{\mathcal{Z}(\boldsymbol{\theta})} \partial_\mu \partial_\nu \mathcal{Z}(\boldsymbol{\theta}) - \frac{1}{\mathcal{Z}^2(\boldsymbol{\theta})} \partial_\mu \mathcal{Z}(\boldsymbol{\theta}) \partial_\nu \mathcal{Z}(\boldsymbol{\theta}), \tag{D.8}$$

where $\mathcal{Z}(\boldsymbol{\theta})$ is the partition function defined in Eq. (4.27). The ranges of parameters sampled are shown in Fig. D.2.

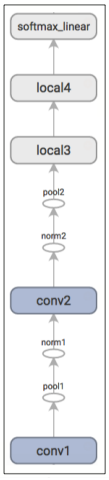


Figure D.3: **Convolutional neural network** used to classify MNIST handwritten digits. Image taken from TensorFlow tutorials [1].

D.3 Convolutional Neural Network with TensorFlow

A two-layer convolutional neural network was constructed in Section 4.6.4 using TensorFlow [1]. The outputs were converted to a probabilities using SoftMax [20]. A schematic of the network is shown in Fig. D.3.

The outputs of the network are turned into probabilities using SoftMax, which effectively treats the output weights as negative energies in a Bolzman distribution. Specifically, if x_i is the vector of network outputs, the probability is given as:

$$\mathcal{L}(x_i) = \frac{e^{x_i}}{\sum_j e^{x_j}} \tag{D.9}$$

D.4 Cosmic Microwave Background

We turn our attention to CMB spectra[1], and describe the manifold which represents all theoretically possible spectra generated by the 6 parameter ΛCDM model visualized in Section 4.6.5. The anisotropy in the CMB can be characterized by a 2×2 direction dependent intensity matrix $I_{ij}(\hat{n})$ whose components can be recognized as 3 of the 4 Stokes parameters. By expanding the various components of the intensity matrix into spherical harmonics, we obtain 3 maps of interest; the temperature fluctuation map T and 2 polarization maps, E and B. These can be expanded into spherical harmonics,

$$X(\hat{n}) = \sum_{\ell m} a^X_{\ell m} Y_{\ell m}(\hat{n}) \quad \text{where } X = T, E, B. \tag{D.10}$$

The angular spectra are defined as the cross correlation of the coefficients in the expansion, written as

$$C^{XY}_\ell \equiv \frac{1}{2\ell + 1} \sum_m \left\langle a^X_{\ell m} a^Y_{\ell m} \right\rangle \quad \text{where } X, Y = T, E, B. \tag{D.11}$$

Using this, we can construct a correlation matrix for the fluctuations,

$$C_\ell = \begin{pmatrix} C^{TT}_\ell & C^{TE}_\ell & 0 \\ C^{TE}_\ell & C^{EE}_\ell & 0 \\ 0 & 0 & C^{BB}_\ell \end{pmatrix}, \tag{D.12}$$

where the C^{TB}_ℓ and C^{EB}_ℓ vanish for symmetry reasons [12]. The values of C_l are parameter dependent, and a likelihood analysis of CMB data fit with such a correlation has been extensively studied, particularly in the case of limited sky coverage, as it is invaluable for fitting CMB measurements [115, 64, 139].

[1] Part of this writeup was initially presented as part of an A-exam questions for James Sethna and Julia Thom.

In the case of perfect full sky coverage we can decompose the measured fluctuations into spherical harmonics, and obtain coefficients $\hat{a}_{\ell m} = \{\hat{a}^T_{\ell m}, \hat{a}^E_{\ell m}, \hat{a}^B_{\ell m}\}$. These are expected to be (approximately) Gaussian, with mean zero. The probability of a fit for this data can be expressed as a huge product of high-dimensional Gaussians:

$$\mathcal{L}(\{\hat{a}_{\ell m}\} \mid \theta) = \prod_{\ell m} \frac{1}{\sqrt{(2\pi)^3 |C_\ell|}} \exp\left(-\frac{1}{2}\hat{a}^\dagger_{\ell m} C_\ell^{-1} \hat{a}_{\ell m}\right). \tag{D.13}$$

This conditional probability defines the likelihood function [115, 139], and so we can generate the FIM. Equation (2.10) can be re-written as

$$\mathcal{I}_{\alpha\beta}(\theta) = -\int \left[\partial_\alpha \partial_\beta \log \mathcal{L}(\theta \mid \mathbf{x})\right] \mathcal{L}(\theta \mid \mathbf{x}) dx. \tag{D.14}$$

We can use the definition of $\mathcal{L}$ for CMB fitting and perform the integral over all $\{\hat{a}_{\ell m}\}$. We begin by looking at the second derivatives of $\mathcal{L}$:

$$\begin{aligned} -\partial_\alpha \partial_\beta \log \mathcal{L}(\theta \mid \{\hat{a}_{\ell m}\}) &= \frac{1}{2} \sum_{\ell m} \partial_\alpha \partial_\beta \left(\log |C_\ell| + \hat{a}_{\ell m} C_\ell^{-1} \hat{a}_{\ell m}\right) \\ &= \frac{1}{2} \sum_{lm} \left(\frac{\partial_\alpha \partial_\beta |C_\ell|}{|C_\ell|} - \frac{\partial_\alpha |C_\ell| \partial_\beta |C_\ell|}{|C_\ell|^2}\right) + \frac{1}{2} \sum_{\ell m} \hat{a}_{\ell m} \partial_\alpha \partial_\beta C_\ell^{-1} \hat{a}_{\ell m}. \end{aligned} \tag{D.15}$$

We can plug this expansion into Eq. (D.14) and pull out all terms independent of the data. Thus, the first 2 terms in the sum can be completely pulled out of the integral. The remaining term is harder, and to evaluate it we make use of the following integral for symmetric, positive definite $M \times M$ matrix A and symmetric $M \times M$ matrix B

$$\sqrt{\frac{|A|}{(2\pi)^M}} \int \mathbf{x}^T B \mathbf{x} \exp\left(-\frac{1}{2}\mathbf{x}^T A \mathbf{x}\right) dx = \mathrm{Tr}(A^{-1}B). \tag{D.16}$$

This allows us to solve Eq. (D.14), setting $A = C_\ell^{-1}$ and $B = \partial_\alpha \partial_\beta C_\ell^{-1}$. We can now combine all the pieces together, and obtain a formula for the FIM

$$\mathcal{I}_{\alpha\beta}(\theta) = \sum_\ell \frac{2\ell+1}{2} \left(\frac{\partial_\alpha \partial_\beta |C_\ell|}{|C_\ell|} - \frac{\partial_\alpha |C_\ell| \partial_\beta |C_\ell|}{|C_\ell|^2}\right) + \sum_\ell \frac{2\ell+1}{2} \mathrm{Tr}\left(C_\ell \partial_\alpha \partial_\beta C_\ell^{-1}\right). \tag{D.17}$$

We can compare this to to previous results for FIM derivations, [112, 139] and confirm that we obtain the same result.

To determine the eigenvalue spread presented in Fig. 2.3, the FIM was calculated numerically using CAMB software and code provided by Francesco De Bernardis.

To visualize the model manifold in Fig. 4.8, we use Eq. (4.7) to compute the cosine-angle between two distributions:

$$\begin{aligned}
\langle\theta_1;\theta_2\rangle &= \int d\{\hat{a}_{\ell m}\}\sqrt{\mathcal{L}(\{\hat{a}_{\ell m}\}\mid\theta_1)}\sqrt{\mathcal{L}(\{\hat{a}_{\ell m}\}\mid\theta_2)} \\
&= \prod_{\ell m}\int d\hat{a}_{\ell m}\frac{1}{\sqrt{(2\pi)^3}|C_{(1)\ell}|^{1/4}|C_{(2)\ell}|^{1/4}}\exp\left(-\frac{1}{2}\hat{a}^{\dagger}_{\ell m}\left(\frac{C^{-1}_{(1)\ell}+C^{-1}_{(2)\ell}}{2}\right)\hat{a}_{\ell m}\right) \\
&= \prod_{\ell}\left(\frac{\left|C^{-1}_{(1)\ell}+C^{-1}_{(2)\ell}\right|^{-2}}{2^{2d}|C_{(1)\ell}||C_{(2)\ell}|}\right)^{\frac{2\ell+1}{4}}
\end{aligned} \tag{D.18}$$

where d reflects the dimension of C_ℓ.

To sample the model manifold, and MC sampling was performed around the best-fit provided by the Planck 2015 data release. The probability of accepting/rejecting a step was determined by the Bhattacharyya distance to the best-fit spectra.

APPENDIX E

INPCA COMPARISONS WITH T-SNE AND DIFFUSION MAPS

We provide a detailed comparison of the Ising model manifold and the outputs of a convolutional neural network trained on the dataset of MNIST handwritten digits using three techniques: (1) the InPCA algorithm developed in Chapter 4, (2) t-SNE [98], and (3) diffusion mapping [31]. Importantly, because t-SNE and diffusion maps are purely visualization techniques that require a distance to be input, we supply our intensive distance to all three methods for consistency and ease of comparison.

E.1 Ising Model Manifold

In this section, we provide a detailed comparison of the Ising model manifold discussed in Section 4.6.3. We look at the model manifold for a 2×2 system with the parameter ranges discussed Section 4.2 and Section D.2. Furthermore, because of the simple nature of manifold (we vary two parameters, external field and nearest-neighbour coupling) two-dimensional visualizations from each method are effectively equivalent. For this reason, we consider the first three components (*i.e.* when the Minkowski-like nature of InPCA has a significant effect).

Fig E.1(a) shows the manifold as visualized with InPCA. Note that the third component (z-axis) is imaginary. In this way, InPCA embeds the manifold in a Minkowski-like space. Because two parameters are varied (field and nearest-neighbour coupling) the manifold is two dimensional, a property that is extracted by InPCA. The t-SNE visualization of the Ising manifold is shown in

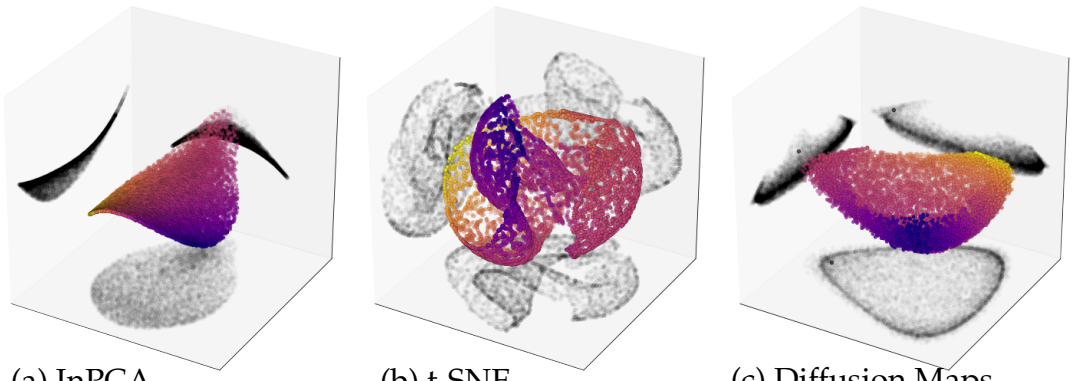
Fig. E.1(b). This embedding technique is best used to reveal clusters and local features, but it fails to fully represent the global features of the manifold. Diffusion maps is used to visualize the Ising manifold in Fig. E.1(c). The two-dimensional nature of the manifold is also revealed through this visualization, however it is still embedding the manifold in a Euclidean space. We see a 'curling' at the edges of the manifold, as the diffusion maps appear to be struggling to capture the important property of large positive/negative fields being very far apart from each other.

How can this visualization be useful? Figure E.1 illustrates the family of behaviors exhibited by Ising models, and could be coarse-grained by sampling a sub-grid of spins in a larger Ising model. The renormalization group tells us that this coarse-grained model can be rescaled to match the original model at renormalized parameters; distance in the intensive metric embedding could be a systematic, principled way of matching these parameters. This is the focus of ongoing research.

E.2 Neural Network

In this section, we provide a detailed comparison of the outputs of a trained neural network constructed with TensorFlow [1]. The outputs are viewed as probabilities through SoftMax [20]. For a well-trained newtork, one expects the outputs to form clusters. Specifically, the number of clusters is a reflection of the number of categories imposed on the network. Here, we have 10 digits, and so we expect any visualization method to reveal 10 clusters.

Figure E.2 shows the outputs visualized with the thee manifold learn-



(a) InPCA (b) t-SNE (c) Diffusion Maps

Figure E.1: **Manifold of the Ising model** visualized with different manifold learning techniques. In all figures, axes reflect aspect ratio of extracted components and colors reflect external magnetic field matching the main text. The manifold is two intrinsically dimensional, as we only vary two parameters (field and nearest-neighbour coupling). (a) InPCA visualization with fist three extracted components. Note that the third component is imaginary. (b) t-SNE visualization in three dimensions and (c) fist three components of diffusion maps.

ing methods, and each method shows clusters. The t-SNE visualization in Fig. E.2(b) shows the cleanest clusters. It is important to note, however, that t-SNE is optimized for local features (and so the distances between clusters is not meaningful). For instance, the digits 6 and 7 are very different, and so should be considered quite distinct in this picture. However, t-SNE places the clusters of 6's and 7's right next to each other. InPCA and diffusion maps have similar visualizations (shown in Fig. E.2(a) and Fig. E.2(c))).

Because InPCA captures global features (as shown in the large distance between 6's and 7's), it will not artificially cluster points and so it is useful for comparing the outputs for trained vs. untrained networks (as shown in the Fig. 4.7).

How can such geometries be useful? By using InPCA to better understand

the global geometry of an initialized neural network, as well as after the first couple epochs, properties of the network can be analyzed (*e.g.* what clusters emerge first? Is there a hierarchical structure?) and is the focus of onging research.

