The Dissertation Committee for Avradeep Bhowmik
certifies that this is the approved version of the following dissertation:

Learning from Aggregated Data

Committee:

Joydeep Ghosh, Supervisor

Haris Vikalo

Sujay Sanghavi

Alexandros Dimakis

Oluwasanmi Koyejo

Learning from Aggregated Data

by

Avradeep Bhowmik

DISSERTATION

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT AUSTIN

December 2018

Dedicated to my parents, my friends and my family

Acknowledgments

Thanks are due to innumerable people whose encouragement and support acted like Theseus's roll of yarn, helping me navigate the labyrinthine corridors of graduate school. I am deeply grateful to my advisor, Prof. Joydeep Ghosh, whose constant support and unwavering trust helped me develop a sense of independence that has benefited me in my research and beyond. His guidance has been indispensable not only in developing technical proficiency, but also in learning to think non-solipsistically about effective communication of ideas, arguments and opinions, to diverse communities in the wider world. I am also indebted to my committee members– Prof. Haris Vikalo, Prof. Sujay Sanghavi, Prof. Alex Dimakis and Prof. Oluwasanmi Koyejo for their insightful comments and suggestions for my research.

In particular, I shall remain deeply indebted to Sanmi for his collaboration and friendship– without his mentorship and guidance this dissertation would not have been possible. I also wish to thank Dr. Suju Rajan for her mentorship, both at Yahoo and Criteo– her advice and feedback helped me develop a pragmatic way of thinking, both about ideas as well as about bringing them to fruition. I am privileged to have had the opportunity to collaborate with some amazing individuals– Nathan Liu, Zhengming Xing, Minmin Chen, Badri Narayan Bhaskar and Erheng Zhong– they have had a massive contribution to making my summer internship experiences both fun as well as fruitful. I was started on this path by one of the most remarkable

and erudite individuals I have had the privilege of knowing– I thank Prof. Borkar for being a constant source of inspiration and wisdom throughout.

My friends in Austin have been a constant source of happiness over the past few years. I thank Shalmali, Vatsal, Suriya and Ayan for always being there for me. I have learned much and more from them, and from my colleagues in IDEA Lab– Sreangsu, Joyce, Yubin, Rajiv, Jette, Diego, Mike, Taewan, Woody, Alan, Dany and Farzan. I have spent many fun evenings with Murat, Eirini, Sam, Siddhant, Soumya, Ajil, Tim, Ashish and Megha. I thank them all for being a part of the good times, the happy hours, the rants, the arguments, the debates, and the congratulations as well as commiserations as apt. The virtual companionship of Nishant and Sham, and my friends back home and in other cities, have all contributed to making grad school a pleasurable experience.

Finally, I thank my parents whose constant encouragement and faith in me throughout has always been my principal source of motivation.

I wish I could individually thank everyone who supported me throughout this process but that would lead to a list longer than the dissertation itself. For those I do not mention here by name, know that you have not been forgotten– you shall always have my sincerest gratitude.

Avradeep Bhowmik

The University of Texas at Austin

November 2018

Learning from Aggregated Data

Publication No. _____

Avradeep Bhowmik, Ph.D.
The University of Texas at Austin, 2018

Supervisor: Joydeep Ghosh

Data aggregation is ubiquitous in modern life. Due to various reasons like privacy, scalability, robustness, etc., ground truth data is often subjected to aggregation before being released to the public, or utilised by researchers and analysts. Learning from aggregated data is a challenging problem that requires significant algorithmic innovation, since naive application of standard techniques to aggregated data is vulnerable to the ecological fallacy. In this work, we explore three different versions of this setting.

First, we tackle the problem of using generalised linear models when features/covariates are fully observed but the targets are only available as histograms- a common scenario in the healthcare domain where many datasets contain both non-sensitive attributes like age, sex, zip-code, etc., as well as privacy sensitive attributes like healthcare records. We introduce an efficient algorithm that uses alternating data imputation and GLM estimation steps to learn predictive models in this setting.

Next, we look at the problem of learning sparse linear models when both features and targets are in aggregated form, specified as empirical estimates of group-wise means computed over different sub-groups of the population. We show that if the true sub-populations are heterogeneous enough, the optimal sparse parameter can be recovered within an arbitrarily small tolerance even in the presence of noise, provided the empirical estimates are obtained from a sufficiently large number of observations.

Third, we tackle the scenario of predictive modelling with data that is subjected to spatio-temporal aggregation. We show that by formulating the problem in the frequency domain, we can bypass the mathematical and representational challenges that arise due to non-uniform aggregation, misaligned sampling periods and aliasing. We introduce a novel algorithm that uses restricted Fourier transforms to estimate a linear model which, when applied to spatio-temporally aggregated data, has a generalisation error that is provably close to the optimal performance by the best possible linear model that can be learned from the non-aggregated data set.

We then focus our attention on the complementary problem that involves designing aggregation strategies that can allow learning, as well as developing algorithmic techniques that can use only the aggregates to train a model that works on individual samples. We motivate our methods by using the example of Gaussian regression, and subsequently extend our techniques to subsume binary classifiers and generalised linear models. We deonstrate the effectiveness of our techniques with empirical evaluation on data from healthcare and telecommunication.

Finally, we present a concrete example of our methods applied to a real life

practical problem. Specifically, we consider an application in the domain of online advertising where the complexity of bidding strategies require accurate estimates of most probable cost-per-click or CPC incurred by advertisers, but the data used for training these CPC prediction models are only available as aggregated invoices supplied by an ad publisher on a daily or hourly basis. We introduce a novel learning framework that can use aggregates computed at varying levels of granularity for building individual-level predictive models. We generalise our modelling and algorithmic framework to handle data from diverse domains, and extend our techniques to cover arbitrary aggregation paradigms like sliding windows and overlapping/non-uniform aggregation. We show empirical evidence for the efficacy of our techniques with experiments on both synthetic data and real data from the online advertising domain as well as healthcare to demonstrate the wider applicability of our framework.

Table of Contents

Acknowledgments v

Abstract vii

List of Tables xv

List of Figures xvi

Chapter 1. Introduction 1

1.1 Data Aggregation in Various Contexts . . . 2
1.1.1 Privacy preservation . . . 2
1.1.2 Communication bottlenecks . . . 2
1.1.3 Limitations of data collection . . . 3
1.1.4 Robustness . . . 3
1.1.5 Spatio-temporal Applications . . . 4
1.1.6 Proprietary Data Protection . . . 4
1.1.7 Security and Legal Regulations . . . 5
1.2 Ecological Fallacy . . . 5
1.3 Learning from Aggregated Data . . . 7
1.3.1 Organisation . . . 8

Chapter 2. Background and Related Work 12

2.1 Sample Statistics and Histograms . . . 12
2.2 Exponential Family Distributions . . . 14
2.3 Bregman Divergences . . . 15
2.4 Generalised Linear Models . . . 16
2.5 Compressed Sensing . . . 17
2.6 Fourier Transforms and Frequency Domain Analysis . . . 20
2.6.1 Decay Rates . . . 22

2.6.2 Multidimensional Fourier Transform . . . 24
2.6.3 Related Work . . . 26

Chapter 3. Learning from Histogram Aggregated Data 28
3.1 Introduction . . . 28
3.2 Problem Description . . . 32
3.2.1 Estimation under a Single Order Statistic constraint . . . 33
3.2.1.1 Solution in terms of $\mathbf{z}$ . . . 35
3.2.2 Histogram Constraints . . . 37
3.2.3 Blockwise Order Statistic Constraints . . . 38
3.3 Experiments . . . 38
3.3.1 Simulated Data . . . 39
3.3.2 DE-SynPUF dataset . . . 42
3.3.3 Texas Inpatient Discharge dataset . . . 43
3.4 Conclusion . . . 44

Chapter 4. Recovery of Sparse Parameter from Group-Wise Aggregated Data 45
4.1 Introduction . . . 45
4.2 Parameter Recovery from Exact Means . . . 48
4.2.1 Estimation from True Means using Compressed Sensing . . . 50
4.2.2 Empirical Mean Estimates and Aggregation Error . . . 52
4.3 Parameter Recovery from Approximate Means . . . 53
4.3.1 Noise-Free Observations . . . 54
4.3.2 Observations with Noise . . . 56
4.3.3 Extension to Histogram Aggregation . . . 58
4.4 Experiments . . . 61
4.4.1 Synthetic Data . . . 62
4.4.2 Real datasets - DE-SynPUF and TxID . . . 62
4.5 Discussion . . . 66
4.5.1 Extensions . . . 66
4.5.2 Higher Order Moments . . . 66
4.6 Conclusion . . . 67

Chapter 5. Frequency Domain Predictive Modelling with Spatio-Temporally Aggregated Data **68**