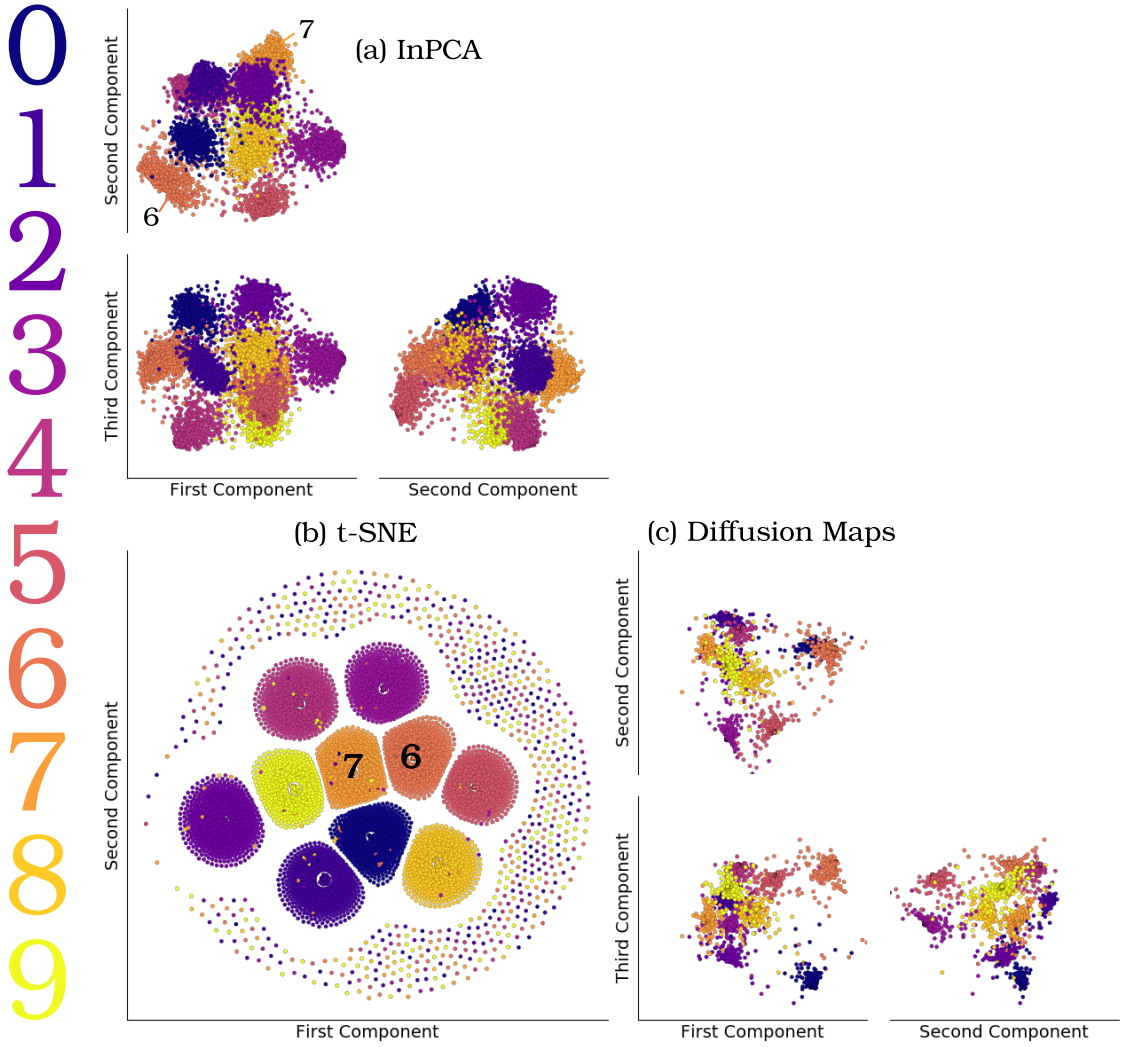


Figure E.2: **Outputs of the trained neural network** visualized with different manifold learning techniques. In all figures, axes reflect aspect ratio of extracted components and colored by digit matching the main text. All projections reveal underlying clusters in the outputs, with one cluster per digit, a reflection of the training of the neural network. (a) InPCA visualization with fist two extracted components. (b) t-SNE visualization. (c) First two components from diffusion maps. t-SNE produces the cleanest visualization of the clusters, however it is important to note that global features are note meaningful. For instance, the digits 6 and 7 are very different. InPCA shows these two clusters as far apart, whereas t-SNE has them next to each other, because InPCA preserves global features.

APPENDIX F

PARAMETER DEGENERACIES AND COMBINATIONS IN THE CMB

There are parameter degeneracies in the CMB spectra which make fitting to data particularly difficult (when solely considering CMB spectra for determining parameters). As a preliminary consideration of this effect on the visualization of the model manifold with InPCA (from Chapter 4), we consider $A_p = A_s e^{-2\tau}$ [140], a degeneracy in the primordial fluctuation amplitude and the optical depth at reionization that affects the amplitude of the CMB spectra. Figure F.1(a) shows how A_p correlates with the first component of InPCA (the first component shows a strong correlation with the primordial fluctuation amplitude A_s, as shown in Fig. 4.9, and so serves as a motivation for this correlation). We see that the Pearson correlation between A_p and the first parameter is $r = 0.98$ [72], indicating a near perfect correlation: the biggest feature that InPCA appears to extract in CMB spectra (from the parameter ranges in Section 4.6.5) is the overall size of the fluctuations. We visualize the manifold of possible CMB spectra using the first two InPCA components in Fig. F.1(b).

Next, we consider[1] the combination of matter density ($\Omega_m = \Omega_b + \Omega_c$), and the reduced Hubble constant (h) given as $\Omega_m h^3$. The constraint on $\Omega_m h^3$ is very tight (as compared to orthogonal directions) [78]. We show the correlation between $\Omega_m h^3$ and the second InPCA component (which is an orthogonal direction in prediction space to the one associated with A_p, and appears correlated with the Hubble constant in Fig. 4.9). Figure F.1(b) shows the correlation, with a Pearson coefficient of $r = 0.83$, indicating a strong correlation [72].

[1] Investigating the connection between $\Omega_m h^3$ and InPCA components was initially suggested to KNQ by David Spergel at the 2019 Aspen winter conference.

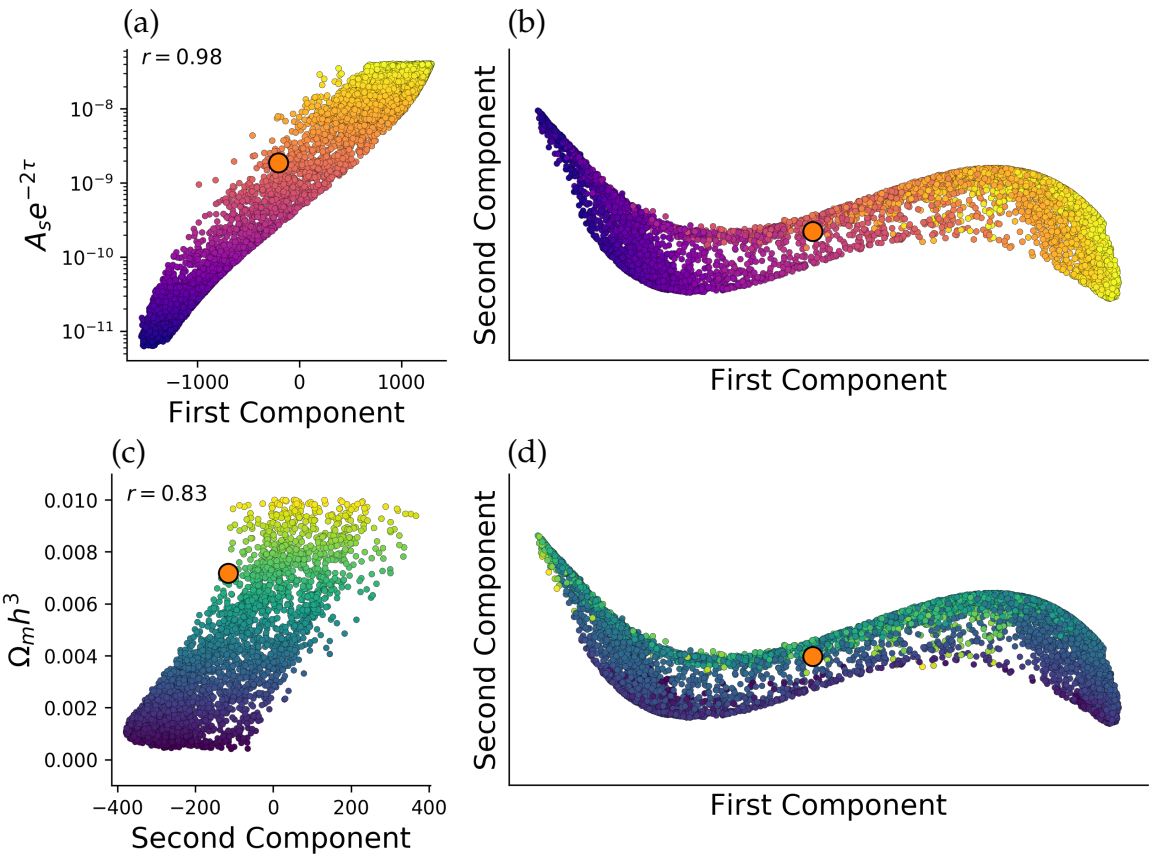


Figure F.1: **The parameter combination correlations with InPCA components** for different CMB predictions of the cosmology model, with the Pearson coefficient to determine the significance of correlations. In all figures, orange dots represent our universe. (a) Shows the correlation of $A_p = A_s e^{-2\tau}$ with the first component, and (b) colors the manifold visualization by A_p. (c) Shows the correlation between $\Omega_m h^3$ and the second component, and (d) shows the model manifold colored in this way.

How can such geometries be useful? By adding new parameters to the model, which are not well known, the full model manifold can be explored and an understanding of the geometry allows for complex non-linearities and parameter degeneracies to become manifest without the need for heuristics. Furthermore, models can be expanded to include foreground features and other properties which influence the observed spectra. Again, this provides a system-

atic way of exploring the impact on fitting to data.

APPENDIX G

STATISTICAL TESTS FOR QUANTIFYING BEHAVIOUR

In this appendix, we provide supporting calculations, examples and figures for material presented in Chapter 5.

G.1 Error Estimates

To obtain the standard error on the fraction of a population (such as in Table 5.1 or Fig. 5.8), we used the following:

$$\texttt{Err}(p, N) = \sqrt{\frac{p(1-p)}{N}} \tag{G.1}$$

where p is the fraction of the population, and N is the size of the total population.

G.2 Statistical Tests

To compare distributions for populations of varying sizes, we performed chi-squared tests on the contingency tables constructed from the total numbers. Note that a series of pair-wise comparisons would be inappropriate in this case, as the different measures are correlated for normalized distributions (*e.g.* since they are all normalizable, if one measure goes up then another must go down).

As an illustration of this method, consider the observation protocol described in Section 5.2.2. A sample graph of the accumulated codes for two observers in a traditional lab section is presented in Fig. G.1. The contingency table constructed from these observations is given by Table G.1. Because the

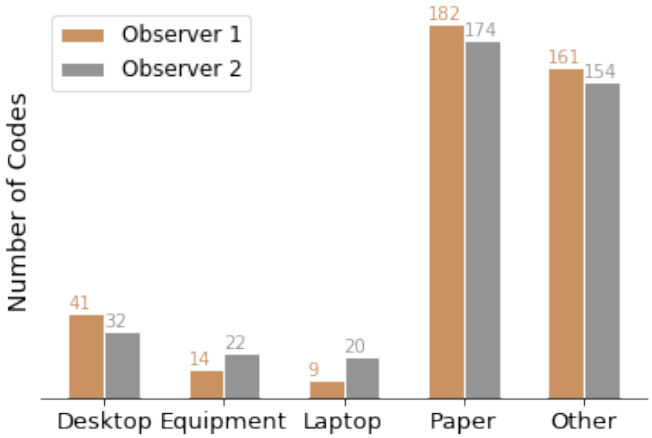


Figure G.1: **Bar plot of code counts from two observers** used to form the basis of a chi-squared test to validate the observation protocol used in the behaviour study from Chapter 5. Two observers documented the same lab period, and the resulting contingency table (given by the raw counts displayed on the graph and shown in Table G.1) was used to determine statistical validity of the method. Here, the two distributions are statistically indistinguishable indicating that the observers captured the same distribution of student actions.

two distributions are statistically indistinguishable, the observers captured the same distribution of student actions.

Table G.1: **Sample contingency table** used to determine if two distributions are statistically different. Two observers documented the same lab period, and a chi-squared test was performed to determine if the resulting distributions are statistically similar or dissimilar. Here, we obtain $p > 0.1$, indicating that the observers captured the same distribution of student actions.

Observer	Desktop	Equipment	Laptop	Paper	Other
1	41	14	9	182	161
2	32	22	20	174	154

As a second illustration of this method, consider the cluster compositions presented in Fig. 5.8. To determine if the distribution of men's profiles in the

inquiry lab were statistically different for mixed-gender versus single-gender groups, we performed a chi-squared test on the contingency table representing the constructed from the number of profiles in each cluster. This table is shown in Table G.2. When comparing these two distributions, we obtained $p = 0.007$, indicating that the two distributions are different: men behaved differently when with men as compared to when they were with women.

Table G.2: **Sample contingency table** showing the distribution of men's profiles in the inquiry lab for mixed- versus single-gender groups. Here, we obtained $p = 0.007$, indicating that the two distributions are significantly different.

Group Type	Desktop	Equipment	Laptop	Paper	Other
Mixed-Gender	17	6	20	0	42
Single-Gender	38	31	45	0	48

G.3 Example of Student Profile Rescaling

To perform a cluster analysis on multidimensional data with k-means, each measure needs to be on a comparable scale. The raw data collected for the study in Chapter 5 was gathered by coding the student's action every five minutes, using printed sheets as shown in Fig. G.2. The quantified behaviour of students measured in this way needed to be rescaled for two reasons in order to satisfy the criteria for clustering.

First, as shown in Fig. 5.3, each of the five observation codes appears to be on different scales, in the sense that they are spread out by different amounts. For instance, the Other code distribution is much wider than that for Equipment.

(a) Traditional Lab

(b) Inquiry Lab

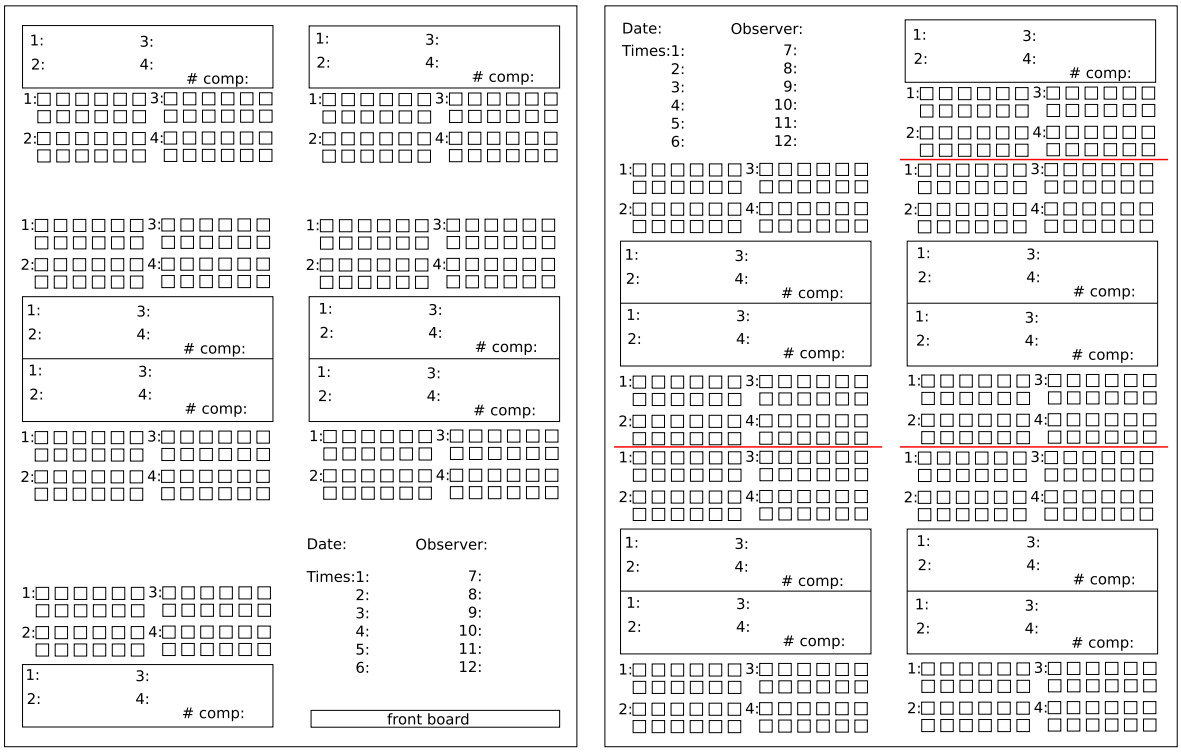


Figure G.2: **Coding sheets** used by observers in this study (scaled for illustrative purposed to fit in figure), for students in (a) the traditional labs and (b) the inquiry labs. Observers mark student names/descriptors at each table, and use the boxes to mark D, E, L, P, O every five minutes to code student behaviors.

Second, students spent very different amounts of time in the lab, as shown in Fig. G.3.

A sample of gathered data for an example student is shown in Fig. G.4(a). Once all the codes were collected over the course of the lab period, they were collapsed to form a normalized distribution representing the student profile in Fig. G.4(b) to account for the large variation in times students spent in lab (shown in Fig. G.3). Finally, once all the data was collected, student profiles were grand mean scaled so that each measure (Desktop, Equipment, Laptop,

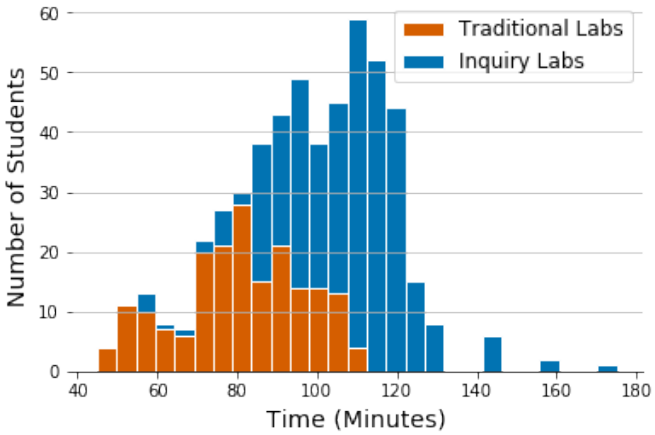


Figure G.3: **Stacked histogram of the amount of coded time students spent in lab**, broken down by lab type. The time students spent is highly variable, from less than 45 minutes to over 175 minutes. Part of this was due to difference in lab type, as students in the traditional labs spent on average 80 ± 6 min, whereas students in the inquiry labs spent on average 107 ± 6 min.

Paper, Other) had mean 0 and standard deviation 1. In this way, naturally different measures (such as Other and Equipment) could be compared on the same scale. The student in Fig. G.4 was in a control lab, reflected in the large Paper measure. They were also assigned to be in the Paper cluster, a reflection of the fact that the Z-score for Paper is the highest.

G.4 Effect of Student Group Sizes

We note that group sizes in the two labs are were the same. Groups in the traditional and inquiry labs were of of varying sizes, as shown in Fig. G.5. Groups in the inquiry labs typically had three or four students, whereas groups in the inquiry labs typically had two or three members. One could expect that, in

(a) Table of Codes

time	0 min	5 min	10 min	15 min	···
code	P	E	P	O	···

(b) Distribution of Codes (c) Z-Scores

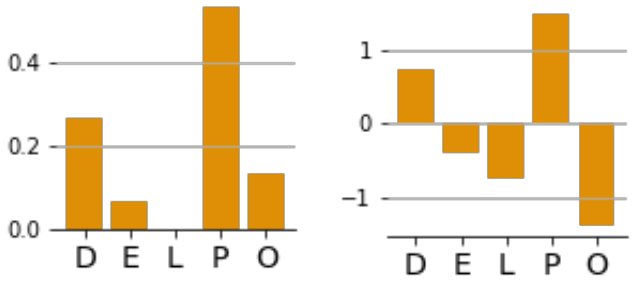


Figure G.4: **Sample student profile** illustrating the two-step rescaling process performed on the raw data to generate student profiles before clustering. (a) Shows a portion of the raw table, where the student's action was documented every five minutes. (b) The collected codes from the course of the entire lab period were collapsed together to form a normalized distribution. (c) Each measure (**D**esktop, **E**quipment, **L**aptop, **P**aper, **O**ther) was grand-mean scaled across all students (so that each measure has mean 0 and standard deviation 1 when averaged over all students).

groups with more members, there is an increased chance of task division occurring. While groups in the traditional labs typically had more members than those in the inquiry labs, Fig. 5.7 in fact shows proportionally fewer groups in the inquiry labs with members in identical clusters, supporting the conclusion that groups in the inquiry labs were more likely to divide tasks.

Figure G.5: **Stacked histogram of the number of groups** with two, three, four or five members, broken down by lab type. Students in the inquiry labs were predominantly in groups of two or three, whereas groups in the traditional labs had three or four members.

APPENDIX H

DIFFERENT CLUSTER ANALYSIS

In this appendix, we sub-divide student profiles from Chapter 5 by lab type and gender prior to clustering to test the robustness of the extracted clusters and insure the validity of performing a cluster analysis on all profiles simultaneously. We consider four student groupings, and perform independent cluster analysis on each following the method described in Section 5.2.4 and Section 5.2.5:

1. Only women's profiles
2. Only men's profiles
3. Only profiles from students in the traditional labs
4. Only profiles from students in the inquiry labs

If the clusters extracted from each of the student groupings is counter to the clusters we found in Chapter 5, then we run the risk of imposing the cluster breakdown of a dominant group (as defined by the group with the most student profiles) onto other groups, thereby obscuring important behaviour differences. We found no significant difference between extracted clusters in all four groupings, and so we perform a single cluster analysis on all student profiles simultaneously in Chapter 5.

We begin by generating an elbow plot for each of the four groupings, to independently determine the optimal number of clusters in each case. Importantly, we rescale student profiles *for each grouping independently* (see Section 5.2.4 for a description of Z-scores and rescaling student profiles). The results of this are

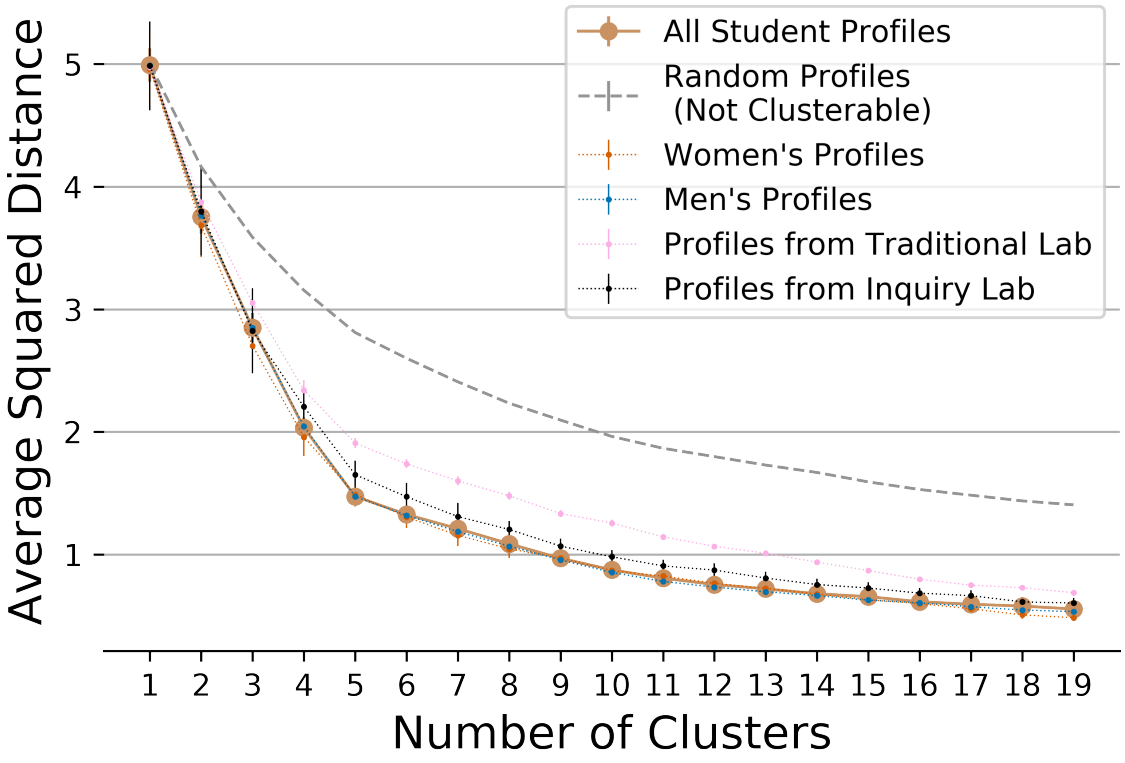


Figure H.1: **Elbow plot** of different groupings of student profiles, testing the robustness of the five clusters. All student profiles (thick, brown line) have a kink at five clusters, as does extracted clusters from considering (1) only women's profiles (red, dotted line), (2) only men's profiles (blue, dotted line), (3) only student profiles from the traditional lab (pink, dotted line) and (4) only student profiles from the inquiry lab (black, dotted line). In all cases, there is a kink at five, indicating that the optimal number of cluster for all cases is five. Moreover, all groupings are well below random, which displays no kink (grey, dashed line).

shown in Fig. H.1. In all cases, the optimal number of clusters is five. For ease of comparison, we superimpose on the figure the plot for all student profiles from Fig. 5.5, which indicate substantial agreement with the initial analysis, as well as the plot for randomly generated profiles to show that the clusters for all groupings are significantly different from random.

Next, we analyse the centers of all five cluster for each of the four student groupings, to characterize each cluster. The results are shown in Fig. H.2. For ease of comparison, we show in Fig. H.2(a) the cluster centers extracted for all student profiles (identical to the results presented in Fig. 5.6. Note that they are all characterized by high engagement in a particular activity (*i.e.* high equipment use). Figure H.2(b) shows the cluster centers when considering only women's profiles, and Fig. H.2(c) shows the cluster centers when considering only men's profiles. In both cases, center peaks align with the five codes, in agreement with the original analysis. Furthermore, the Z-score characterizing the peak in each center matches in magnitude with the original analysis.

Figure H.2(d) shows the cluster centers when only considering profiles from students in the traditional labs, and Fig. H.2(e) shows the cluster centers when only considering profiles from students in the inquiry labs. Again, we see that the centers align with those of the original analysis, falling along the five observation codes. However, there appears to be an important feature characterizing the center of the Laptop cluster in Fig. H.2(d) and the Paper cluster in Fig. H.2(e). Specifically, the Z-score characterizing the peak for these two centers is significantly larger in both cases than it is in Fig. H.2(a). This is because students in the traditional labs fill out paper worksheets, and nearly all students never handle a laptop or personal devise. However, a non-zero number of students did occasionally use their laptop in lab, and so this skews the resulting Z-score for those students (because it is scaled by the standard deviation, which in this case is very small). Similarly, students in the inquiry labs use electronic notebooks via laptops and personal devices (and the lab desktop computer) and so most never use paper. However, a non-zero number of students still brought a notepad to lab to write on, and so highly skews the resulting Z-scores for paper,

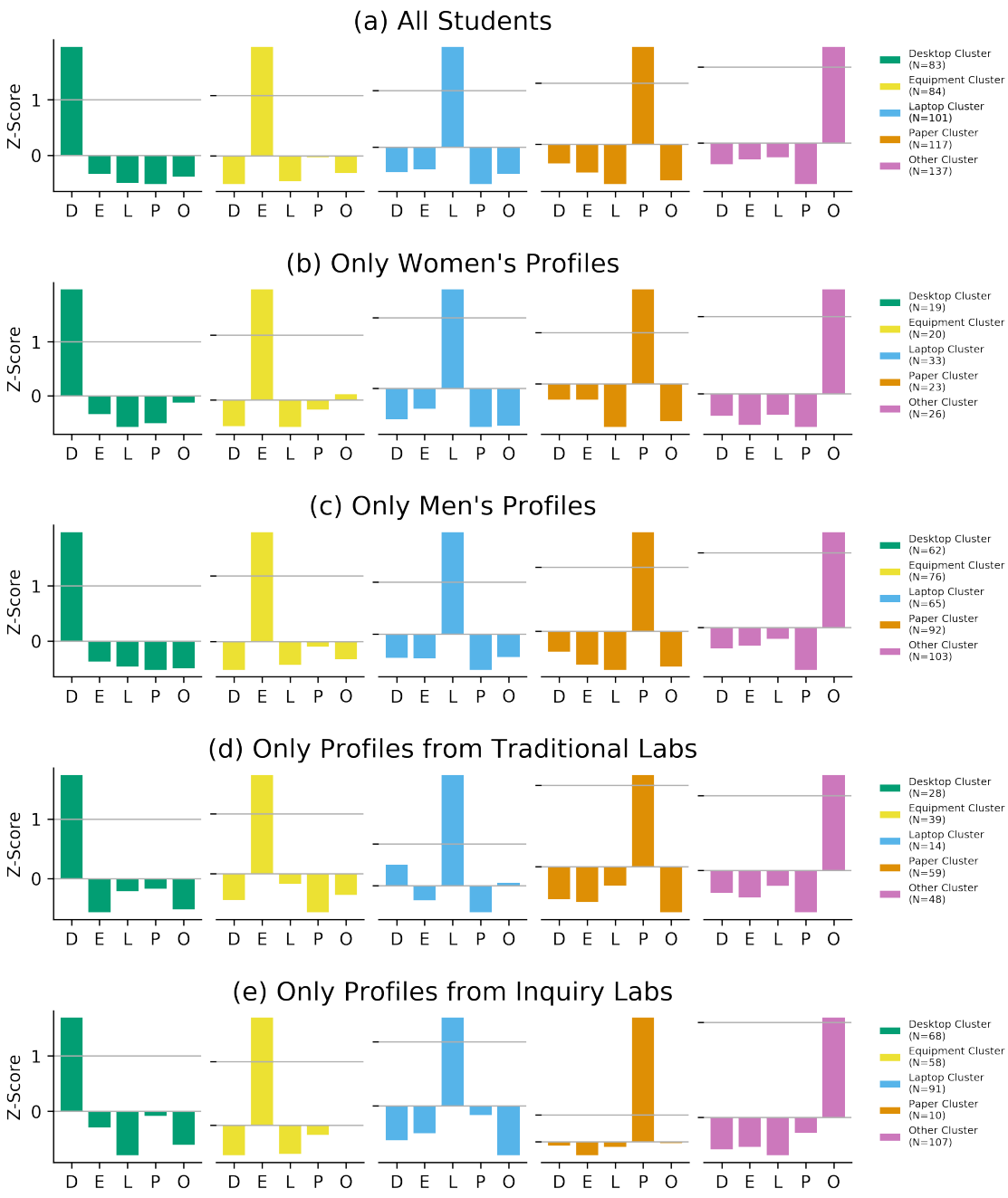


Figure H.2: **Cluster centers** of different groupings of student profiles, testing the robustness of the five clusters. (a) All student profiles have five clusters, matching the five codes used in observation. We then consider (b) only women's profiles, (c) only men's profiles, (d) only student profiles from the traditional lab and (e) only student profiles from the inquiry lab. In all cases, cluster centers match up with observation codes, indicating that in all groupings the same clusters are present.

again because Z-scores are scaled by standard deviation, which is very small in this case.

Because the extracted clusters are the same regardless of student grouping, in that they all line up with the five codes used in the observation protocol and therefore all characterize high engagement in a particular activity, we perform a single cluster analysis on all student profiles in Chapter 5. Because we are interested in comparing demographic composition of the resulting clusters (*e.g.* who are the high equipment users?), it is necessary to include all students in the Z-score rescaling. In this way, we can more transparently highlight important behavioural differences between students, since they are all on the same scale.

APPENDIX I

INPCA VISUALIZATION OF STUDENT BEHAVIOURS

In this appendix, we visualize the collection of student profiles and reveal underlying geometric structure in the data from Chapter 5, using both the qualitative method in the original analysis and quantitative method from Chapter 4. We use t-SNE in Fig. I.1(a) to qualitatively visualize the data, in a way which shows primarily the underlying clusters. Figure I.1(b) shows the data visualized using InPCA (only using the first two components), which preserves global and local features, and reveals additional underlying features in the data. The five main clusters follow the five codes, and the average profile from each cluster[1] is shown in Fig. I.1(c).