5.1 Introduction . . . 68

5.1.1 Contributions . . . 70

5.1.2 Related Work . . . 71

5.2 Problem Setup . . . 73

5.2.1 Data Aggregation in Time Series . . . 75

5.3 Frequency Domain Parameter Estimation from Spatio-Temporally Aggregated Data . . . 75

5.3.1 Frequency Domain Representation of Aggregated Time-Series Data . . . 76

5.3.2 Formulation and Estimation Algorithm . . . 77

5.3.3 Aliasing and Approximation Effects . . . 81

5.4 Discussion and Extensions . . . 83

5.5 Experiments . . . 86

5.6 Conclusion . . . 90

Chapter 6. Aggregation Paradigms and Learning with Sensitive Data **91**

6.1 Introduction . . . 91

6.2 Problem Definition . . . 95

6.3 Aggregation Design Paradigms . . . 96

6.3.1 Binary Classifiers . . . 97

6.3.2 Generalised Linear Models and Exponential Family Distributions . . . 99

6.4 Discussion on Privacy . . . 104

6.5 Experiments . . . 107

6.5.1 Binary Classification: Churn in Telecom: . . . 108

6.5.2 Real-valued data: Healthcare . . . 111

6.6 Conclusion . . . 112

Chapter 7. Predicting Cost-per-Click in Online Advertising from Aggregated Invoices **113**
7.1 Introduction . . . 113
7.2 Problem Description . . . 117
7.3 General Formulation . . . 120
7.3.1 Generalised Linear Models . . . 121
7.3.1.1 Loss Function: Bregman Divergences . . . 121
7.3.2 Extension to Overlapping Aggregation . . . 123
7.3.3 Learning Algorithm for the General Case . . . 124
7.4 Experiments . . . 127
7.4.1 Synthetic Data . . . 129
7.4.2 Real Data: CPC in Online Advertising . . . 131
7.4.3 Real Data: Healthcare . . . 133
7.5 Conclusion . . . 136

Chapter 8. Conclusion and Future Work **138**
8.1 Summary of the Dissertation . . . 138
8.2 Future Directions of Research . . . 141

Appendices **145**

Appendix A. Recovery of Sparse Parameter from Group-Wise Aggregated Data: Appendix **146**
A.1 General Remarks on the Results . . . 146
A.2 Proofs of Main Results . . . 149
A.2.1 Proof of Theorem 4.3.1 . . . 153
A.2.2 Proof of Theorem 4.3.2 . . . 155
A.2.3 Proof of Theorem 4.3.3 . . . 156
A.3 Higher Order Moments . . . 157

Appendix B. Frequency Domain Predictive Modelling with Spatio-Temporally Aggregated Data: Appendix **162**
B.1 Frequency Domain Formulation . . . 162
B.2 Proofs of Main Results . . . 164

B.2.1 Proof of Theorem 5.3.1 169
B.2.2 Proof of Theorem 5.3.2 171
B.3 Discussion: Decay Rates 179

Appendix C. Aggregation Paradigms and Learning with Sensitive Data: Appendix **181**
C.1 Proof of Proposition 6.3.1 181
C.2 Proof of Proposition 6.3.2 183

Bibliography **185**

List of Tables

2.1 Examples of Bregman Divergences 15

6.1 Final Training and Test Error on all three datasets (with full data used) for learner with non-aggregate data, our method, SGD and naive averaging. Our method outperforms baseline and has performance very close to learner with full, non-aggregated dataset 110

List of Figures

3.1 Permutation tests under Poisson, Gaussian and Binomial Estimation for 2, 5, 25 bins (top left, top right, bottom left) and "No Relationship" (bottom right) . . . 38
(a) Poisson Fit Error . . . 38
(b) Gaussian Fit Error . . . 38
(c) Binomial Fit Error . . . 38
3.2 Training Error under Poisson, Gaussian and Binomial Estimation . . . 39
(a) Poisson Training Error . . . 39
(b) Gaussian Training Error . . . 39
(c) Binomial Training Error . . . 39
3.3 Test Set Error under Poisson, Gaussian and Binomial Estimation . . . 39
(a) Poisson Test Error . . . 39
(b) Gaussian Test Error . . . 39
(c) Binomial Test Error . . . 39
3.4 Performance on SynPUF dataset . . . 40
(a) Training Set Error . . . 40
(b) Test Set Error . . . 40
3.5 Performance on Texas Inpatient Discharge dataset . . . 40
(a) Training Set Error . . . 40
(b) Test Set Error . . . 40
3.6 Recovered Histograms of both datasets (true histograms on the left) . . . 41
(a) Recovered Histogram of DE-SynPUF Data . . . 41
(b) Recovered Histogram of Texas Discharged Data . . . 41

4.1 Performance on Gaussian model with increasing number of datapoints in each group . . . 63
(a) Parameter Recovery . . . 63
(b) Signed Support Recovery . . . 63

4.2 Performance on Bernoulli model with increasing number of datapoints in each group . . . 63
(a) Parameter Recovery . . . 63
(b) Signed Support Recovery . . . 63
4.3 Performance on DESynPUF dataset with increasing number of datapoints in each group . . . 65
(a) Parameter Recovery Error . . . 65
(b) Support Recovery Error . . . 65
4.4 Performance on TxID dataset with increasing number of datapoints in each group . . . 65
(a) Parameter Recovery Error . . . 65
(b) Support Recovery Error . . . 65

5.1 Results on Synthetic Data – Mean Estimation Error with increasing Fourier Window ω_0 for uniform aggregation (5.1a), and non-uniform aggregation with increasing discrepancy among aggregation periodicities (5.1b through 5.1d). Frequency domain parameter estimation outperforms naive application of time domain methods . . . 86
(a) Uniform Sampling . . . 86
(b) Low Discrepancy . . . 86
(c) Medium Discrepancy . . . 86
(d) High Discrepancy . . . 86
5.2 Results on Forest Fires Dataset, Las Rosas Datasets show that frequency domain parameter estimation outperforms naive application of time domain methods and approaches the optimal for high enough ω_0. If ω_0 is too large, however, aliasing effects can lead to deteriorated performance as in Figure 5.2b . . . 87
(a) Las Rosas Dataset . . . 87
(b) Forest Fires Dataset . . . 87
(c) CCDS Dataset . . . 87

6.1 Training and Test error vs Number of Data Chunks on IBM, Kaggle and DESynPUF datasets: even with very few data chunks, our algorithm performs comparable to or even better than that obtained from a non-aggregated dataset **Note 1:** Our algorithm does better than a binary classifier trained with non-aggregated data, exactly as predicted by Prop 6.3.1. **Note 2:** Training error is shown here since our algorithm does not has full access to the training dataset . . . 109

(a) Train Error on IBM (churn) dataset . . . 109
(b) Train Error on Kaggle (churn) dataset . . . 109
(c) Train Error on DESynPUF . . . 109
(d) Train Error on IBM (churn) dataset . . . 109
(e) Train Error on Kaggle (churn) dataset . . . 109
(f) Train Error on DESynPUF . . . 109

7.1 Synthetic Data (Gaussian): Error on predictions for test data, error in reconstructed training data and estimation error for parameter recovery plotted vs iteration for Gaussian Model. Our model outperforms the baseline in all three metrics and converges within very few iterations 128
(a) Gaussian Test Error . . . 128
(b) Gaussian Train Error . . . 128
(c) Gaussian Parameter Estimation . . . 128

7.2 Synthetic Data (Poisson): Error on predictions for test data, error in reconstructed training data and estimation error for parameter recovery plotted vs iteration for a Poisson Regression Model. Our model outperforms the baseline in all three metrics and converges within very few iterations . . . 128
(a) Poisson Test Error . . . 128
(b) Poisson Train Error . . . 128
(c) Poisson Parameter Estimation . . . 128

7.3 Real Data: Estimating CPC for Online Advertising: Error on predictions for test data plotted vs iteration for a Log-Normal Model, for aggregation period limited to 30 clicks (figure 7.3a) and 35 clicks (figure 7.3b). Errors for both cases shown in figure 7.3c, scaled for ease of comparison. Our model outperforms the baseline, leading to nearly 4-5% improvement in predictive performance . . . 131
(a) Test Error for Aggregation over 30 clicks . . . 131
(b) Test Error for Aggregation over 35 clicks . . . 131
(c) Scaled Test Error Comparison . . . 131

7.4 Real Data: Estimation of Medicare Reimbursement Using CMS Data: Error on predictions for test data and error in reconstructed training data plotted vs iteration, as estimated using a Gaussian Model. Our model outperforms the baseline and converges within very few iterations, with a reasonably faithful reconstruction of the training data 134
(a) DESynPUF Test Error . . . 134

(b) DESynPUF Train Error 134
7.5 Real Data: Estimation of Texas State Hospital Charges: Error on predictions for test data and error in reconstructed training data plotted vs iteration, as estimated using a Poisson Regression Model. Our model outperforms the baseline and converges within very few iterations, with a reasonably faithful reconstruction of the training data . 134
(a) TxID Test Error 134
(b) TxID Train Error 134

Chapter 1

Introduction

Modern life is highly data driven. Datasets with potential for granular, individual level predictive modelling and inference are generated every day in large volumes in fields as diverse as healthcare (Park & Ghosh, 2014; Armstrong et al., 1999), econometrics (Davidson et al., 1978), climate science (Lozano et al., 2009; Liu et al., 2010), financial forecasting (Taylor, 2007), Internet of Things (IoTs) (Da Xu et al., 2014; Li et al., 2013). This creates an opportunity for researchers and policy-makers to analyze the data and draw individual level inferences using machine learning and data mining models trained on the data.

The traditional machine learning training paradigm involves models, both parametric and non-parametric, that are learned by training them to fit a dataset of individual level training samples as best as they can, up to generalisation constraints. Therefore, the typical machine learning setup requires access to data points that are in the form of individual-level records, e.g., account information for individual customers, or health records for individual patients, etc. However, in many domains, access to such individual records are severely restricted due to a variety of reasons. Instead, the data available to practitioners for training their machine learning models is often available only in an aggregated form or as summaries.

1.1 Data Aggregation in Various Contexts

Data aggregation and other statistical disclosure limitation techniques are an immensely popular technique in varied contexts including, but not limited to, the following

1.1.1 Privacy preservation

Aggregation is a common strategy for sharing of sensitive data in the healthcare industry where regulations and ethics guidelines protecting the privacy of patients restricts public access to granular information about individuals. Sensitive patient information is subject to various Statistical Disclosure Limitation (SDL) techniques (Armstrong et al., 1999; Duncan et al.) before public release, and data aggregation is a common statistical disclosure limitation technique to enable learning while preserving user anonymity. There are lots of applications in domains like healthcare (e.g. patient records), or users and communities in social media (e.g. purchases on e-commerce websites, or sample statistics about consumed content on Hulu, Netflix, Spotify, etc. together with information about social network graphs).