To quantitatively visualize the collection of student profiles using InPCA, we first perform an eigenvalue decomposition on the cross covariance matrix from the data (see Section 4.4.1) to determine the optimal number of components to keep. From Fig. I.2, we see that the first four components are real and the third is imaginary. To capture 60% of the variation in the data, we use the first three components, and to capture over 85% we use the first five in Fig. I.4(a).

From InPCA, we see that there are dense regions, reflecting many profiles that are similar to one another, as well as less dense regions, reflecting profiles that are somewhat similar to only a few others. There are also three dense stripes, emanating from the central region. Note that, using these visualization methods, we see that the Other cluster reflects points in this central core, with the remaining clusters surrounding it. In particular, the Desktop, Laptop, and

[1]Here, we do not show the average Z-score such as in Fig. 5.6 but rather the average distribution. We do so to preserve the probabilistic nature of the data, so that it can be used with InPCA.

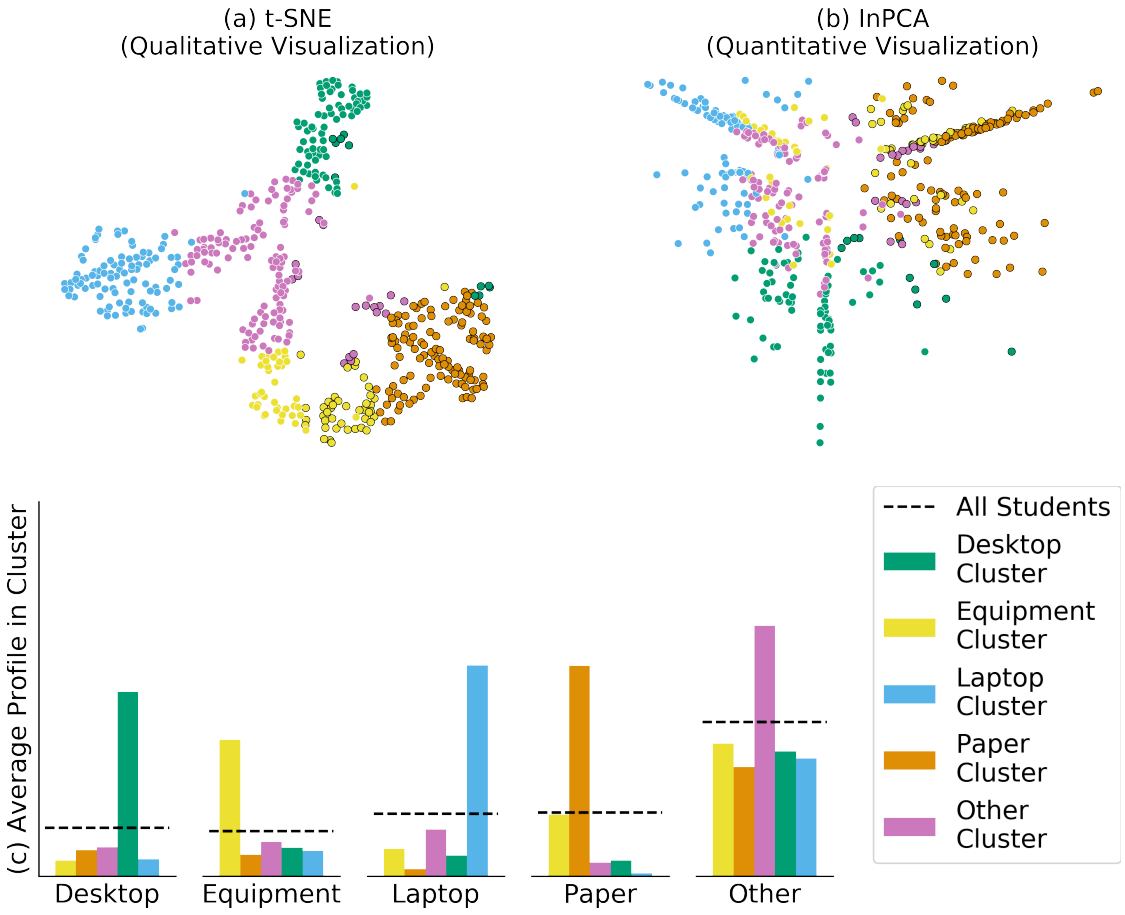


Figure I.1: **Visualization** of student profiles using (a) t-SNE, revealing underlying clusters in the data and (b) with InPCA, revealing additional structure. In particular, there are three dense stripes that intersect several clusters. Points are coloured by cluster, and edge-color reflects lab type (black border is a profile in the traditional lab, white border is a profile in the inquiry lab). (c) Average profile of each cluster, compared to overall student average (dashed line), for each of the five codes used (Desktop, Equipment, Laptop, Paper, Other).

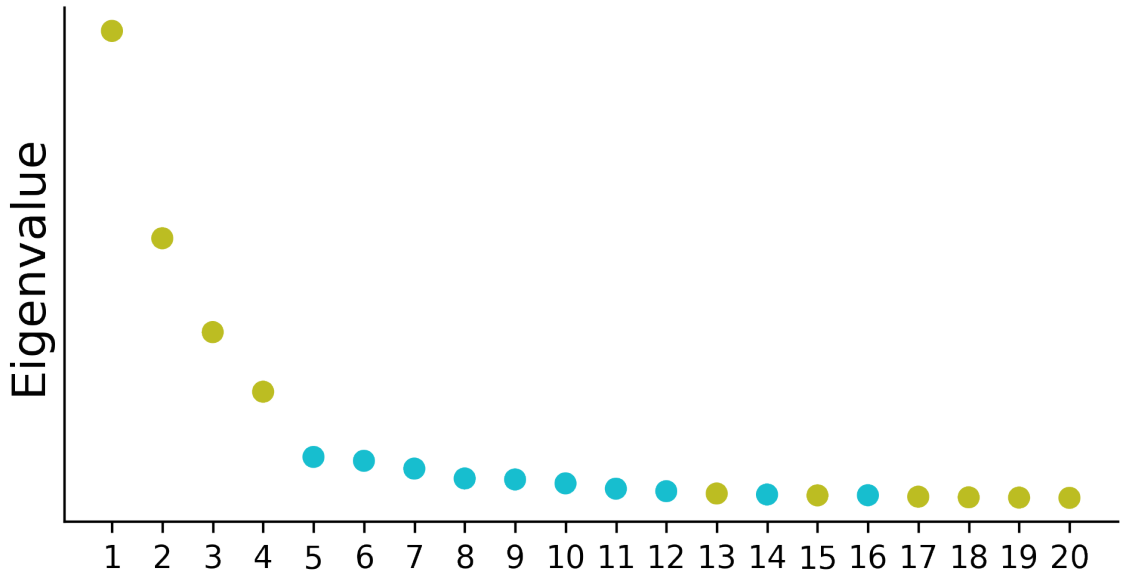


Figure I.2: **Ordered InPCA eigenvalues** of quantitative projections, revealing five dimensions needed to capture important features. Green points are positive eigenvalues, blue are negative. Because we need five dimensions to accurately visualize the data, we plot the orthogonal components in the triangle plot in Fig. I.4.

Paper clusters are highly distinct from one another and are located in very different regions.

To determine if these features are reflect data structure or measurement methods, we compare the features to 10,000 randomly generated student profiles. We generate each random profile through a uniform sampling of integers from 0 to 20 for four measures, and from 1 to 20 for one measure[2] that are then rescaled following the same protocol as in Section 5.2.4. We do so to see if the stripes/dense regions are a feature of the sampling, discretization, or 5-D nature of the data. Figure I.3(b) shows an InPCA visualization of random data, form-

[2]We ensure that one measure is nonzero, to ensure that no random profile is exactly orthogonal from another, otherwise they will be an infinite distance apart and impossible to visualize. The one measure that is sampled from 1 to 20 reflects the Other code in observations (nearly all students were observed doing other at least once).

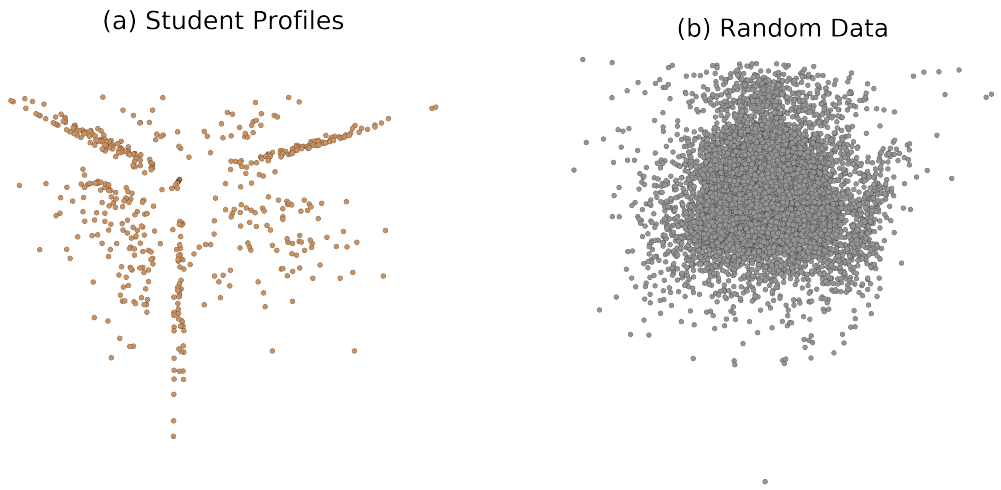


Figure I.3: **InPCA visualization of profiles** as compared to random data. (a) Real data collected in this study show structure. (b) Random profiles form a blob of points. Note that we are showing the first two projections here, and both are real. Aspect ratio reflects the actual ratio of projections.

ing a cloud of points with a dense central region. The first four components are real, and the fifth is imaginary. Note that the features of random data are dissimilar to our measured data in two important ways: (1) real data have dense stripes/jets emanating from the central region, whereas random data do not, and (2) random data have a dense central core, whereas our real data do not. There is one similar feature, namely that the fifth projected component in both cases is imaginary. Thus the imaginary fifth component shown in Fig. I.4(a) may be a reflection of the measurements method rather than a statement about the underlying data.

Confident that the stripes are a significant feature of the data (and not a reflection of the method of data collection), we use DBSCAN [47] to isolate and study the different regions of interest (stripes, dense regions). DBSCAN (density-based spatial clustering of applications with noise) is a clus-

tering method, where clusters are determined by connected simplices whose vertices are the data points. We identified nine regions of interest, by manually tuning the clustering parameters until the stripes were captured. Four of these regions contain profiles that are only in the traditional labs, and five of these regions contain profiles that are only in the inquiry labs: indicating that the most significant factor impacting student behaviour is the pedagogical structure of the labs. More regions are associated with the inquiry labs, reflecting the greater diversity of behaviour occurring in these labs.

About one third the student profiles (32.5%) are not associated with a particular region (*i.e.* they are not densely packed with other student profiles). We show the InPCA projections of student profiles colored by their region in Fig. I.4(a). The first InPCA component (largest, most important direction) directly lines up with the pedagogical structure of the lab. Students from the inquiry labs are in the negative direction of this component (left-hand side) whereas students in the traditional labs are in the positive direction of this component (right-hand side). From Fig. I.4(b), we see that all regions have profiles engaged in other activities to some degree. Interestingly, all regions also have students never engaging in a particular activity (*e.g.* no student in the purple region is observed ever handling a laptop or paper). The regions therefore reflect areas of "reduced dimension", in the sense that have 4 or less non-zero measures. For example, the blue region is described by students solely writing on paper or engaging in other activities to varying degrees (and who never handle the desktop computer, equipment, or a laptop or personal device).

To understand how the regions interact with the main clusters, we look at the average student profile in each region, shown in Fig. I.4(b). Two jets are

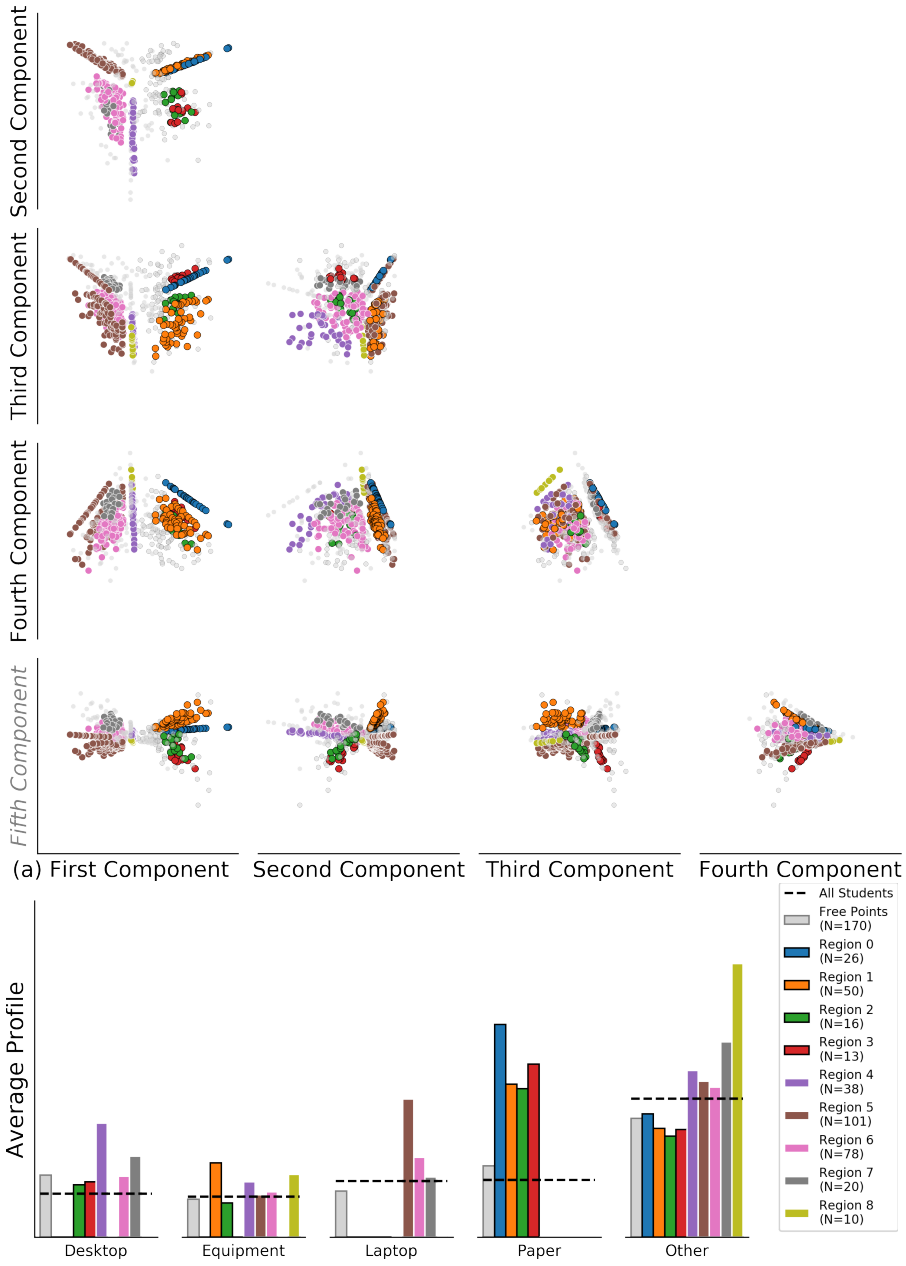


Figure I.4: **Quantitative visualization** of student profiles using (a) InPCA, revealing structure in the data. Note that the fifth projected component is imaginary. Points with black border are profiles from the traditional labs, and white border are points from the inquiry labs. (b) Average profile in each region, with light grey representing points not assigned a region, for each of the five codes used (Desktop, Equipment, Laptop, Paper, Other). Black border represents regions with profiles only from traditional lab, and white border represents regions with profiles only from inquiry labs. From (a), we see that the most dominant effect is lab type, with the inquiry and traditional labs occupying different ends of the first component.

solely associated with the inquiry labs, and one jet is solely associated with the traditional labs. Students along these jet appear to engage in two activities to varying degrees:

1. *The purple jet:* corresponds to students in the inquiry labs to handle the desktop, equipment, or engage in other activities more than average (and who never handle a laptop or personal device, nor write on paper).
2. *The brown jet:* corresponds to students in the inquiry labs who use their laptops or personal devices or engage in other activities more than average, and handle equipment, but never touch the desktop or paper.
3. *The blue jet:* corresponds to students in the traditional labs who write on paper the most, engage in other activities less than average, and who never touch the desktop, equipment, or a laptop or personal device.

We further explore this relationship by considering the overlap between clusters and regions, as shown in Fig. I.5. Here, we consider only profiles in a particular cluster, and show their average broken down by region. Figure I.5(a) shows the Desktop cluster, decomposed by overlapping regions. For all regions, student profiles have an above-average fraction of time spend on the desktop code. On of the jets in the data (see Fig. I.4(a)) is highlighted in purple, with a large number of profiles in the Desktop cluster. Note that the purple region overlaps with the Other cluster as well (Fig. I.5(e)), here with profiles that spend above-average amount of time engaged in other activities as well as desktop. Similarly, the brown jet reaches from the Other cluster, to the Laptop cluster, and a slightly overlaps the Equipment cluster, representing students to engage primarily in handling a laptop or other activities. Finally, the blue jet only goes from the Other to Paper cluster (and never overlaps with the other clusters).

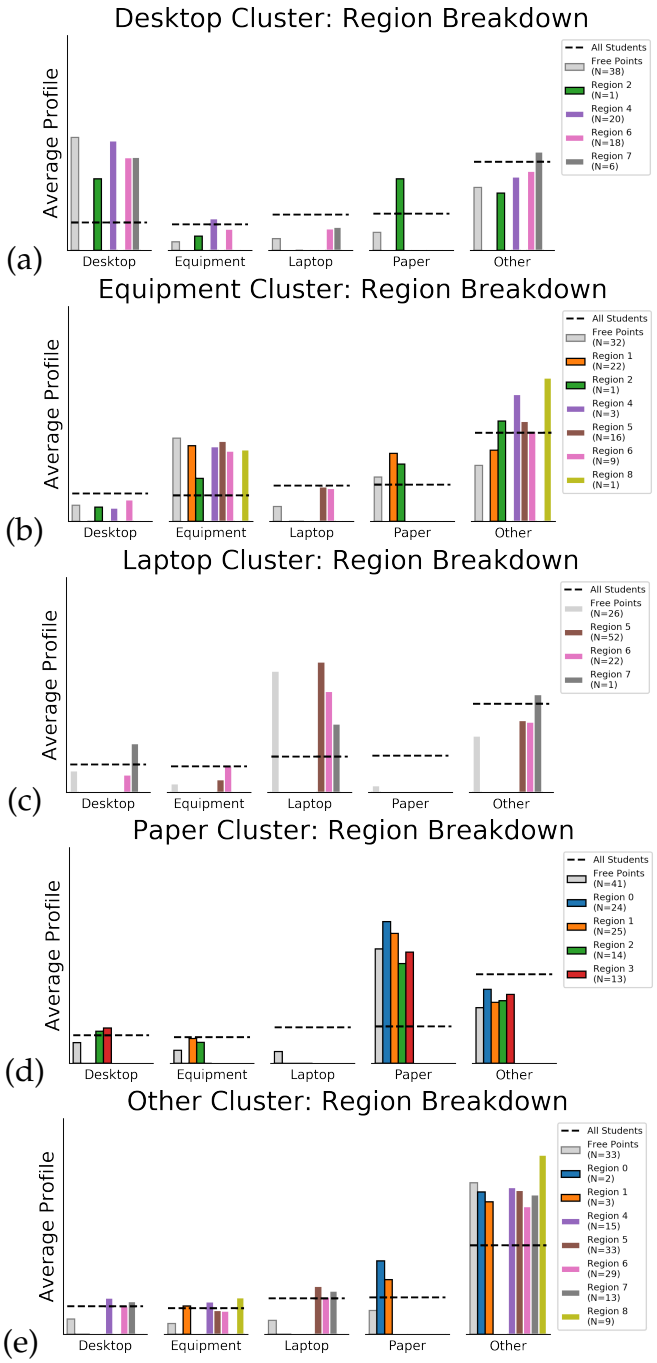


Figure I.5: **Cluster breakdown** of student profiles overlapping the regions extracted from Fig. I.4. Bars with black border are regions with profiles only in the traditional labs, and white border are regions with profiles only from the inquiry labs (light grey border represents free points that are not clustered). Almost all regions overlap with the Other cluster in (e), the central core of the distributions of profiles in Fig. I.1. The horizontal axis in each case represents the distribution over observation codes (Desktop, Equipment, Laptop, Paper, Other).

Understanding the demographic breakdown of these three jets, as well as less dense regions (representing students engaging in very dissimilar behaviour from their peers) is the focus of ongoing work.

APPENDIX J

VIDEO CODING

In this appendix, we show the coded time series for 23 decomposed student profiles (discussed in Chapter 5), grouped by lab type. Because of the gendered differences observed in the inquiry lab, as highlighted by the cluster analysis, only profiles from the inquiry labs were decomposed. Furthermore, emphasis was placed on understanding roles in mixed-gender groups, and so the majority of groups analyzed through single-group video are mixed gender. The importance of social dynamics in single-gender groups is the focus of future work.

The detailed time series for each lab appears very qualitatively different for each group for a couple reasons. First, each week students were engaged in different activities. Fig. J.1,J.2, J.3,J.4 show the time series for students performing the Bouncing Ball experiment. Note that in all figures, a large portion of the lab at the beginning is dedicated to whole-class discussions and an invention activity involving whiteboards. Fig. J.5 shows the time series for students performing the Hooke's Law experiment. Fig. J.6,J.7 sow the time series for students in the Pendulum experiment. Fig. J.9,J.8 show the time series for students in the final Project Lab section, where they have designed their own experiment.

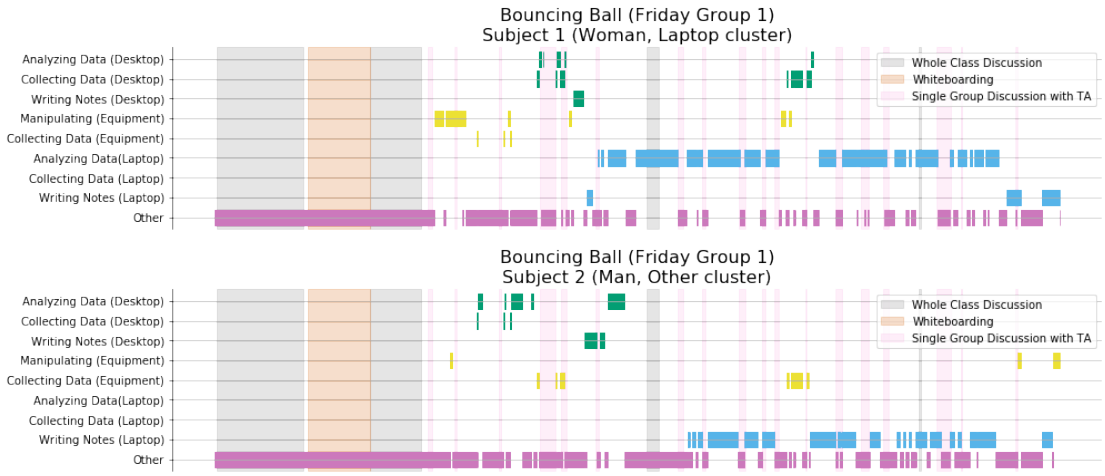


Figure J.1: **Decomposed student profiles** of two students in the same group during the Bouncing Ball experiment, with the x-axis representing time.

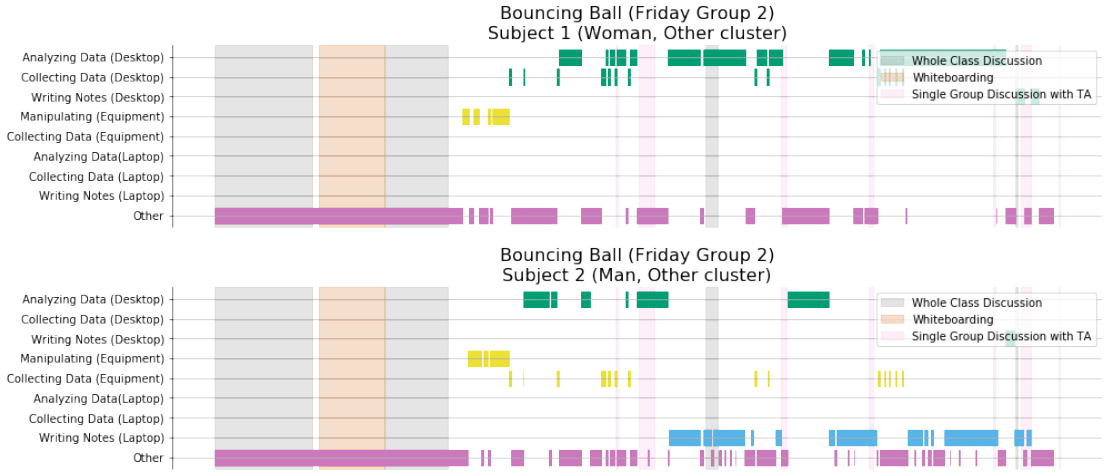


Figure J.2: **Decomposed student profiles** of two students in the same group during the Bouncing Ball experiment, with the x-axis representing time.

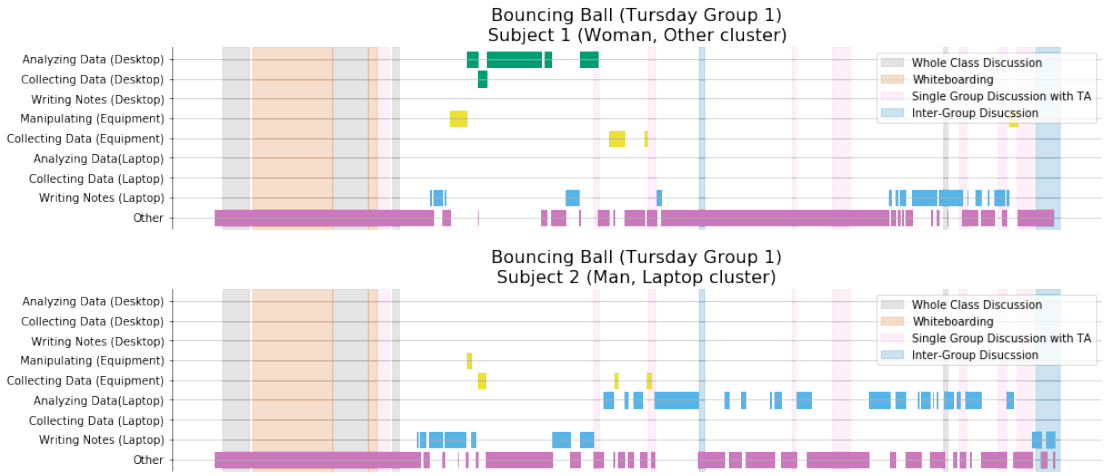


Figure J.3: **Decomposed student profiles** of two students in the same group during the Bouncing Ball experiment, with the x-axis representing time.

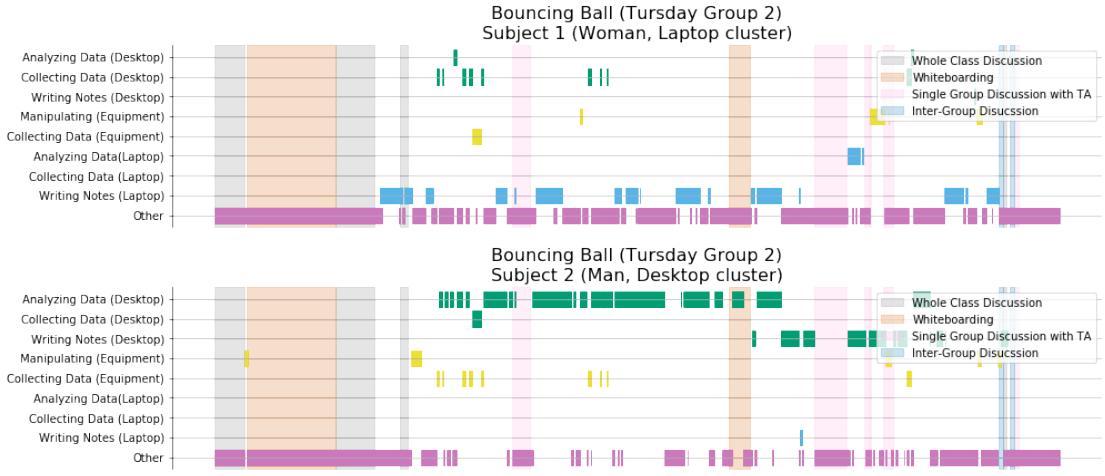


Figure J.4: **Decomposed student profiles** of two students in the same group during the Bouncing Ball experiment, with the x-axis representing time.

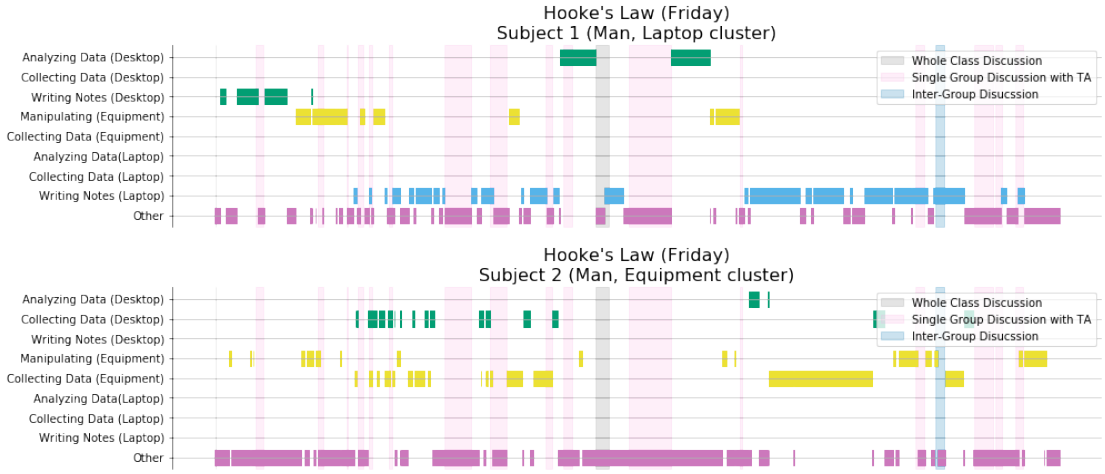


Figure J.5: **Decomposed student profiles** of two students in the same group during the Hooke's Law experiment, with the *x*-axis representing time..