1.1.2 Communication bottlenecks

Large scale data is inherently difficult to transport, hence data is often aggregated before transferring between different nodes. Census data and other large scale data collection programs like the General Social Survey (GSS) collect and report data in aggregated form (NORC). Similarly, data collected and released by the Bureau of Labour Statistics (US Department of Labour) and Bureau of Economic

Analysis (US Department of Commerce) are often aggregated for ease of use.

1.1.3 Limitations of data collection

Some kinds of data simply cannot be collected with sufficient information granularity. This is especially true for dynamic data collection, where point by point snapshots may not be available, or reliable (for example, a person may rate his experience with an airline or an e-commerce portal as "overall unsatisfactory", without elaborating on specific positives or negative- the timeline of his interactions with the portal might still be available, but not his "rating" for each step in the timeline).

1.1.4 Robustness

Aggregated data is known to be more robust to noise and interference. A lot of decentralised sensor networks or user behavious therefore report data in aggregated form, which are more reliable for analysis and inferential application than raw non-aggregated data. Potential applications include physical sensor networks that monitor temperature, humidity, etc. for agriculture or meteorology, or crowdsourced information collection for, say restaurants ratings or real time waiting periods, or current traffic information, for example. Data from IoTs and other distributed sensor networks are often collected in aggregated form to mitigate communication costs, and improve robustness to noise and malicious interference (Wagner, 2004; Zhao et al., 2003)

1.1.5 Spatio-temporal Applications

This is another context in which aggregation shows up. In many real life cases (Burrell et al., 2004; Lozano et al., 2009; Davidson et al., 1978) instead of releasing granular datasets involving individual samples with localised measurements, the data that is collected is publicly reported only as spatio-temporally aggregated averages, collected over specific intervals and released periodically. For example, data released by the Bureau of Labour Statistics (US Department of Labour) and Bureau of Economic Analysis (US Department of Commerce) are often in this form. Analysis of such data with complex strucutral correlations is an important and ever present problem in many disciplines with applications in diverse and wide-ranging fields.

1.1.6 Proprietary Data Protection

It is very common in industry for two or more different companies work together on a common platform to provide a service to their clients or customers. In such a case, effective service would require data sharing between the companies, but to protect proprietary ownership and the integrity of private data, a company cannot provide carte blanche access to their own individual records to third parties. In such a case, companies often aggregate the data before they share it with external partners. For example, in the multi-billion dollar online advertising industry (Zeff & Aronson, 1999; Yan et al., 2009), ad publishers (like Google or Facebook) have to work with advertisers (like Criteo) to ensure that the right ad reaches the right customer. The backbone of the entire process is a partially observed auction mechanism, where

advertisers bid for ad space on publishers' domains using their proprietary bidding algorithm, and publishers charge advertisers using their own cost-per-click (CPC) formula. To protect their algorithms and retain leverage during negotiation, each side only provides aggregated user propensity or CPC data to the other side, and models for each have to be learned at the individual ad or individual user level using only these aggregates.

1.1.7 Security and Legal Regulations

Data security is a critical consideration in many sensitive applications, including healthcare, e-commerce, and Telecom. In these domains, storage and transportation of individual level records can be problematic due to both ethical concerns as well as vulnerability to leakage or hacks. In such domains, data can only be accessed at the individual level in small chunks, after which the data needs to be aggregated and all individual records purged from the system. For example, in the telecom domain (Breyer, 2005; Brown, 2010), federal regulations prohibit the storage of user data in identifiable form beyond a specific period of time, and all future models need to be learned only using these aggregates.

1.2 Ecological Fallacy

One of the main challenges in learning from aggregated is the phenomenon of Ecological Fallacy, wherein inferences drawn by analysing a system at the group or aggregate level differs significantly from the ground truth at the individual level. An especially egregious variation of this phenomenon is Simpson's paradox (Wagner,

1982; Kievit et al., 2013), where trends that can be observed in data at a lower level of granularity reverse themselves when the data is viewed at an aggregate level.

A particularly famous example of Simpson's paradox occurred in the context of undergraduate admissions at the University of California in Berkeley for Fall of 1973 (Bickel et al., 1975), when comparison of the proportions of women accepted for admission to the University as a whole was found to be less than the corresponding proportion of men admitted. This analysis pointed towards the possibility of gender bias against women. However, when the data was broken down by Department, this inference was found to be highly misleading– in fact, the proportion of women accepted turned out to be comparable or marginally higher for most departments. The discrepancy was caused by the fact that during that particular semester, most women happened to apply for admission to highly competitive departments, while most men applied to departments that admitted a larger proportion of students overall. Hence, when the data was aggregated across all applicants to the University as a whole, the proportion of men admitted was inflated.

Another well known example of Simpson's paradox occurred (Cohen & Nagel, 1957) in the context of tuberculosis related deaths in the cities of Richmond, Virginia and New York, New York, during the year 1910. While New York city had a lower mortality rate overall compared to Richmond, when the data was separated by race into white and non-white categories, the mortality rate for each category was found to be lower in Richmond. The resolving explanation was the same as in the Berkeley case– unequal racial distribution for the two cities resulted in the discrepancy between the aggregate trend and the trend at the sub-group level.

These examples, and many others throughout history, demonstrate the necessity of caution in using aggregate level data analysis to make inferences at the individual level without adequate due diligence in ensuring the transferability of the analysis. To avoid these pitfalls, while we train our models using aggregated data, for all methods presented throughout this thesis we compare the performance of our model when applied to data at the individual level.

1.3 Learning from Aggregated Data

This aggregated data setup is a relatively new form of semi-supervision, which requires novel techniques and significant algorithmic innovation on the part of data analysts to perform modeling and inference. However, despite its ubiquity, there has not been sufficient attention devoted to problems in this domain. In this thesis we investigate machine learning and data mining in various settings involving data aggregation, and rectify in part the lacunae in research on aggregated data. We consider various forms of aggregation paradigms and develop novel algorithmic techniques to offset some of the challenges that accompany data analysis under each such setup. We also design new aggregation frameworks as well as corresponding learning protocols that allow for training a model while preserving privacy. Finally, to bridge the gap between esoteric mathematical theory and out-of-domain stakeholders, we introduce new criteria for measuring privacy that is both stringent yet easily comprehensible without a strong foundation in rigorous mathematical disciplines. We further show how our aggregation paradigms satisfy these new criteria in privacy preservation.

1.3.1 Organisation

The rest of the thesis is organised as follows:

Chapter 2 introduces the requisite background and preliminaries that form the foundation of most of the work in the thesis. It also describes some related work that form complementary reading to the material covered by the thesis.

Chapter 3 studies the case where all covariates or feature variables are fully observed at individual level (e.g. demographic information), but the target variables of interest is in aggregated form (e.g. privacy-sensitive patient information). We study the variation of this problem is when the target variables are aggregated in a non-linear manner, specifically when they are available as order statistics like medians and quartiles, or available as histograms of arbitrary granularity. We develop novel techniques to learn under such constraints in the context of generalised linear models, and demonstrate its efficacy using empirical evaluation of parameter estimation fidelity and estimation error on synthetic data, as well as predictive accuracy on real data from the healthcare domain.

Chapter 4 looks at the setup where both the features/covariates and targets are aggregated group-wise– specifically, they only known up to their respective group-wise means. This section investigates when parameter recovery is possible in such a scenario, in the context of linear regression models. The main result is that under certain standard incoherence conditions on the matrix of feature variable, and structural constraints on the regression parameter, parameter recovery is still possible from only the aggregated data. Recovery is exact for the noiseless case, and approx-

imate up to arbitrary tolerance when the measurements are corrupted with random noise. In the special case where the data is aggregated into histograms, recovery is possible within a tolerance that depends on the granularity of the histograms. Experimental results on synthetic data is used to corroborate the theoretical results, and further empirical evaluation on two healthcare datasets demonstrated the relevance of the results on real life applications.

Chapter 5 investigates the problem of building predictive models that can be trained using data that is subjected to a spatio-temporal aggregation procedure before being publicly released. Aggregated time series data is ubiquitous in domains like econometrics and healthcare, and also in recommendation systems that use user history. The problem with its arbitrary correlation structures poses its unique set of challenges, but also provides exploitable structural properties. This chapter shows how to bypass the inherent structural problems in aggregated spatio-temporal data by introducing a novel framework to perform predictive modeling in the frequency domain, and uses duality properties of Fourier analysis to design a new algorithm with strong theoretical guarantees. We provide experimental results on synthetically generated data to corroborate our theoretical results. Empirical evaluation on three real world datasets showed significant improvement in performance compared to naive time-domain techniques.

Chapter 6 looks at the complementary problem, where the objective is to design aggregation strategies that can still allow learning predictive models. The motivation for this comes from concrete practical problems that are common in

many real-life modern applications, where considerations like privacy, security and legal doctrines like the GDPR put limitations on data storage and sharing with third parties. We bypass these constraints by designing aggregation paradigms that conform to privacy or non-identifiability requirements, while at the same time designing learning algorithms that can nevertheless be used to learn from only the aggregates. We delineate our framework for the case of Gaussian regression, and extend our techniques to subsume arbitrary binary classifiers and generalised linear models. We provide theoretical results and empirical evaluation of our methods on real data from healthcare and telecom. Finally, noting that existing metrics for privacy can often be too esoteric for wide applicability in non-mathematically grounded domains, we introduce a new criterion for measuring privacy that is more stringent but easily comprehensible with minimal mathematical background, and we demonstrate how our framework satisfies these new constraints.