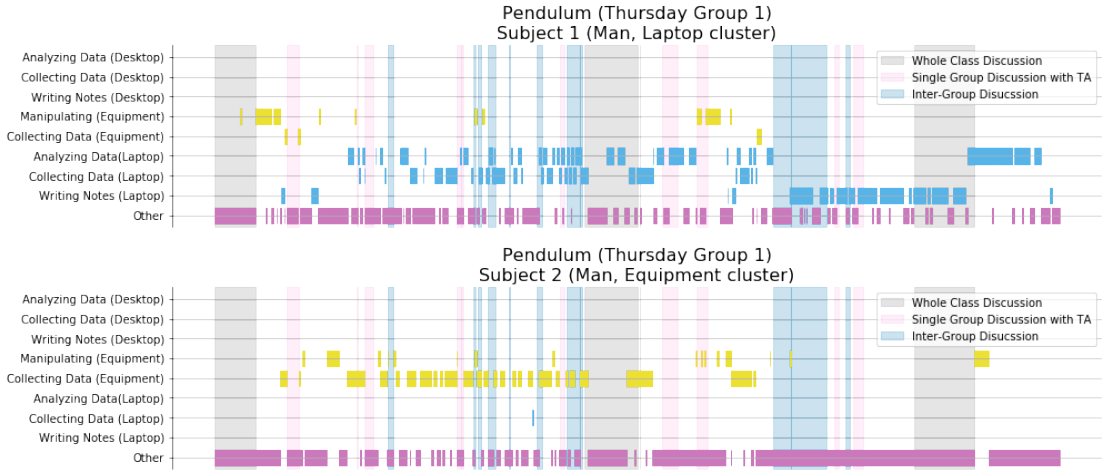


Figure J.6: **Decomposed student profiles** of two students in the same group during the Pendulum experiment, with the *x*-axis representing time.

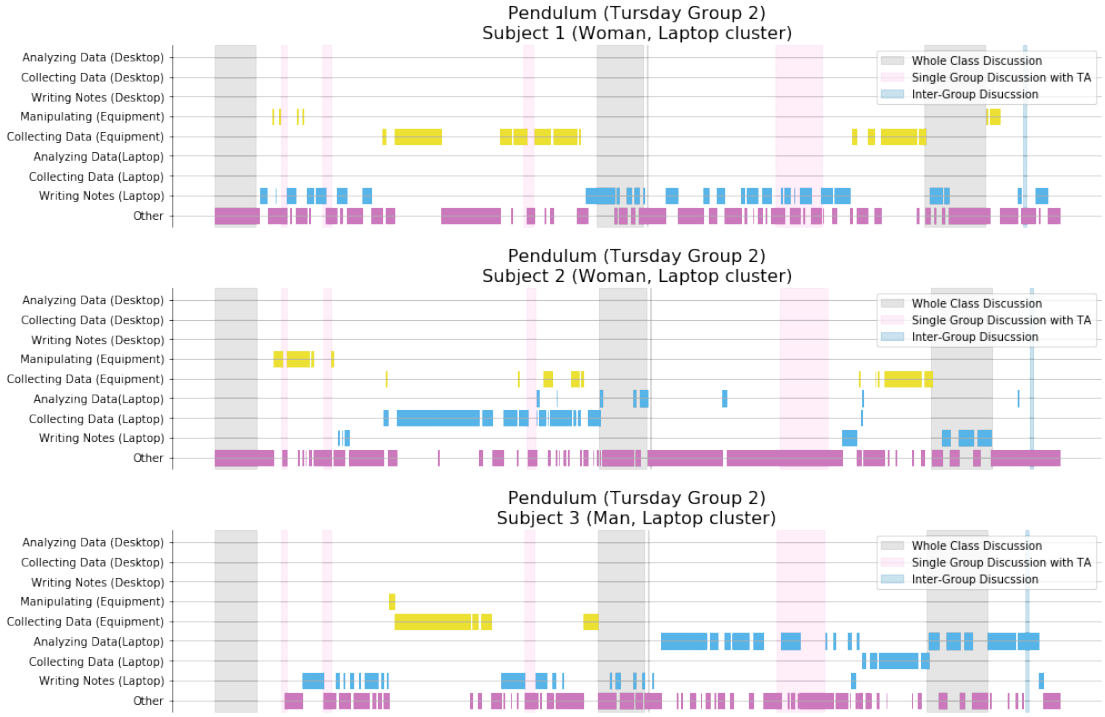


Figure J.7: **Decomposed student profiles** of three students in the same group during the Pendulum experiment, with the x-axis representing time.

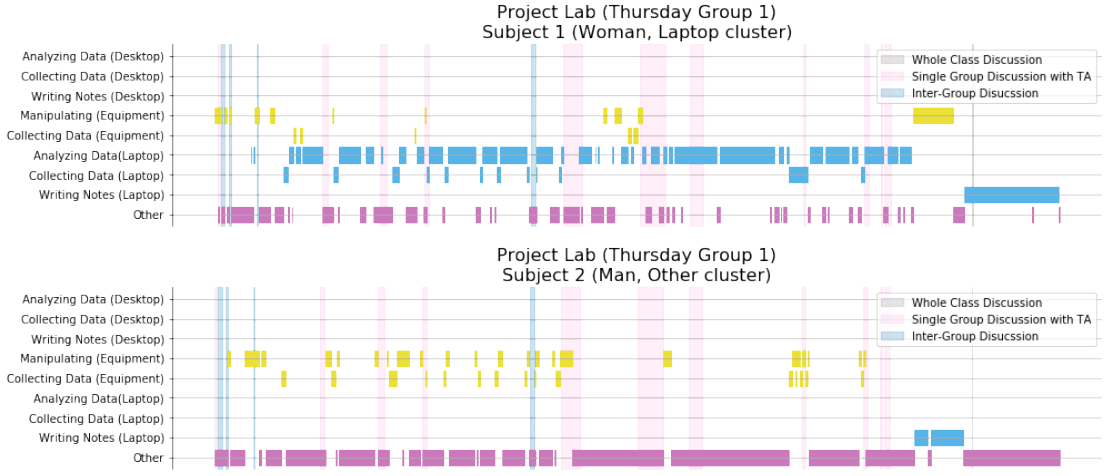


Figure J.8: **Decomposed student profiles** of two students in the same group during the Project Lab experiment, with the x-axis representing time.

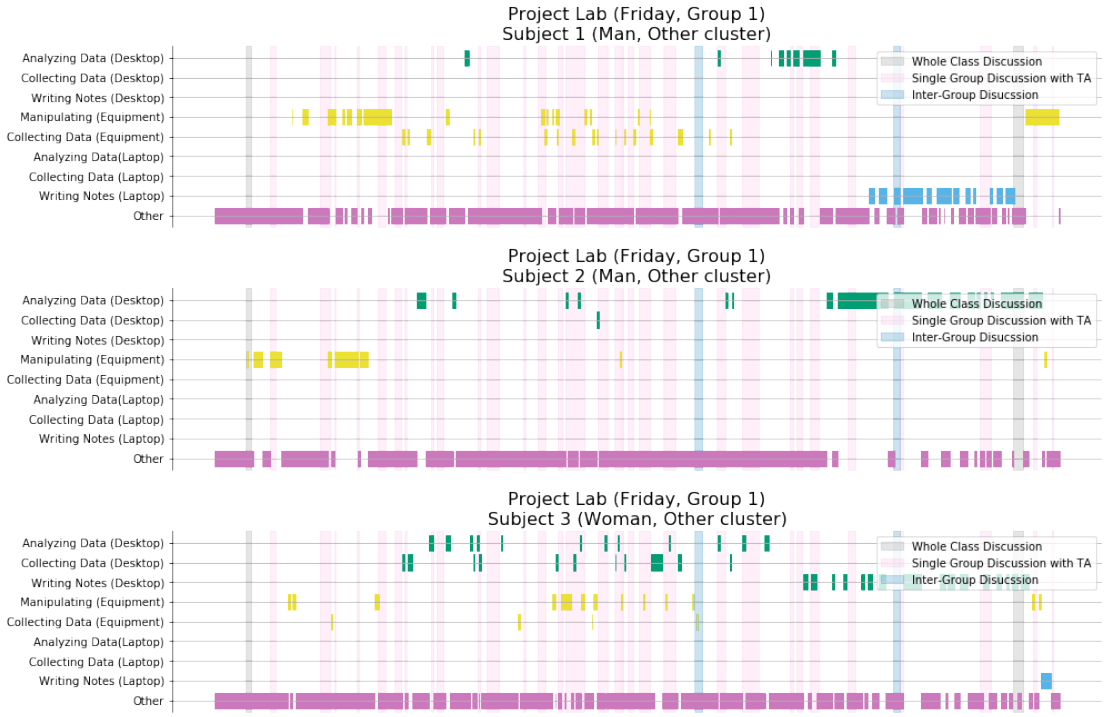


Figure J.9: **Decomposed student profiles** of three students in the same group during the Project Lab experiment, with the x-axis representing time.

BIBLIOGRAPHY

[1] ABADI, M., ET AL. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. Software available from https://tensorflow.org.

[2] AITKENHEAD, D. Peter higgs: I wouldn't be productive enough for today's academic system. *The Guardian 7* (2013). https://www.theguardian.com/science/2013/dec/06/peter-higgs-boson-academic-system.

[3] ALI, S. M., AND SILVEY, S. D. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society, Series B. 28*, 1 (1966), 131–142. https://www.jstor.org/stable/2984279.

[4] AMARI, S.-I. Information geometry and its applications. In *Applied Mathematical Sciences* (2016), vol. 194, Springer, Japan.

[5] AMARI, S.-I., AND NAGAOKA, H. Methods of information geometry. In *Translations of Mathematical Monographs* (2000), vol. 191, Oxford University Press.

[6] ANDERSON, C., AND BROWN, C. E. The functions and dysfunctions of hierarchy. *Research in Organizational Behavior 30* (2010), 55 – 89. https://doi.org/10.1016/j.riob.2010.08.002.

[7] ANDERSON, P. W. More is different. *Science 177*, 4047 (1972), 393–396. https://doi.org/10.1126/science.177.4047.393.

[8] ANDERSSON, S., AND JOHANSSON, A. Gender gap or program gap? students' negotiations of study practice in a course in electromagnetism. *Phys. Rev. Phys. Educ. Res. 12* (Aug 2016), 020112. https://doi.org/10.1103/PhysRevPhysEducRes.12.020112.

[9] ANICICH, E. M., SWAAB, R. I., AND GALINSKY, A. D. Hierarchical cultural values predict success and mortality in high-stakes teams. *Proceedings of the National Academy of Sciences 112*, 5 (2015), 1338–1343. https://doi.org/10.1073/pnas.1408800112.

[10] ARCHER, L., MOOTE, J., FRANCIS, B., DEWITT, J., AND YEOMANS, L. The "exceptional" physics girl: A sociological analysis of multimethod

data from young women aged 10–16 to explore gendered patterns of post-16 participation. *American Educational Research Journal 54*, 1 (2017), 88–126. https://doi.org/10.3102/0002831216678379.

[11] AVERICK, B. M., CARTER, R. G., MORÉ, J. J., AND XUE, G.-L. The MINIPACK-2 test problem collection. Tech. Rep. MCS-P153-0692, Mathematics and Computer Science Division, Argonne National Laboratory, June 1992.

[12] BAUMANN, D., ET AL. Probing inflation with cmb polarization. *AIP Conference Proceedings 1141*, 1 (2009), 10–120. https://doi.org/10.1063/1.3160885.

[13] BECKERMANN, B., AND TOWNSEND, A. On the singular values of matrices with displacement structure. *SIAM Journal on Matrix Analysis and Applications 38*, 4 (2017), 1227–1248. https://doi.org/10.1137/16M1096426.

[14] BENDERLY, B. L. Rosalind franklin and the damage of gender harassment. *Science* (Aug 2018). https://doi.org/10.1126/science.caredit.aau9709.

[15] BENNETT, M. Men's and women's self-estimates of intelligence. *The Journal of Social Psychology 136*, 3 (1996), 411–412. https://doi.org/10.1080/00224545.1996.9714021.

[16] BERGE, M., AND DANIELSSON, A. T. Characterising learning interactions: A study of university students solving physics problems in groups. *Research in Science Education 43*, 3 (Jun 2013), 1177–1196. https://doi.org/10.1007/s11165-012-9307-0.

[17] BERMAN, G. J., AND WANG, Z. J. Energy-minimizing kinematics in hovering insect flight. *Journal of Fluid Mechanics 582* (2007), 153–168. https://doi.org/10.1017/S0022112007006209.

[18] BEYER, K., GOLDSTEIN, J., RAMAKRISHNAN, R., AND SHAFT, U. When is "nearest neighbor" meaningful? In *Database Theory — ICDT'99* (Berlin, Heidelberg, 1999), Springer Berlin Heidelberg, pp. 217–235.

[19] BHATTACHARYYA, A. On a measure of divergence between two multinomial populations. *Sankhyā: The Indian Journal of Statistics (1933-1960) 7*, 4 (1946), 401–406. https://www.jstor.org/stable/25047882.

[20] BISHOP, C. M. *Pattern Recognition and Machine Learning*. Springer, NY, 2006.

[21] BRICKHOUSE, N. W. Embodying science: A feminist perspective on learning. *Journal of Research in Science Teaching 38*, 3 (2001), 282–295. https://doi.org/10.1002/1098-2736(200103)38:3<282::AID-TEA1006>3.0.CO;2-0.

[22] BROWN, K. S., HILL, C. C., CALERO, G. A., MYERS, C. R., LEE, K. H., SETHNA, J. P., AND CERIONE, R. A. The statistical mechanics of complex signaling networks: nerve growth factor signaling. *Physical Biology 1*, 3 (oct 2004), 184–195. https://doi.org/10.1088%2F1478-3967%2F1%2F3%2F006.

[23] BROWN, K. S., AND SETHNA, J. P. Statistical mechanical approaches to models with many poorly known parameters. *Phys. Rev. E 68* (Aug 2003), 021904. https://doi.org/10.1103/PhysRevE.68.021904.

[24] BUG, A. Has feminism changed physics? *Signs: Journal of Women in Culture and Society 28*, 3 (2003), 881–899. https://doi.org/doi.org/10.1086/345323.

[25] BUTLER, J. *Gender Trouble: Feminism and the Subversion of Identity*. Routledge, New York, 1999.

[26] CARLONE, H. B. The cultural production of science in reform-based physics: Girls' access, participation, and resistance. *Journal of Research in Science Teaching 41*, 4 (2004), 392–414. https://doi.org/10.1002/tea.20006.

[27] CARLONE, H. B., AND JOHNSON, A. Understanding the science experiences of successful women of color: Science identity as an analytic lens. *Journal of Research in Science Teaching 44*, 8 (2007), 1187–1218. https://doi.org/10.1002/tea.20237.

[28] CHACHRA, R., TRANSTRUM, M. K., AND SETHNA, J. P. Structural susceptibility and separation of time scales in the van der pol oscillator. *Phys. Rev. E 86* (Aug 2012), 026712. https://doi.org/10.1103/PhysRevE.86.026712.

[29] CHEN, S., BILLINGS, S., AND GRANT, P. Non-linear system identification using neural networks. *International Journal of Control 51*, 6 (1990), 1191–1214. https://doi.org/10.1080/00207179008934126.

[30] CHIS, O.-T., BANGA, J. R., AND BALSA-CANTO, E. Structural identifiability of systems biology models: A critical comparison of methods. *PLOS ONE 6*, 11 (11 2011), 1–16. https://doi.org/10.1371/journal.pone.0027755.

[31] COIFMAN, R. R., LAFON, S., LEE, A. B., MAGGIONI, M., NADLER, B., WARNER, F., AND ZUCKER, S. W. Geometric diffusions as a tool for harmonic analysis and structure definition of data: Diffusion maps. *Proceedings of the National Academy of Sciences 102*, 21 (2005), 7426–7431. https://doi.org/10.1073/pnas.0500334102.

[32] CORPUS, J. H., AND WORMINGTON, S. V. Profiles of intrinsic and extrinsic motivations in elementary school: A longitudinal analysis. *The Journal of Experimental Education 82*, 4 (2014), 480–501. https://doi.org/10.1080/00220973.2013.876225.