Chapter 7 describes the application of our techniques for learning from aggregated data on a concrete real world problem. Specifically, we consider the problem of learning individual level predictive models when the target variables used for training are only available averaged over varying aggregation windows. In particular, this problem is a critical bottleneck in designing effective bidding strategies in the context of online advertising where ground-truth cost-per-click (CPC) data is aggregated to protect proprietary information before being released to advertisers. While accurate estimates of cost-per-click or CPC expenses at individual click level are required in the bidding process, for various reasons the ground-truth CPC data for training these models is usually available only as aggregates over a certain time period (daily,

hourly, etc.). We introduce a novel learning framework that can use aggregates computed at varying levels of granularity for building individual-level predictive models. We generalise our modelling and algorithmic framework to handle data from diverse domains, and extend our techniques to cover arbitrary aggregation paradigms like sliding windows and overlapping/non-uniform aggregation. We show empirical evidence for the efficacy of our techniques with experiments on both synthetic data and real data from the online advertising domain as well as healthcare to demonstrate the wider applicability of our framework.

Finally, Chapter 8 rounds out the thesis with a summary of the main details covered in the thesis and provides concluding remarks putting our results in broader context. Aggregated data arises in a vast number of domains and in a wide variety of avatars, and it is impossible to cover all possible bases in a single dissertation, if they can be solved at all, regardless of the size and scope of the study. We therefore include some thoughts on future directions of research in this area and provide some potential directions and ideas on how to tackle the challenges that are yet open problems.

Chapter 2

Background and Related Work

The objective in a standard machine learning program is to learn a mapping function $\boldsymbol{\phi} : \mathbf{X} \mapsto y$ that maps features $\mathbf{X}$ to targets y. In our work we mostly concern ourselves with parametric models, where the targets y are related to the features $\mathbf{X}$ via a vector parameter $\boldsymbol{\beta}$, that is, $y \sim \boldsymbol{\phi}(\mathbf{X}\boldsymbol{\beta})$. The standard setup– one that has been studied extensively in existing literature– deals with the case where both y and X are available as individual level samples, in non-aggregated form.

The next few sections provide a short overview of concepts and techniques that we shall use in our work.

2.1 Sample Statistics and Histograms

Aggregated data is often summarized using a sample statistic, which provides a succinct descriptive summary (Wilks, 1962). Examples of sample statistics include the average, median and various other quantiles. While the mean is still the most common choice, the best choice for summarizing a sample generally depends on the distribution the sample has been generated from. In many cases, the use of histograms Scott (1979b) or order statistic summaries is much more "natural" e.g. for

categorical data, binary data, count valued data, etc.

Order Statistics: Given a sample of n real valued datapoints, the τ^{th} order statistic of the sample is the τ^{th} smallest value in the sample. For example, the first order statistic is the minimum value of the sample, the $\frac{n}{2}^{th}$ order statistic is the median and the n^{th} order statistic is the maximum value of the sample. We specifically design a framework which makes it relatively straightforward to work with order statistics.

Histograms : Given a finite sample of n items from a set $\mathcal{C}$, a histogram is a partition of the set $\mathcal{C}$ into disjoint bins $C_i : \cup_i C_i = \mathcal{C}$ and the respective count or percentage of elements from the sample in each bin. Seen this way, for any sample from $\mathcal{C} \subseteq \mathbb{R}$, a histogram is essentially a set of order statistics for that sample. Histograms can sometimes be specified without their boundary values (eg. "$x < 30$" as opposed to "$0 < x < 30$")- this is equivalent to leaving out the first and the n^{th} order statistic. Further, a set of sample statistic summaries are easily converted to (and from) a discrete cumulative distribution by identifying the quantile value as the cumulative histogram boundary, and the quantile identity as the height. This cumulative histogram is easily converted to a standard histogram by differencing of adjacent bins. A similar strategy is also applicable to unbounded domains using abstract *max* and *min* boundaries of $\pm\infty$. Based on this bijection, we we will refer to a histogram as a generalization of the order statistic for the remainder of this manuscript.

2.2 Exponential Family Distributions

The exponential family of distributions are a large class of ubiquitously probability measures that share a common parametric form. As a generalisation of the Gaussian distribution, the exponential family subsumes many standard distributions like Poisson, binomial, negative binomial, chi-squared, pareto, etc.

A random variably $Y \in \mathcal{Y}$ is said to be in the exponential family in canonical form if its probability distribution takes the following form for any $y \in \mathcal{Y}$:

$$P_{\boldsymbol{\phi}}(y|\mu) \propto h(y)\ exp\left(t(y) \cdot \mu - \mathbf{G}_{\boldsymbol{\phi}}(\mu)\right) \tag{2.1}$$

where $\boldsymbol{\mu} \in t(\mathcal{Y})$ is called the mean parameter, h is known as the base measure, $t(\cdot)$ is known as the sufficient statistic, and $\mathbf{G}_{\boldsymbol{\phi}}$ is a function known as the log-partition function. In particular, whenever defined, we have

$$\mathbf{G}_{\boldsymbol{\phi}}(\mu) = -log\left(\int_{y \in \mathcal{Y}} h(y) exp(t(y) \cdot \mu)\right)$$

A key property of the log-partition function for exponential family distributions is that all moments of the sufficient statistic of the random variable $t(Y)$ can be obtained by successive differentiation of the log partition function. In particular, we use the property that

$$g_{\phi}(\mu) \equiv \nabla \mathbf{G}_{\boldsymbol{\phi}}(\mu) = E[t(Y)|\mu]$$

Without loss of generality, for the rest of this thesis, we shall assume $t(\cdot)$ to be the identity function. A detailed analysis of exponential family distributions can be found in Banerjee et al. (2005).

2.3 Bregman Divergences

Let $\boldsymbol{\phi} : \Theta \mapsto \mathbb{R}$ be a strictly convex, closed function on the domain $\Theta \subseteq \mathbb{R}^m$ which is differentiable on $\text{int}(\Theta)$. Then, the Bregman divergence $D_{\boldsymbol{\phi}}(\cdot\|\cdot)$ corresponding to the function $\boldsymbol{\phi}$ is defined as

$$D_{\boldsymbol{\phi}}(\mathbf{y}\|\mathbf{x}) \triangleq \boldsymbol{\phi}(\mathbf{y}) - \boldsymbol{\phi}(\mathbf{x}) - \langle \nabla\boldsymbol{\phi}(\mathbf{x}), \mathbf{y} - \mathbf{x}\rangle$$

From strict convexity, it follows that $D_{\boldsymbol{\phi}}(\mathbf{y}\|\mathbf{x}) \geq 0$ and $D_{\boldsymbol{\phi}}(\mathbf{y}\|\mathbf{x}) = 0$ if and only if $\mathbf{y} = \mathbf{x}$. Bregman divergences are strictly convex in their first argument but not necessarily in their second argument. In this paper we only consider convex functions of the form $\boldsymbol{\phi}(\cdot) : \mathbf{R}^m \ni \mathbf{x} \mapsto \sum_i \phi(x^{(i)})$ that are sums of identical scalar convex functions applied to each component of the vector $\mathbf{x}$. We refer to this class as *identically separable* (**IS**). Square loss, Kullback-Leibler (KL) divergence and generalized I-Divergence (GI) are members of this family (Table 2.1).

Table 2.1: Examples of Bregman Divergences

$\boldsymbol{\phi}(\mathbf{x})$	$D_{\boldsymbol{\phi}}(\mathbf{y}\|\mathbf{x})$
$\frac{1}{2}\|\mathbf{x}\|^2$, with $\mathbf{x} \in \mathbb{R}^n$	$\text{SQ}(\mathbf{y}\|\mathbf{x}) = \frac{1}{2}\|\mathbf{y} - \mathbf{x}\|^2$ (Square Loss)
$\sum_i (x^{(i)} \log x^{(i)})$ with $\mathbf{x} \in$ Prob. Simplex	$\text{KL}(\mathbf{y}\|\mathbf{x}) = \sum_i \left(y^{(i)} \log(\frac{y^{(i)}}{x^{(i)}}) \right)$ (Kullback-Leibler Divergence)
$\sum_i x^{(i)} \log x^{(i)} - x^{(i)}$ with $\mathbf{x} \in \mathbb{R}^n_+$	$\text{GI}(\mathbf{y}\|\mathbf{x}) = \sum_i y^{(i)} \log(\frac{y^{(i)}}{x^{(i)}}) - y^{(i)} + x^{(i)}$ (Generalised Itakura-Saito Distance)

As shown in Banerjee et al. (2005), each Bregman divergence has a one-to-one relationship with a distribution in the exponential family, parametrized by the convex function $\boldsymbol{\phi}$.