[33] COSTA, S. I. R., SANTOS, S. A., AND STRAPASSON, J. E. Fisher information distance: A geometrical reading. *Discrete Applied Mathematics 197* (2015), 59 – 69. https://doi.org/10.1016/j.dam.2014.10.004.

[34] CRAMÉR, H. *Mathematical Methods of Statistics*. Princeton Univ. Press., 1946.

[35] CSISZÁR, I. Eine informationstheoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoffschen ketten. *Magyar. Tud. Akad. Mat. Kutato Int. Kozl. 8* (1963), 85–108.

[36] CYRANOSKI, D., GILBERT, N., LEDFORD, H., NAYAR, A., AND YAHIA, M. Education: the phd factory. *Nature 472* (April 2011), 276–279. https://doi.org/doi:10.1038/472276a.

[37] DANIELS, B. C., CHEN, Y.-J., SETHNA, J. P., GUTENKUNST, R. N., AND MYERS, C. R. Sloppiness, robustness, and evolvability in systems biology. *Current Opinion in Biotechnology 19*, 4 (2008), 389 – 395. https://doi.org/10.1016/j.copbio.2008.06.008.

[38] DANIELSSON, A. T., AND LINDER, C. Learning in physics by doing laboratory work: towards a new conceptual framework. *Gender and Education 21*, 2 (2009), 129–144. https://doi.org/10.1080/09540250802213081.

[39] DAVIES, B., AND R., H. Positioning: The discursive production of selves. *Journal for the Theory of Social Behaviour 20*, 1 (1990), 43–63. https://doi.org/10.1111/j.1468-5914.1990.tb00174.x.

[40] DAY, J., STANG, J. B., HOLMES, N. G., KUMAR, D., AND BONN, D. A. Gender gaps and gendered action in a first-year physics laboratory. *Phys. Rev. Phys. Educ. Res. 12* (Aug 2016), 020104. https://doi.org/10.1103/PhysRevPhysEducRes.12.020104.

[41] DE OLIVEIRA, M. F., AND LEVKOWITZ, H. From visual data exploration to visual data mining: a survey. *IEEE Transactions on Visualization and Computer Graphics 9*, 3 (July 2003), 378–394. https://doi.org/10.1109/TVCG.2003.1207445.

[42] DEMANET, L., AND TOWNSEND, A. Stable extrapolation of analytic functions. *Found. Comput. Math.* (2018). https://doi.org/10.1007/s10208-018-9384-1.

[43] DSILVA, C. J., TALMON, R., RABIN, N., COIFMAN, R. R., AND KEVREKIDIS, I. G. Nonlinear intrinsic variables and state reconstruction in multiscale simulations. *Journal of Chemical Physics 139*, 18 (Nov 2013), 184109. https://doi.org/10.1063/1.4828457.

[44] DWECK, C. S. *Mindsets and math/science achievement*. Institute for Advanced Study, 2014.

[45] DYSON, F. A meeting with enrico fermi. *Nature 427*, 6972 (2004), 297. https://doi.org/10.1038/427297a.

[46] EDDY, S. L., AND BROWNELL, S. E. Beneath the numbers: A review of gender disparities in undergraduate education across science, technology, engineering, and math disciplines. *Phys. Rev. Phys. Educ. Res. 12* (Aug 2016), 020106. https://doi.org/10.1103/PhysRevPhysEducRes.12.020106.

[47] ESTER, M., KRIEGEL, H.-P., SANDER, J., XU, X., ET AL. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (1996), vol. 96, pp. 226–231. https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf.

[48] EVANS, T. M., BIRA, L., GASTELUM, J. B., WEISS, L. T., AND VANDERFORD, N. L. Evidence for a mental health crisis in graduate education. *Na-*

ture biotechnology 36, 3 (2018), 282. https://doi.org/10.1038/nbt.4089.

[49] FEYERABEND, P. *Tyranny of Science*. Polity Press, 2011.

[50] FEYNMAN, R. The character of physical law, 1964. Speech available at http://www.cornell.edu/video/playlist/richard-feynman-messenger-lectures.

[51] FISHER, R. A. On an absolute criterion in systems of nonlinear equations. *Messenger Mathematics 41* (1912), 155–160.

[52] FOUXON, I., AND OZ, Y. Conformal field theory as microscopic dynamics of incompressible euler and navier-stokes equations. *Phys. Rev. Lett. 101* (Dec 2008), 261602. https://link.aps.org/doi/10.1103/PhysRevLett.101.261602.

[53] FRANCIS, B., ARCHER, L., MOOTE, J., DEWITT, J., MACLEOD, E., AND YEOMANS, L. The construction of physics as a quintessentially masculine subject: Young people's perceptions of gender issues in access to physics. *Sex Roles 76*, 3 (Feb 2017), 156–174. https://doi.org/10.1007/s11199-016-0669-z.

[54] FRANCIS, B., AND PAECHTER, C. The problem of gender categorisation: addressing dilemmas past and present in gender and education research. *Gender and Education 27*, 7 (2015), 776–790. https://doi.org/10.1080/09540253.2015.1092503.

[55] FRIARD, O., AND GAMBA, M. Boris: a free, versatile open-source event-logging software for video/audio coding and live observations. *Methods in Ecology and Evolution 7*, 11 (2016), 1325–1330. https://doi.org/10.1111/2041-210X.12584.

[56] GOLUB, G. H., AND VAN LOAN, C. F. *Matrix Computations*. Johns Hopkins University Press, Baltimore, 1996.

[57] GONSALVES, A. J., DANIELSSON, A., AND PETTERSSON, H. Masculinities and experimental practices in physics: The view from three case studies. *Phys. Rev. Phys. Educ. Res. 12* (Aug 2016), 020120. https://doi.org/10.1103/PhysRevPhysEducRes.12.020120.

[58] GOSLING, C. Identity as a research lens in science and physics education. *Journal of Belonging, Identity, Language, and Diversity 1*, 1 (2017). https://bild-lida.ca/journal/volume_1_1_2017/identity-as-a-research-lens-in-science-and-physics-education/.

[59] GOULD, S. J. *The mismeasure of man*. WW Norton & Company, 1996.

[60] GROMOV, M. In a search for a structure, part 1: On entropy, 2013. https://www.ihes.fr/gromov.

[61] GUTENKUNST, R. *Sloppiness, modeling, and evolution in biochemical networks*. PhD thesis, Cornell University, 2007. https://ecommons.library.cornell.edu/handle/1813/8206.

[62] GUTENKUNST, R. N., WATERFALL, J. J., CASEY, F. P., BROWN, K. S., MYERS, C. R., AND SETHNA, J. P. Universally sloppy parameter sensitivities in systems biology models. *PLOS Comput. Bio 3*, 10 (10 2007), 1–8. https://doi.org/10.1371/journal.pcbi.0030189.

[63] HALEVY, N., CHOU, E. Y., AND GALINSKY, A. D. A functional model of hierarchy: Why, how, and when vertical differentiation enhances group performance. *Organizational Psychology Review 1*, 1 (2011), 32–52. https://doi.org/10.1177/2041386610380991.

[64] HAMIMECHE, S., AND LEWIS, A. Likelihood analysis of cmb temperature and polarization power spectra. *Phys. Rev. D 77* (May 2008), 103013. https://doi.org/10.1103/PhysRevD.77.103013.

[65] HARDING, S. *Whose science? Whose knowledge?: Thinking from women's lives*. Cornell University Press, 2016.

[66] HARDING, S. G. *The science question in feminism*. Cornell University Press, 1986.

[67] HARTIGAN, J. A., AND WONG, M. A. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics) 28*, 1 (1979), 100–108. https://jstor.org/stable/2346830.

[68] HARTMAN, E. J., KEELER, J. D., AND KOWALSKI, J. M. Layered neural networks with gaussian hidden units as universal approximations. *Neural*

Computation 2, 2 (1990), 210–215. https://doi.org/10.1162/neco.1990.2.2.210.

[69] HASSE, C. Learning and transition in a culture of playful physicists. *European Journal of Psychology of Education 23*, 2 (Jun 2008), 149. http://doi.org/10.1007/BF03172742.

[70] HELLINGER, E. Neue Begründung der Theorie quadratischer Formen von unendlichvielen Veränderlichen. *J. Reine Angew. Math. 136* (1909), 210–271.

[71] HETHCOTE, H. W. The mathematics of infectious diseases. *SIAM Rev. 42*, 4 (Dec. 2000), 599–653. https://doi.org/10.1137/S0036144500371907.

[72] HINKLE, D. E., WIERSMA, W., AND JURS, S. G. *Applied statistics for the behavioral sciences*. Houghton Mifflin Boston, 1988.

[73] HOCHANADEL, A., AND FINAMORE, D. Fixed and growth mindset in education and how grit helps students persist in the face of adversity. *Journal of International Education Research 11*, 1 (2015), 47–50. https://doi.org/https://eric.ed.gov/?id=EJ1051129.

[74] HOLMES, N. G., ROLL, I., AND BONN, D. A. Participating in the physics lab: Does gender matter? *Physics in Canada 70*, 2 (2014), 1–3. https://pic-pac.cap.ca/static/downloads/fb5c638abcd67ec1a718ff76146159606bc7a928.pdf.

[75] HOOKWAY, C. *Truth, rationality, and pragmatism: Themes from Peirce*. Oxford University Press on Demand, 2002.

[76] HORNBOSTEL, K., LEPAGE, G. P., DAVIES, C. T. H., DOWDALL, R. J., NA, H., AND SHIGEMITSU, J. Fast fits for lattice qcd correlators. *Phys. Rev. D 85* (Feb 2012), 031504. https://doi.org/10.1103/PhysRevD.85.031504.

[77] HOTELLING, H. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology 24*, 6 (1933), 417. https://doi.org/10.1037/h0071325.

[78] HOWLETT, C., LEWIS, A., HALL, A., AND CHALLINOR, A. CMB power spectrum parameter degeneracies in the era of precision cosmology. *Jour-*

nal of Cosmology and Astroparticle Physics 2012, 04 (apr 2012), 027–027. https://doi.org/10.1088/1475-7516/2012/04/027.

[79] HU, W. Cmb tutorials, 2001. http://background.uchicago.edu/index.html.

[80] IRVING, P. W., AND SAYRE, E. C. Becoming a physicist: The roles of research, mindsets, and milestones in upper-division student perceptions. *Phys. Rev. ST Phys. Educ. Res. 11* (Sep 2015), 020120. https://doi.org/10.1103/PhysRevSTPER.11.020120.

[81] IRVING, P. W., AND SAYRE, E. C. Identity statuses in upper-division physics students. *Cultural Studies of Science Education 11*, 4 (Dec 2016), 1155–1200. https://doi.org/10.1007/s11422-015-9682-8.

[82] JOVANOVIC, J., AND KING, S. S. Boys and girls in the performance-based science classroom: Who's doing the performing? *American Educational Research Journal 35*, 3 (1998), 477–496. https://doi.org/10.3102/00028312035003477.

[83] KAVOURAS, A., GEORGAKIS, C., KELLEY, C. T., SIETTOS, C., AND KEVREKIDIS, I. G. Steady states for chemical process plants: A legacy code, time-stepping approach. *Aiche Journal 59*, 9 (Sep 2013), 3308–3321. https://doi.org/10.1002/aic.14199.

[84] KIM, J. Making sense of emergence. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition 95*, 1/2 (1999), 3–36. https://www.jstor.org/stable/4320946.

[85] KOWALIK, J., AND MORRISON, J. F. Analysis of kinetic data for allosteric enzyme reactions as a nonlinear regression problem. *Mathematical Biosciences 2*, 1 (1968), 57 – 66. https://doi.org/10.1016/0025-5564(68)90006-0.

[86] KOZMINSKI, J., LEWANDOWSKI, H., BEVERLY, N., LINDAAS, S., DEARDORFF, D., REAGAN, A., DIETZ, R., TAGG, R., WILLIAMS, J., HOBBS, R., ET AL. Aapt recommendations for the undergraduate physics laboratory curriculum. *American Association of Physics Teachers* (2014), 29. https://www.aapt.org/resources/upload/labguidlinesdocument_ebendorsed_nov10.pdf.

[87] KREUTZ, C., AND TIMMER, J. Systems biology: experimental design. *The FEBS Journal 276*, 4 (2009), 923–942. https://doi.org/10.1111/j.1742-4658.2008.06843.x.

[88] KRIEGEL, H.-P., KRÖGER, P., AND ZIMEK, A. Clustering high-dimensional data: A survey on subspace clustering, pattern-based clustering, and correlation clustering. *ACM Trans. Knowl. Discov. Data 3*, 1 (Mar. 2009), 1:1–1:58. https://doi.acm.org/10.1145/1497577.1497578.

[89] KULLBACK, S., AND LEIBLER, R. A. On information and sufficiency. *Ann. Math. Statist. 22*, 1 (03 1951), 79–86. https://doi.org/10.1214/aoms/1177729694.

[90] LECUN, Y., CORTES, C., AND BURGES, C. J. Mnist database. Tech. rep., Courant Institute, Google Lab, Microsoft Research, 2018. http://yann.lecun.com/exdb/mnist/.

[91] LEE, J. A., AND VERLEYSEN, M. *Nonlinear Dimensionality Reduction*. Springer, NY, 2007. https://doi.org/10.1007/978-0-387-39351-3.

[92] LEPAGE, G. P., CLARK, B., DAVIES, C. T., HORNBOSTEL, K., MACKENZIE, P. B., MORNINGSTAR, C., AND TROTTIER, H. Constrained curve fitting. *Nuclear Physics B - Proceedings Supplements 106-107* (2002), 12 – 20. https://doi.org/10.1016/S0920-5632(01)01638-3.

[93] LESLIE, S.-J., CIMPIAN, A., MEYER, M., AND FREELAND, E. Expectations of brilliance underlie gender distributions across academic disciplines. *Science 347*, 6219 (2015), 262–265. https://doi.org/10.1126/science.1261375.

[94] LEWIS, A., CHALLINOR, A., AND LASENBY, A. Efficient computation of cosmic microwave background anisotropies in closed friedmann-robertson-walker models. *The Astrophysical Journal 538*, 2 (Aug 2000), 473–476. https://doi.org/10.1086%2F309179.

[95] LEWIS, K. L., STOUT, J. G., POLLOCK, S. J., FINKELSTEIN, N. D., AND ITO, T. A. Fitting in or opting out: A review of key social-psychological factors influencing a sense of belonging for women in physics. *Phys. Rev. Phys. Educ. Res. 12* (Aug 2016), 020110. https://doi.org/10.1103/PhysRevPhysEducRes.12.020110.

[96] LIU, P., SAFFORD, H. R., COUZIN, I. D., AND KEVREKIDIS, I. G. Coarse-grained variables for particle-based models: diffusion maps and animal swarming simulations. *Computational Particle Mechanics 1*, 4 (Dec 2014), 425–440. https://doi.org/10.1007/s40571-014-0030-7.

[97] LIU, S., MALJOVEC, D., WANG, B., BREMER, P.-T., AND PASCUCCI, V. Visualizing high-dimensional data: Advances in the past decade. *IEEE Transactions on Visualization and Computer Graphics 23*, 3 (March 2017), 1249–1268. https://doi.org/10.1109/TVCG.2016.2640960.

[98] MAATEN, L. V. D., AND HINTON, G. Visualizing data using t-sne. *Journal of machine learning research 9*, Nov (2008), 2579–2605. https://www.jmlr.org/papers/volume9/vandermaaten08a/vandermaaten08a.pdf.