2.4 Generalised Linear Models

While least squares regression is useful for modeling continuous real valued data generated from a Gaussian distribution. This is not always a valid assumption. In many cases, the data of interest may be binary valued or count valued. A generalized linear model (GLM) (McCullagh & Nelder, 1989; Nelder & Wedderburn, 1972) is a generalization of linear regression that subsumes various models like Poisson regression, logistic regression, etc. as special cases. A generalized linear model assumes that the response variables, y are generated from a distribution in the exponential family with the mean parameter related via a monotonic link function to a linear function of the predictor $\mathbf{x}$. The model therefore is specified completely by a distribution $P_{\boldsymbol{\phi}}(\cdot \mid \boldsymbol{\theta})$ from the exponential family, a linear predictor $\eta = \mathbf{x}\boldsymbol{\theta}$, and a link function $\mathbf{g}_{\boldsymbol{\phi}}^{-1} \equiv (\nabla\phi)^{-1}(\cdot)$ which connects the expectation parameter of the response variable to a linear function of the predictor variables as $E[y|\mathbf{x}, \boldsymbol{\theta}] = (\nabla\phi)^{-1}(\mathbf{x}\boldsymbol{\theta})$

In particular, given a predictor $\mathbf{x}$, a parameter $\boldsymbol{\theta}$, and an exponential family distribution $P_{\boldsymbol{\phi}}$, the likelihood of the target y has the following form:

$$P_{\boldsymbol{\phi}}(y|\mathbf{x}, \boldsymbol{\theta}) = h(y) \cdot \ exp\left(y \cdot \mathbf{x}^\top \boldsymbol{\theta} - \mathbf{G}_{\boldsymbol{\phi}}(\mathbf{x}^\top \boldsymbol{\theta})\right) \tag{2.2}$$

where $h(\cdot)$ is the base measure associated with the distribution $P_{\boldsymbol{\phi}}$

As explored in great detail in Banerjee et al. (2005), Bregman Divergences have a very close relationship with generalized linear models. In particular, maximum likelihood parameter estimation for a generalized linear model is equivalent to minimizing a corresponding Bregman divergence. For example, maximum likelihood for a Gaussian corresponds to squares loss, for Poisson the corresponding divergence is generalized I-divergence and for Binomial, the corresponding divergence is the KL divergence (see Banerjee et al. (2005) for details). GLMs have been successfully applied in a wide variety of fields including biological surveys (Nicholls, 1989), image segmentation and reconstruction (Paul et al., 2013), analysis of medical trials (Dias et al., 2013), studying species-environment relationships in ecological sciences (Jamil et al., 2013), virology (Gart, 1964) and estimating mortality from infectious diseases (Hardelid et al., 2013), among many others, and are widely prized for the interpretability of their results and the extendability of their methods in a plethora of domain specific variations (Song et al., 2013). They are easy to use and implement and many off-the-shelf software packages are available for most major programming platforms.

2.5 Compressed Sensing

The case of under-determined linear systems appears in a large number of practical applications and is a very well studied problem in the compressed sensing domain. The standard form of such problems involves estimating the linear model parameter $\boldsymbol{\beta}_0$ given a feature matrix $\mathbf{M} \in \mathbb{R}^{k \times d}$ and a target vector $\mathbf{y}$, such that $\mathbf{y} = \mathbf{M}\boldsymbol{\beta}_0$, when the number of rows in the feature matrix is smaller than the dimen-

sionality of the problem, $k < d$, resulting in an under-determined system.

A line of work including (Candes & Tao, 2006; Donoho, 2006), among others, have shown that subject to certain sparsity conditions on $\boldsymbol{\beta}_0$ and *restricted isometry* constraints on the matrix $\mathbf{M}$, the parameter $\boldsymbol{\beta}_0$ can be recovered uniquely as the solution to an ℓ_1 minimisation optimisation problem under observation constraints.

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1 \quad \text{s.t. } \mathbf{M}\boldsymbol{\beta} = \mathbf{y} \tag{2.3}$$

The principal condition that allows the recovery of a sparse parameter in such under-determined systems is called the restricted isometry property.

Definition 2.5.1. For a $k \times d$ matrix $\mathbf{M}$ and a set $T \subseteq \{1, 2, \cdots, d\}$, suppose $\mathbf{M}_T$ is the $k \times |T|$ matrix consisting of the columns of $\mathbf{M}$ corresponding to T. Then, the *s-restricted isometry constant* δ_s of the matrix $\mathbf{M}$ is defined as the smallest quantity δ_s such that the matrix $\mathbf{M}_T$ obeys

$$(1 - \delta_s)\|c\|_2^2 \leq \|\mathbf{M}_T c\|_2^2 \leq (1 + \delta_s)\|c\|_2^2$$

for every subset $T \subset \{1, 2, \cdots, d\}$ of size $|T| < s$ and all real $c \in \mathbb{R}^{|T|}$

Restricted isometry is a common and standard assumption in the sparse parameter recovery literature. Intuitively, this property means that when $\mathbf{M}$ satisfies Definition 2.5.1 with a small δ_s, every sub-matrix of small enough size constructed out of the columns of the matrix behaves approximately like an orthonormal system.

In fact, a number of random matrices satisfy this property including the Gaussian ensemble and the Bernoulli ensemble (Donoho, 2006; Candès et al., 2006).

Suppose we had access to the true mean matrices $(\mathbf{M}, \mathbf{y})$. First, we consider the case when observations are noise-free, i.e. $\epsilon = 0$. Suppose $\boldsymbol{\beta}_0$ is known to be κ_0-sparse and $\mathbf{M}$ satisfies the restricted isometry hypothesis, then the following result applies:

Theorem 2.5.1 (Exact Recovery (Foucart, 2010)). *Let* $\Theta_0 = \frac{3}{4+\sqrt{6}} \approx 0.465$. *If there exists an* s_0 *such that* $\delta_{2s_0} < \Theta_0$ *for* $\mathbf{M}$, *then as long as* $\kappa_0 < s_0$, *the constraint* $\mathbf{M}\boldsymbol{\beta}_0 = \mathbf{y}$ *is sufficient to uniquely recover any* κ_0*-sparse* $\boldsymbol{\beta}_0$ *exactly as the solution of the following optimization problem:*

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1 \quad s.t.\ \mathbf{M}\boldsymbol{\beta} = \mathbf{y} \tag{2.4}$$

A similar result for approximate recovery holds for the case when the observations are corrupted with noise $\boldsymbol{\epsilon}$, i.e., instead of $\mathbf{y} = \mathbf{M}\boldsymbol{\beta}_0$, we are given $\mathbf{y}_\epsilon = \mathbf{M}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}$.

Theorem 2.5.2 (Approximate Recovery (Candes, 2008)). *Let* $\Theta_1 = \sqrt{2}-1 \approx 0.414$. *If there exists an* s_0 *for* $\mathbf{M}$ *such that* $\delta_{2s_0} < \Theta_1$, *then as long as* $\kappa_0 < s_0$ *and the noise* $\boldsymbol{\epsilon}$ *in observations* $\mathbf{y}_\epsilon = \mathbf{M}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}$ *is bounded as* $\|\boldsymbol{\epsilon}\|_2 < \xi$, *any* κ_0*-sparse* $\boldsymbol{\beta}_0$ *can be recovered within an* ℓ_2 *distance of* $C_{s_0}\xi$ *from the true parameter* $\boldsymbol{\beta}_0$ *using the noisy measurements* $(\mathbf{M}, \mathbf{y}_\epsilon)$. *That is, the solution* $\widehat{\boldsymbol{\beta}}$ *to the following optimization problem:*

$$\min_{\boldsymbol{\beta}_0} \|\boldsymbol{\beta}\|_1 \ s.t.\ \|\mathbf{M}\boldsymbol{\beta} - \mathbf{y}_\epsilon\|_2 < \xi \tag{2.5}$$

satisfies $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 < C_{s_0}\xi$ *where the constant* C_{s_0} *depends only on* δ_{2s_0} *and is well-behaved (for example when* $\delta_{2s_0} = 0.2$*, the constant is less than 8.5).*

The aforementioned results are a sampling of the existing breadth of literature in the compressed sensing domain, and while we use these extensively in Chapter 4, we note that other results from this field can be used very easily to extend the analyses presented in this dissertation. In particular, various alternative frameworks like non-sparse parameter, alternative estimators to LASSO, beyond sub-gaussian assumptions on different marginals, etc. can be analysed in an identical manner, and the main results we present in this dissertation would still continue to hold, albeit with slightly different sample complexity.

2.6 Fourier Transforms and Frequency Domain Analysis

A random signal $\{z(t) \in \mathbb{R} : t \in \mathbb{R}\}$ is said to be **centered** and **weakly stationary** with **finite variance** if the following conditions hold:

1. the process is centred, $E[z(t)] = 0$ for all t
2. for any time-stamps t, t', we have $E[z(t)z(t')] = \rho_z(\|t - t'\|)$ for a non-negative real valued auto-correlation function $\rho_z(\cdot) : \mathbb{R}^+ \mapsto \mathbb{R}^+$
3. at every point, the noise process has finite variance, $E[z(t)^2] = \rho(0) < +\infty$

Stationarity is a standard assumption in time series analysis and very common in many real life applications (see (Granger & Newbold, 2014; Dzhaparidze, 2012;

Feige & Pearce, 1974)). In particular, many predictive models for standard time series analysis use methods like filtering out trend lines and differencing to ensure stationarity in the data before analysis (Hibon & Makridakis, 1997).

Given a continuous signal $z(t)$, the **Fourier Transform** of the signal with respect to a particular frequency $\omega \in \mathbb{R}$ is given by

$$Z(\omega) = \int_{\mathbb{R}} z(t) e^{-\iota\omega t} dt \tag{2.6}$$

For a signal $z(t)$, we use both $Z(\omega)$ and $\mathcal{F}z(\omega)$ to denote its Fourier transform.