[99] MACHTA, B. B., CHACHRA, R., TRANSTRUM, M. K., AND SETHNA, J. P. Parameter space compression underlies emergent theories and predictive models. *Science 342*, 6158 (2013), 604–607. https://doi.org/0.1126/science.1238723.

[100] MADSEN, A., MCKAGAN, S. B., AND SAYRE, E. C. Gender gap on concept inventories in physics: What is consistent, what is inconsistent, and what factors influence the gap? *Phys. Rev. ST Phys. Educ. Res. 9* (Nov 2013), 020121. https://doi.org/10.1103/PhysRevSTPER.9.020121.

[101] MADSEN, A., MCKAGAN, S. B., AND SAYRE, E. C. How physics instruction impacts students' beliefs about learning physics: A meta-analysis of 24 studies. *Phys. Rev. ST Phys. Educ. Res. 11* (Jun 2015), 010115. https://doi.org/10.1103/PhysRevSTPER.11.010115.

[102] MCINNES, L., HEALY, J., AND MELVILLE, J. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426* (2018). https://arxiv.org/abs/1802.03426.

[103] MEHTA, M., FOX, R. O., AND PEPIOT, P. Reduced chemical kinetics for the modeling of tio2 nanoparticle synthesis in flame reactors. *Industrial & Engineering Chemistry Research 54*, 20 (2015), 5407–5415. https://doi.org/10.1021/acs.iecr.5b00130.

[104] MÉZARD, M., PARISI, G., AND VIRASORO, M. *Spin Glass Theory and Beyond*. WORLD SCIENTIFIC, 1986. https://doi.org/10.1142/0271.

[105] MONTUORI, A., AND PURSER, R. E. Deconstructing the lone genius myth: Toward a contextual view of creativity. *Journal of Humanistic Psychology 35*, 3 (1995), 69–112. https://doi.org/10.1177/00221678950353005.

[106] MORIMOTO, T. Markov processes and the h-theorem. *Journal of the Physical Society of Japan 18*, 3 (1963), 328–331. https://doi.org/10.1143/JPSJ.18.328.

[107] MURPHY, K. P. *Machine Learning: A Probabilistic Perspective*. The MIT Press, 2012.

[108] NATIONAL SCIENCE FOUNDATION. Report to the national science board on the national science foundation's merit review process fiscal year 2014. Tech. Rep. NSB-2015-14, National Science Foundation, May 2015.

[109] NISSEN, J. M., AND SHEMWELL, J. T. Gender, experience, and self-efficacy in introductory physics. *Phys. Rev. Phys. Educ. Res. 12* (Aug 2016), 020105. https://doi.org/10.1103/PhysRevPhysEducRes.12.020105.

[110] PARISI, G. Infinite number of order parameters for spin-glasses. *Phys. Rev. Lett. 43* (Dec 1979), 1754–1756. https://doi.org/10.1103/PhysRevLett.43.1754.

[111] PASCANU, R., MIKOLOV, T., AND BENGIO, Y. On the difficulty of training recurrent neural networks. In *Proceedings of the 30th International Conference on International Conference on Machine Learning* (2013), vol. 28 of *ICML'13*, JMLR.org, pp. III–1310–III–1318. https://dl.acm.org/citation.cfm?id=3042817.3043083.

[112] PEROTTO, L., LESGOURGUES, J., HANNESTAD, S., TU, H., AND WONG, Y. Y. Y. Probing cosmological parameters with the CMB: forecasts from monte carlo simulations. *Journal of Cosmology and Astroparticle Physics 2006*, 10 (oct 2006), 013–013. https://doi.org/10.1088%2F1475-7516%2F2006%2F10%2F013.

[113] PETTERSSON, H. Making Masculinity in Plasma Physics. *Science Studies 24*, 1 (2011), 47–65. https://sciencestudies.fi/system/files/v24n1Pettersson.pdf.

[114] PLANCK COLLABORATION. Planck 2015 results - i. overview of products and scientific results. *A&A 594* (2016), A1. https://doi.org/10.1051/0004-6361/201527101.

[115] PLANCK COLLABORATION. Planck 2015 results - xi. cmb power spectra, likelihoods, and robustness of parameters. *A&A 594* (2016), A11. https://doi.org/10.1051/0004-6361/201526926.

[116] POPPER, K. *Conjectures and refutations: The growth of scientific knowledge*. Routledge, 1962.

[117] POWELL, A., BAGILHOLE, B., AND DAINTY, A. How women engineers do and undo gender: Consequences for gender equality. *Gender, Work & Organization 16*, 4 (2009), 411–428. https://doi.org/10.1111/j.1468-0432.2008.00406.x.

[118] QUINN, K. N., CLEMENT, C. B., DE BERNARDIS, F., NIEMACK, M., AND SETHNA, J. Visualizing probabilistic models: Intensive principal component analysis. *arXiv preprint arXiv:1810.02877* (2018). https://arxiv.org/abs/1810.02877.

[119] QUINN, K. N., MCGILL, K. L., KELLEY, M. M., SMITH, E. M., AND HOLMES, N. Who does what now? how physics lab instruction impacts student behaviors. In *Physics Education Research Conference 2018* (Washington, DC, August 1-2 2018), PER Conference. https://doi.org/10.1119/perc.2018.pr.Quinn.

[120] QUINN, K. N., MCGILL, K. L., KELLEY, M. M., SMITH, E. M., AND HOLMES, N. G. Who does what now? how physics lab instruction impacts student behaviors. *arXiv preprint arXiv:1807.09724* (2018). https://arxiv.org/abs/1807.09724.

[121] QUINN, K. N., WILBER, H., TOWNSEND, A., AND SETHNA, J. Chebyshev approximation and the global geometry of sloppy models. *arXiv preprint arXiv:1809.08280* (2018). https://arxiv.org/abs/1809.08280.

[122] QUINN, K. N., WILBER, H., TOWNSEND, A., AND SETHNA, J. P. Chebyshev approximation and the global geometry of model predictions. *Phys. Rev. Lett. 122* (Apr 2019), 158302. https://link.aps.org/doi/10.1103/PhysRevLett.122.158302.

[123] RAJU, A. *Aspects of the Renormalization Group*. PhD thesis, Cornell University, 2018. https://search.proquest.com/docview/2107819051?accountid=10267.

[124] RAJU, A., MACHTA, B. B., AND SETHNA, J. P. Information loss under coarse graining: A geometric approach. *Phys. Rev. E 98* (Nov 2018), 052112. https://doi.org/10.1103/PhysRevE.98.052112.

[125] RAMAN, D. V., ANDERSON, J., AND PAPACHRISTODOULOU, A. Delineating parameter unidentifiabilities in complex models. *Phys. Rev. E 95* (Mar 2017), 032314. https://doi.org/10.1103/PhysRevE.95.032314.

[126] RAO, C. R. *Information and the Accuracy Attainable in the Estimation of Statistical Parameters*. Springer New York, New York, NY, 1992. https://doi.org/10.1007/978-1-4612-0919-5_16.

[127] REES, T. Mainstreaming gender equality in science in the european union: The 'etan report'. *Gender and Education 13*, 3 (2001), 243–260. https://doi.org/10.1080/09540250120063544.

[128] ROLIN, K. Gender and physics: feminist philosophy and science education. *Science & Education 17*, 10 (Nov 2008), 1111–1125. https://doi.org/10.1007/s11191-006-9065-3.

[129] RORTY, R. *Consequences of Pragmatism*. University of Minnesota Press, 1982.

[130] ROSA, K., AND MENSAH, F. M. Educational pathways of black women physicists: Stories of experiencing and overcoming obstacles in life. *Phys. Rev. Phys. Educ. Res. 12* (Aug 2016), 020113. https://doi.org/10.1103/PhysRevPhysEducRes.12.020113.

[131] ROSSITER, M. W. The matthew matilda effect in science. *Social Studies of Science 23*, 2 (1993), 325–341. https://doi.org/10.1177/030631293023002004.

[132] SCHERR, R. Editorial: Never mind the gap: Gender-related research in physical review physics education research, 2005–2016. *Phys. Rev. Phys. Educ. Res. 12* (Nov 2016), 020003. https://doi.org/10.1103/PhysRevPhysEducRes.12.020003.

[133] SCHMIDT, J. A., ROSENBERG, J. M., AND BEYMER, P. N. A person-in-context approach to student engagement in science: Examining learning activities and choice. *Journal of Research in Science Teaching 55*, 1 (2017), 19–43. https://doi.org/10.1002/tea.21409.

[134] SETHNA, J. P., BIERBAUM, M. K., DAHMEN, K. A., GOODRICH, C. P., GREER, J. R., HAYDEN, L. X., KENT-DOBIAS, J. P., LEE, E. D., LIARTE, D. B., NI, X., QUINN, K. N., RAJU, A., ROCKLIN, D. Z., SHEKHAWAT, A., AND ZAPPERI, S. Deformation of crystals: Connections with statistical physics. *Annual Review of Materials Research 47*, 1 (2017), 217–246. https://doi.org/10.1146/annurev-matsci-070115-032036.

[135] SNYDERMAN, M., AND ROTHMAN, S. *The IQ controversy, the media and public policy.* Transaction Publishers, 1988.

[136] STEELE, C. M., AND ARONSON, J. Stereotype threat and the intellectual test performance of african americans. *Journal of personality and social psychology 69*, 5 (1995), 797. 10.1037/0022-3514.69.5.797.

[137] SWAAB, R. I., SCHAERER, M., ANICICH, E. M., RONAY, R., AND GALINSKY, A. D. The too-much-talent effect: Team interdependence determines when more talent is too much or not enough. *Psychological Science 25*, 8 (2014), 1581–1591. https://doi.org/10.1177/0956797614537280.

[138] TAO, T. Topics in random matrix theory. In *Graduate Studies in Mathematics* (2012), vol. 132, American Mathematical Soc.

[139] TEGMARK, M., AND DE OLIVEIRA-COSTA, A. How to measure cmb polarization power spectra without losing information. *Phys. Rev. D 64* (Aug 2001), 063001. https://doi.org/10.1103/PhysRevD.64.063001.

[140] TEGMARK, M., ET AL. Cosmological parameters from sdss and wmap. *Phys. Rev. D 69* (May 2004), 103501. https://doi.org/10.1103/PhysRevD.69.103501.

[141] THORNDIKE, R. L. Who belongs in the family? *Psychometrika 18*, 4 (Dec 1953), 267–276. https://doi.org/10.1007/BF02289263.

[142] TOFTESKOV, J., TØRNGREN, M. A., BAILEY, N. P., AND HANSEN, J. S. Modelling headspace dynamics in modified atmosphere packaged meat. *Journal of Food Engineering 248* (2019), 46 – 52. https://doi.org/10.1016/j.jfoodeng.2018.12.013.

[143] TONSO, K. L. Student engineers and engineer identity: Campus engineer identities as figured world. *Cultural Studies of Science Education 1*, 2 (Sep 2006), 273–307. https://doi.org/10.1007/s11422-005-9009-2.

[144] TORGERSON, W. S. Multidimensional scaling: I. theory and method. *Psychometrika 17*, 4 (1952), 401–419. https://doi.org/10.1007/BF02288916.

[145] TOWNSEND, A., AND WILBER, H. On the singular values of matrices with high displacement rank. *Lin. Alg. & Appl. 548* (2018), 19–41. https://doi.org/10.1016/j.laa.2018.02.025.

[146] TRANSTRUM, M. K., MACHTA, B. B., BROWN, K. S., DANIELS, B. C., MYERS, C. R., AND SETHNA, J. P. Perspective: Sloppiness and emergent theories in physics, biology, and beyond. *The Journal of Chemical Physics 143*, 1 (2015), 010901. https://doi.org/10.1063/1.4923066.

[147] TRANSTRUM, M. K., MACHTA, B. B., AND SETHNA, J. P. Why are nonlinear fits to data so challenging? *Phys. Rev. Lett. 104* (Feb 2010), 060201. https://doi.org/10.1103/PhysRevLett.104.060201.

[148] TRANSTRUM, M. K., MACHTA, B. B., AND SETHNA, J. P. Geometry of nonlinear least squares with applications to sloppy models and optimization. *Phys. Rev. E 83* (Mar 2011), 036701. https://doi.org/10.1103/PhysRevE.83.036701.

[149] TRANSTRUM, M. K., AND QIU, P. Model reduction by manifold boundaries. *Phys. Rev. Lett. 113* (Aug 2014), 098701. https://doi.org/10.1103/PhysRevLett.113.098701.

[150] TRANSTRUM, M. K., SARIĆ, A. T., AND STANKOVIĆ, A. M. Measurement-directed reduction of dynamic models in power systems. *IEEE Transactions on Power Systems 32*, 3 (May 2017), 2243–2253. https://doi.org/10.1109/TPWRS.2016.2611511.

[151] TRANSTRUM, M. K., SARIĆ, A. T., AND STANKOVIĆ, A. M. Information geometry approach to verification of dynamic models in power systems. *IEEE Transactions on Power Systems 33*, 1 (Jan 2018), 440–450. https://doi.org/10.1109/TPWRS.2017.2692523.

[152] TRAWEEK, S. *Beamtimes and lifetimes: The World of High Energy Physicists*. Harvard University Press, Cambridge, MA, 1988.

[153] TRAXLER, A. L., CID, X. C., BLUE, J., AND BARTHELEMY, R. Enriching gender in physics education research: A binary past and a complex future. *Phys. Rev. Phys. Educ. Res. 12* (Aug 2016), 020114. https://doi.org/10.1103/PhysRevPhysEducRes.12.020114.

[154] TREFETHEN, L. N. *Approximation Theory and Approximation Practice*. SIAM, Philadelphia, 2013.

[155] WADE, N. Discovery of pulsars: A graduate student's story. *Science 189*, 4200 (1975), 358–364. http://www.jstor.org/stable/1740548.

[156] WANDS, D., PIATTELLA, O. F., AND CASARINI, L. Physics of the cosmic microwave background radiation. In *The Cosmic Microwave Background*. Springer, 2016, pp. 3–39. https://doi.org/10.1007/978-3-319-44769-8_1.

[157] WATERFALL, J. J., CASEY, F. P., GUTENKUNST, R. N., BROWN, K. S., MYERS, C. R., BROUWER, P. W., ELSER, V., AND SETHNA, J. P. Sloppy-model universality class and the vandermonde matrix. *Phys. Rev. Lett. 97* (Oct 2006), 150601. https://doi.org/10.1103/PhysRevLett.97.150601.

[158] WEIDEMAN, J., AND TREFETHEN, L. N. The kink phenomenon in fejér and clenshaw–curtis quadrature. *Numerische Mathematik 107*, 4 (2007), 707–727. https://doi.org/10.1007/s00211-007-0101-2.

[159] WEST, C., AND ZIMMERMAN, D. H. Doing gender. *Gender and Society 1*, 2 (1987), 125–151. https://www.jstor.org/stable/189945.

[160] WIGNER, E. P. On the distribution of the roots of certain symmetric matrices. *Annals of Mathematics 67*, 2 (1958), 325–327. http://www.jstor.org/stable/1970008.

[161] WILLIAMS, R. January 1, 1925: Cecilia payne-gaposchkin and the day the universe changed. *APS News 24*, 1 (2015). https://www.aps.org/publications/apsnews/201501/physicshistory.cfm.

[162] WOOLSTON, C. Feeling overwhelmed by academia? you are not alone. *Nature 557*, 7703 (2018), 129. https://doi.org/10.1038/d41586-018-04998-1.

[163] ZIMEK, A., SCHUBERT, E., AND KRIEGEL, H.-P. A survey on unsupervised outlier detection in high-dimensional numerical data. *Statistical Analysis and Data Mining: The ASA Data Science Journal 5*, 5 (2012), 363–387. https://doi.org/10.1002/sam.11161.