We can similarly define the T-restricted **Finite Fourier Transform** $Z_T(\omega)$ for the signal $z(t)$ as

$$Z_T(\omega) = \int_{-T}^{T} z(t) e^{-\iota\omega t} dt \tag{2.7}$$

The **Power Spectral Density** $P_Z(\omega)$ of a signal $z(t)$ with respect to a particular frequency $\omega \in \mathbb{R}$ is given by

$$P_Z(\omega) = \lim_{T\uparrow\infty} \frac{1}{T} E\left[\| \int_{-T}^{T} z(t) e^{-\iota\omega t} dt \|^2 \right] \tag{2.8}$$

Let $z(t)$ be a weakly stationary process with autocovariance function $\rho_z(\tau) = E[z(t)z(t+\tau)]$. Let $\rho_z(0) = E[z(t)^2] < \infty$ be the variance of the process. We simply state the following well known results (Grafakos, 2004) without proof:

1. (Wiener-Khinchin Theorem)The power spectral density of a stationary process $z(t)$ is the Fourier Transform of its autocovariance function

$$P_Z(\omega) = \int_{-\infty}^{\infty} \rho_z(\tau) e^{-\iota\omega(\tau)} d\tau \tag{2.9}$$

2. (Corrollary of above) For a stationary process, the integral of the power spectral density gives the instantaneous variance

$$\int_{-\infty}^{\infty} P_Z(\omega)d\omega = \rho_z(0) = E[z(t)^2] \tag{2.10}$$

2.6.1 Decay Rates

In our work, we assume that the power spectral density and autocovariance function for every signal of interest exists finitely for each ω. We further assume that the autocovariance function decays rapidly with lag for all processes involved in our analysis. In essence this means that the value of the time series at any given point is highly correlated with values at points close to it in time, but the correlation decreases rapidly with values farther away in time.

For example, we may assume that $\rho_{(\cdot)}(\cdot)$ is a Schwartz function (Terzioğlu, 1969), that is $\rho(\cdot)$ and all its derivatives decay at least as fast as any inverse polynomial. That is, $\forall \alpha, \beta \in \mathbb{Z}_+^n$ we have

$$|\zeta^\alpha \frac{\partial^\beta \rho(\zeta)}{\partial \zeta^n}| \to 0 \quad as \quad |\zeta| \to \infty$$

Examples of Schwartz functions are exponential functions like $e^{-a\zeta^2}$ for $a > 0$, or any polynomial $\wp(\zeta)$ multiplied with an exponential function like $\wp(\zeta)e^{-a\zeta^2}$, or any smooth domain-restricted function $f(\zeta)$ which is 0 outside of a bounded compact subset $\zeta \in \Im \subset \mathbb{R}^n$, e.g. all time limited signals are automatically Schwartz functions.

A key property of Schwartz functions is that the Fourier Transform of a Schwartz function is itself a Schwartz function (Gröchenig & Zimmermann, 2001; Strichartz, 2003). Therefore, if we assume that the covariance functions $\rho_{(.)}(\tau)$ decays rapidly with τ for each of our signals, then their corresponding power spectral densities $P_{(.)}(\omega)$ will decay rapidly with ω, since $P = \mathcal{F}\rho$. Therefore, most of the power for our signals will be concentrated around $\omega = 0$.

We note that unlike traditional signal processing applications, we do not consider a flat power spectral density (e.g. white noise) for our noise process. This is because traditional signal processing applications assume band-limited signals of interest. Properties of the noise process outside the band are irrelevant since outputs are going to be filtered regardless, and analysis only needs to focus on effects of additive noise within the frequency band of interest. In our case, we can make no such assumption– signals need not be bandlimited and therefore we have to consider effects of noise throught the entire spectrum[1].

As we have seen, for a stationary process $z(t)$, the power spectral density P_z and the autocorrelation function ρ_z are Fourier Transform pairs.

$$P_Z(\omega) = \int_{-\infty}^{\infty} \rho_z(\tau) e^{-\iota\omega(\tau)} d\tau \tag{2.11}$$

[1]Note that a true white noise process is unrealistic because it implies infinite variance for the noise process which renders any attempt at parameter learning futile.

Rates of decay will vary from case to case depending on the exact functional forms for the autocorrelation function (or, equivalently, the power spectral density). Standard expressions can be obtained by using the fact that for a signal $z(t)$ with finite variance, $\frac{P_z(\omega)}{\int_{-\infty}^{\infty} P_z(\omega)d\omega}$ is a valid probability density function, and then using the tail probability results for the corresponding probability distribution.

For example, if ρ exhibits a Gaussian decay (analogous to normal distribution), that is, $\rho(\tau) \sim exp(-O(\tau^2))$, then P_{ε_β} also exhibits a Gaussian decay, that is $P_{\varepsilon_\beta}(\omega) \sim exp(-O(\omega^2))$, and therefore, $\xi_{\omega_0} \sim exp(-O(\omega_0^2))$. Similarly, if ρ exhibits power law/ Lorentzian decay (analogous to Cauchy distribution), that is, $\rho(\tau) \sim \frac{1}{O(\tau^2)}$, then P_{ε_β} exhibits exponential decay (Laplace distribution), that is $P_{\varepsilon_\beta}(\omega) \sim exp(-O(|\omega|))$, and therefore $\xi_{\omega_0} \sim exp(-O(|\omega_0|)$. Similar arguments can be made for other decay rates using Fourier duality.

This makes intuitive sense because the more spread out $\rho(\tau)$ is, the more "peaky" $P_{\varepsilon_\beta}(\omega)$ is and the smaller the value of ω_0 required. This means that if the error terms are well-correlated, most of the instantaneous power will be concentrated within a very small range of frequencies.

2.6.2 Multidimensional Fourier Transform

Let $\mathbb{R}^p$ be the p-dimensional Euclidean space (called the interaction space), discrete points in which are indexed by the vector $\mathbf{v} \in \mathbb{R}^p$. Signals $\mathbf{z}$ in the interaction space are random processes $\mathbf{z}(\mathbf{v})$ operating at each point in $\mathbf{v} \in \mathbb{R}^p$. For

example, for spatial processes, $p = 2$ or 3, and $\mathbf{v}$ denotes the spatial coordinates, and a typical signal $\mathbf{z}(\mathbf{v})$ may indicate the temperature or pressure at a particular point $\mathbf{v}$.

Similar to the unidimensional case, a random signal $\{\mathbf{z}(\mathbf{v}) \in \mathbb{R} : \mathbf{v} \in \mathbb{R}^p\}$ is said to be **centered** and **weakly stationary** with **finite variance** if the following conditions hold:

1. $E[\mathbf{z}(\mathbf{v})] = 0$ for all $\mathbf{v} \in \mathbb{R}^p$
2. $E[\mathbf{z}(\mathbf{v})\mathbf{z}(\mathbf{v}')] = \rho(\|\mathbf{v} - \mathbf{v}'\|)$ for a non-negative real valued auto-correlation function $\rho(\cdot)$, for every $\mathbf{v}, \mathbf{v}' \in \mathbb{R}^p$
3. $E[\mathbf{v}(\mathbf{v})^2] = \rho(0) < +\infty$

Observations for any signal $z(\mathbf{v})$ are obtained as aggregates over periodically translated (similar to a sliding window) bounded connected set $A \subset \mathbb{R}^p$ as

$$z[\mathbf{k}] = \frac{1}{Vol(A)} \int_{\mathbf{v} \in A + \mathbf{k}} z(\mathbf{v}) d\mathbf{v}$$

Given a continuous signal $z(\mathbf{v})$, for any point $\boldsymbol{\theta} = [\theta_1, \theta_2, \cdots \theta_p] \in \mathbb{R}^p$ (the "frequency" vector), the **Multidimensional Fourier Transform** is defined in a way very similar to the one-dimensional case (Tangirala, 2014; Easton; Smith & Smith, 1995), provided the integral exists

$$Z(\boldsymbol{\theta}) = \int_{\mathbb{R}^p} z(\mathbf{v}) e^{-\iota \langle \boldsymbol{\theta}, \mathbf{v} \rangle} d\mathbf{v} \tag{2.12}$$

where $\langle \cdot, \cdot \rangle$ represents the standard inner product. All the properties of Fourier Transforms that are required within the scope of our work follow exactly as in the unidimensional case (see (Easton; Smith & Smith, 1995)).

2.6.3 Related Work

Learning from aggregated data is a challenging problem, and there is extremely limited prior art on this topic. This is a new and relatively unexplored form of semi-supervision, and there are significant gaps in analysis and understanding, rectifying which is the main objective of this dissertation. Here, we delineate some of this existing work but restrict ourselves to the topic of learning from aggregated data. Further related literature that is specific to the material covered in each chapter is included within the content of the chapters themselves.

The problem of imputing individual level records from the sample mean has been studied in (Park & Ghosh, 2012) and (Park & Ghosh, 2014) among others. In particular, the paper Park & Ghosh (2014) attempts to reconstruct the individual level matrix by assuming a low rank structure and compares their framework with other approaches which include an extension of the neighborhood model (Freedman et al., 1991b) and a variation of ecological regression (Freedman et al., 1991a) for the task of imputing individual level records of the response variable. In particular, these works focus on data reconstruction rather than predictive modelling, and they further make structural assumptions on their data generation procedure for this purpose.

In the classification literature, learning from label proportions (LLP) (Quadrianto et al., 2009; Patrini et al., 2014) involves estimation of classifiers given the proportion of discrete valued labels in groups or bags of labeled targets. The authors further introduced an estimation algorithm that used the idea of sufficient statistics for paranmeter estimation through an implicit label imputation step. The main drawback of this approach is that it is restricted to cases where the targets variables can

only take values from a finite alphabet, and therefore their methods can no longer be applied to wider regression where the targets can be real-valued.

Chapter 3

Learning from Histogram Aggregated Data

3.1 Introduction

In many domains like healthcare, census reports, voting and political data, etc. it is common for applications to work with both sensitive records (like patient records, income data, etc.) as well as information in the public domain (e.g. census records, voter files, etc.) Most agencies in such areas usually report both individual level information for non-sensitive attributes together with the aggregated information in the form of sample statistics. Care must be taken in the analysis of such data, as naïve modeling with aggregated data may significantly diminish the accuracy of inferences at the individual level due the problem of ecological fallacy (Robinson, 2009) that was extensively discussed in Chapter 1. Without this due diligence, resulting conclusions at the group level may be misleading to researchers and policy makers interested in individual level inferences.

Portions of this chapter has been published as: A Bhowmik, J Ghosh, O Koyejo, "Generalized Linear Models for Aggregated Data", International Conference on Artificial Intelligence and Statistics, (AISTATS) 2015, San Diego, USA

Co-authors have participated extensively in model formulation and research methods, and have contributed in reviewing the final manuscript.

Aggregated data is often summarized using a sample statistic, which provides a succinct descriptive summary (Wilks, 1962). Examples of sample statistics include the average, median and various other quantiles. While the mean is still the most common choice, the best choice for summarizing a sample generally depends on the distribution the sample has been generated from. In many cases, the use of histograms (Scott, 1979b) or order statistic summaries (see section 2.1) is much more "natural" e.g. for categorical data, binary data, count valued data, etc. In particular, when the data has a long tail, where means and other moment-based summaries can be misleading due to being heavily affected by outliers.

Aggregated data in the form of histograms and other sample statistics are becoming more and more common. Further, most of the data that is collected relates to questions for which the respondents have only a few discrete options from which to select their answer. For example, data available from the Generalized Social Survey (NORC) are often in this form. This chapter addresses the scenario where features are provided at the individual level, but the target variables are only available as histogram aggregates or order statistics. Despite the prevalence of order-statistic and histogram aggregated data, to the best of our knowledge, this problem has not been addressed in the literature.

We consider a limiting case of generalized linear modeling when the target variables are only known up to permutation, and explore how this relates to permutation testing (Good, 2005); a standard technique for assessing statistical dependency. Based on this relationship, we propose a simple algorithm to estimate the model parameters and individual level inferences via alternating imputation and standard

generalized linear model fitting. Our results suggest the effectiveness of the proposed approach when, in the original data, permutation testing accurately ascertains the veracity of the linear relationship. The framework is extended to general histogram data with larger bins - with order statistics such as the median as a limiting case. Our experimental results suggest a diminishing returns property - when a linear relationship holds in the original data, the targets can be predicted accurately given relatively coarse histograms. Our results also suggest caution in in the widespread use of aggregation for ensuring the privacy of sensitive data.

In summary, the main contributions of this chapter are as follows:

1. We propose a framework for estimating the response variables of a generalized linear model given only a histogram aggregate summary by formulating it as an optimization problem that alternates between imputation and generalized linear model fitting.

2. We examine a limiting case of the framework when all the data is known up to permutation. Our examination suggests the effectiveness of the proposed approach when, in the original data, permutation testing accurately ascertains the veracity of the linear relationship.

3. We examine a second limiting case where only a few order statistics are provided. Our experimental results suggest a diminishing returns property - when a linear relationship holds in the original data, the targets can be predicted accurately given relatively coarse histograms.

The proposed approach is applied to the analysis of simulated datasets. In addition, we examine the Texas Inpatient Discharge dataset from the Texas Department of State Health Services (TxID, 2014) and a subset of the SynPUF dataset (DESynPUF, 2008).

Notation

Matrices are denoted by boldface capital letters, vectors by boldface lower case letters and individual elements of the vector by the same lowercase letter with the boldface removed and the index added as a superscript. $\mathbf{v}^\top$ refers to the transpose of the column vector $\mathbf{v}$. We denote column partitions using semicolons, that is, $\mathbf{M} = [\mathbf{X}; \mathbf{Y}]$ implies that the columns of the submatrices $\mathbf{X}$ and $\mathbf{Y}$ are, in order, the columns of the full matrix $\mathbf{M}$. We use $\|\cdot\|$ to denote the L_2 norm for vectors and Frobenius norm for matrices. The vector $\mathbf{v}$ is said to be in increasing order if $v^{(i)} \leq v^{(j)}$ whenever $i \leq j$, and the set of all such vectors in $\mathbb{R}^n$ is denoted with a subscripted downward pointing arrow as $\mathbb{R}^n_\downarrow$. Two vectors $\mathbf{v}$ and $\mathbf{w}$ are said to be isotonic, $\mathbf{v} \sim_\downarrow \mathbf{w}$, if $v^{(i)} \geq v^{(j)}$ if and only if $w^{(i)} \geq w^{(j)}$ for all i, j.

Chapter 2 contains an in-depth treatment of all the preliminaries required for this chapter. In particular, we encourage the reader to refer to section 2.4 for a description of Generalised Linear Models, section 2.3 for a note on Bregman Divergences, and section 2.1 for order statistics and histograms the way we use them in this chapter.

3.2 Problem Description

Consider a set of fully observed covariates $\mathbf{X} = [\mathbf{x_1}; \mathbf{x_2}; \cdots; \mathbf{x_{d-p}}] \in \mathbb{R}^{n \times (d-p)}$, and columns of response variables, $\mathbf{Z} = [\mathbf{z_1}; \mathbf{z_2}; \cdots \mathbf{z_p}] \in \mathbb{R}^{n \times p}$, which are only known only up to the respective histograms of their values (i.e., up to order statistics).

We assume that each element of $\mathbf{z_i}$ has been generated from covariates $\mathbf{X}$ according to some generalized linear model (as described in section 2.1) with parameters $\boldsymbol{\beta}_i$. The objective is to estimate the $\boldsymbol{\beta}_i$ together with $\mathbf{Z} = [\mathbf{z_1}; \mathbf{z_2}; \cdots \mathbf{z_p}]$ subject to the given order statistic constraints. Since maximum likelihood estimation in a generalized linear model is equivalent to minimizing a corresponding Bregman divergence, we choose the loss function $\mathcal{L}(\mathbf{Z}, \boldsymbol{\beta}) = D_{\boldsymbol{\phi}}\left(\mathbf{Z} \| (\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta})\right)$ to be minimized over the variables $\mathbf{Z}, \boldsymbol{\beta}$ while satisfying order statistics constraints on $\mathbf{Z}$.

Without additional structure, the regression problem for each column can be solved independently, therefore without loss of generality we assume $\mathbf{Z}$ is a single column $\mathbf{z}$. We denote the $\tau_i{}^{th}$ order statistic of $\mathbf{z}$ as s_{τ_i}, with $\tau_i \in \{\tau_1, \tau_2, \cdots \tau_h\} \subseteq [n]$, which is the set of h order statistics specified via the histogram. For simplicity, in the following section we consider estimation under a single order statistic which has been computed over the entire column. We extend it subsequently to the more general case of multiple order statistics computed over disjoint partitions.

Therefore, with Frobenius regularization terms $\mathcal{R}(\boldsymbol{\beta}) = \lambda\|\boldsymbol{\beta}\|^2$, the overall problem statement boils down to the following optimization problem:

$$\begin{aligned} \min_{\mathbf{z},\boldsymbol{\beta}} \quad & D_{\boldsymbol{\phi}}\left(\mathbf{z} \| (\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta})\right) + \lambda\|\boldsymbol{\beta}\|^2 \\ \text{s.t.} \quad & \tau_i^{th} \text{ order statistic of } \mathbf{z_i} = s_{\tau_i} \end{aligned} \tag{3.1}$$

3.2.1 Estimation under a Single Order Statistic constraint

Estimating under order statistics constraints is in general a highly non-trivial problem. It is easy to see that the set of vectors with a given order statistic is not a convex set. Therefore, the above optimization problem looks especially difficult to even represent in a concise manner in terms of $\mathbf{z}$. However, it turns out that with the following reformulation, the analysis of the problem becomes much more manageable.

We rewrite $\mathbf{z} = \mathbf{P}\mathbf{y}$ where $\mathbf{P} \in \mathbb{P}$ is a permutation matrix and $\mathbf{y}$ is a vector sorted in increasing order. Note the following-

(i) For a $\mathbf{y} \in \mathbb{R}^n_{\downarrow}$, if $\mathbf{e}^{\tau_i}$ is a row vector with 1 in the ${\tau_i}^{th}$ index and 0 everywhere else, then $\mathbf{e}^{\tau_i}\mathbf{y}$ represents the ${\tau_i}^{th}$ order statistic of $\mathbf{y}$. Since permutation does not change the value of order statistics, this is also the ${\tau_i}^{th}$ order statistic of $\mathbf{z}$

(ii) If $\mathbf{\Lambda}$ is the matrix with $\mathbf{\Lambda}_{j,j+1} = -1, \mathbf{\Lambda}_{j,j} = 1$ and $\mathbf{\Lambda}_{j,k} = 0$ for all other $j, k : (k - j) \neq 0, \pm 1$, the condition that $\mathbf{y}$ is sorted in increasing order is equivalent to the linear constraint $\mathbf{\Lambda}\mathbf{y} \leq 0$.

Putting all this together, the optimization problem (3.1) becomes the following

$$\begin{aligned} &\min_{\mathbf{P},\mathbf{y},\boldsymbol{\beta}} \; D_{\boldsymbol{\phi}}\left(\mathbf{P}\mathbf{y} \| (\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta})\right) + \mathcal{R}(\boldsymbol{\beta}) \\ &\text{s.t. } \mathbf{e}^{\tau_i}\mathbf{y} = s_{\tau_i}, \; \mathbf{\Lambda}\mathbf{y} \geq 0, \; \mathbf{P} \in \mathbb{P} \end{aligned} \tag{3.2}$$

The above optimization problem is jointly convex in $\mathbf{y}$ and $\boldsymbol{\beta}$ for a fixed $\mathbf{P}$, but the presence of $\mathbf{P}$ as a variable makes the problem much more complicated.

Therefore, we attempt to solve it iteratively for each variable in an alternating minimization framework. The update steps consist of the following for each timestep:

(i) $\boldsymbol{\beta_t} = \underset{\boldsymbol{\beta}}{\text{argmin}}\ D_{\boldsymbol{\phi}}\left(\mathbf{P_{t-1}y_{t-1}} \| (\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta})\right) + \mathcal{R}(\boldsymbol{\beta})$

(ii) $\mathbf{y_t} = \underset{\mathbf{y}}{\text{argmin}} D_{\boldsymbol{\phi}}\left(\mathbf{P_{t-1}y} \| (\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta_t})\right)$ such that $\boldsymbol{\Lambda}\mathbf{y} \leq 0$ and $\mathbf{e}^{\boldsymbol{\tau_i}}\mathbf{y} = s_{\tau_i}$

(iii) $\mathbf{P_t} = \underset{\mathbf{P}\in\mathbb{P}}{\text{argmin}}\ D_{\boldsymbol{\phi}}\left(\mathbf{Py_t} \| (\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta_t})\right)$

Step (i) is a standard generalized linear model parameter estimation problem. This problem has been studied in great detail in literature and a variety of off-the-shelf GLM solvers can be used for this. We focus instead on steps (ii) and (iii) which are much more interesting.

For (ii), note that since we assumed that $\boldsymbol{\phi}$ is identically separable, the same permutation applied to both arguments of the corresponding Bregman divergence $D_\phi(\cdot\|\cdot)$ does not change its value. For any constraint set $\mathcal{C}$, we have $\underset{\mathbf{y}\in\mathcal{C}}{\text{argmin}}\ D_{\boldsymbol{\phi}}\left(\mathbf{Py} \| (\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta_t})\right) = \underset{\mathbf{y}\in\mathcal{C}}{\text{argmin}}\ D_{\boldsymbol{\phi}}\left(\mathbf{y} \| \mathbf{P}^{-1}(\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta_t})\right)$ given[1] a fixed $\mathbf{P}, \mathbf{X}, \boldsymbol{\beta}$. Following this fact, step (ii) is a convex optimization problem in $\mathbf{y}$ and can be solved very easily.

Step (iii) is a non-convex optimization problem in general. However, for an identically separable Bregman divergence it turns out that the solution to this is remarkably simple.

Lemma 3.2.1. *The (set of) optimal permutation(s) in step (iv) above is given by-*

[1] note that for a permutation matrix $\mathbf{P}$, $\mathbf{P^{-1}} = \mathbf{P}^\top$

$$\underset{\mathbf{P}\in\mathbb{P}}{\textit{argmin}}\ D_{\boldsymbol{\phi}}\left(\mathbf{P}\mathbf{y_t}\|(\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta}_t)\right) \equiv \hat{\mathbf{P}}\ \textit{such that}\ \hat{\mathbf{P}}\mathbf{y_t} \sim_{\downarrow} (\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta}_t)$$

In other words, the optimal permutation is the one which makes $\mathbf{y_{i,t}}$ isotonic with $(\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta}_t)$. Note that the optimal permutation is not unique if $(\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta}_t)$ is not totally ordered. This is a direct application of the following result which appeared as Lemma 3 in the paper Acharyya et al. (2012).

Lemma 3.2.2. *If $x_1 \geq x_2$ and $y_1 \geq y_2$ and $\boldsymbol{\phi}(\cdot)$ is identically separable, then*

$$\begin{aligned} D_{\boldsymbol{\phi}}(\left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right]\|\left[\begin{smallmatrix} y_1 \\ y_2 \end{smallmatrix}\right]) &\leq D_{\boldsymbol{\phi}}(\left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right]\|\left[\begin{smallmatrix} y_2 \\ y_1 \end{smallmatrix}\right]),\ \textit{and} \\ D_{\boldsymbol{\phi}}(\left[\begin{smallmatrix} y_1 \\ y_2 \end{smallmatrix}\right]\|\left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right]) &\leq D_{\boldsymbol{\phi}}(\left[\begin{smallmatrix} y_2 \\ y_1 \end{smallmatrix}\right]\|\left[\begin{smallmatrix} x_1 \\ x_2 \end{smallmatrix}\right]) \end{aligned}$$

3.2.1.1 Solution in terms of z

Lemmata 3.2.1 and 3.2.2 suggest that we can optimize jointly over $\mathbf{P}$ and $\mathbf{y}$ instead of separately, since for any $\mathbf{y}$ we already know the optimal $\mathbf{P}$. Combining the optimization steps (ii) and (iii) in terms of $\mathbf{P}$ and $\mathbf{y}$, our update step for $\mathbf{z}$ in the original optimization problem is the following

$$\begin{aligned} \hat{\mathbf{z}}_\mathbf{t} &= \underset{\mathbf{z}}{\text{argmin}}\ D_{\boldsymbol{\phi}}\left(\mathbf{z}\|(\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta}_t)\right) \\ &\text{s.t.}\quad \tau_i^{th} \text{ order statistic of } \mathbf{z} = s_{\tau_i} \end{aligned} \tag{3.3}$$

It is not immediately obvious how to approach the solution to this since the constraint set for $\mathbf{z}$ is not convex. However, note that as a result of Lemma 3.2.2 it

is clear that given a fixed $\mathbf{X}$ and $\boldsymbol{\beta_t}$ if $\hat{\mathbf{z}}_\mathbf{t}$ is a solution to the subproblem (3.3), we must have $\hat{\mathbf{z}}_\mathbf{t} \sim_\downarrow (\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta}_t)$.

Therefore, instead of searching over the set of all vectors in $\mathbb{R}^n$, it is sufficient to search only in the subset of vectors that are isotonic with $(\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta}_t)$. It turns out that not only is this set convex given a fixed $\mathbf{X}, \boldsymbol{\beta}_t$, the solution for $\mathbf{z}_t$ is readily available in closed form.

Let $\boldsymbol{\Gamma_t}$ $=(\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta}_t)$. Since the Bregman Divergence is IS, without loss of generality we can assume that $\boldsymbol{\Gamma_t}$ is in increasing order, therefore the constraint set for $\mathbf{z}$ becomes $\mathbf{z} \in \mathbb{R}^n_\downarrow$ and order statistics constraints for $\mathbf{z}$ becomes the linear constraint $\mathbf{e}^{\boldsymbol{\tau_i}}\mathbf{z} = s_{\tau_i}$.

Therefore, the optimization problem (3.3) over $\mathbf{z}$ is equivalent, up to a simple re-permutation step, to the following

$$\begin{aligned} &\min_{\mathbf{z}} \; D_\phi(\mathbf{z}\|\boldsymbol{\Gamma_t}) \\ &\text{s.t. } \mathbf{z} \in \mathbb{R}^n_\downarrow, \; \mathbf{e}^{\boldsymbol{\tau_i}}\mathbf{z} = s_{\tau_i} \end{aligned} \tag{3.4}$$

Lemma 3.2.3. *Let $\hat{\mathbf{z}}$ be the solution to the optimization problem (3.4). Then, $\hat{\mathbf{z}}$ is given by-*

$$\hat{z}_t^{(j)} = \begin{cases} s_{\tau_i} & j = \tau_i \\ \max(\Gamma_t^{(j)}, s_{\tau_i}) & j > \tau_i \\ \min(\Gamma_t^{(j)}, s_{\tau_i}) & j < \tau_i \end{cases} \tag{3.5}$$

Sketch of Proof In the space of all $\mathbf{z}$ ordered in increasing order, the τ_i^{th} order statistic constraint simply becomes $\hat{z}_t^{(j)} < s_{\tau_i}$ for $j < \tau_i$ and vice versa for $j > \tau_i$. Suppose we were to optimize over all space instead of $\mathbb{R}^n_\downarrow$- because the Bregman divergence is identically separable, the optimization problem separates out

over different coordinates j as $\hat{z}_t^{(j)} = \arg\min_z D_\phi(z\|\Gamma_t^{(j)})$ such that $z < (>) s_{\tau_i}$ for $j < (>)\tau_i$. This is a unidimensional convex optimization problem the solution to which is given by equation (3.5) above.

Finally we note that $\hat{\mathbf{z}}_\mathbf{t}$, automatically lies in $\mathbb{R}^n_\downarrow$ since $\mathbf{\Gamma_t} \in \mathbb{R}^n_\downarrow$, and hence, is also the solution to the optimisation problem (3.4). □

Now, note that since we are performing iterative minimization, the cost function is non-increasing at every step. As the cost function is bounded below by 0, the algorithm converges to a stationary point. We now extend the framework to include histogram constraints and blockwise partitioning.

3.2.2 Histogram Constraints

In case there are multiple order statistics constraints (histogram), the solution can be obtained by repeated application of equation (3.5).

Suppose for the column $\mathbf{z}$ we have constraints as ${\tau_i}^{th}$ order statistic of $\mathbf{z} = s_{\tau_i}$ for $\tau_i \in \{\tau_1, \tau_2, \cdots \tau_h\} \subseteq \{1, 2, \cdots n\}$, the solution is given by the following-

1. For all $j < \tau_1$, $\hat{z}^{(j)} = \min(\Gamma_i^{(j)}, s_{\tau_1})$; similarly, for all $j > \tau_h$, $\hat{z}^{(j)} = \max(\Gamma_i^{(j)}, s_{\tau_h})$

2. For all $1 \leq k < h$, and $j \ : \tau_k \leq j \leq \tau_{k+1}$,

$$\hat{z}^{(j)} = \begin{cases} s_{\tau_k} & j = \tau_k \\ s_{\tau_{k+1}} & j = \tau_{k+1} \\ \min\left(s_{\tau_{k+1}}, \max(\Gamma_i^{(j)}, s_{\tau_k})\right) & \tau_k \leq j \leq \tau_{k+1} \end{cases}$$

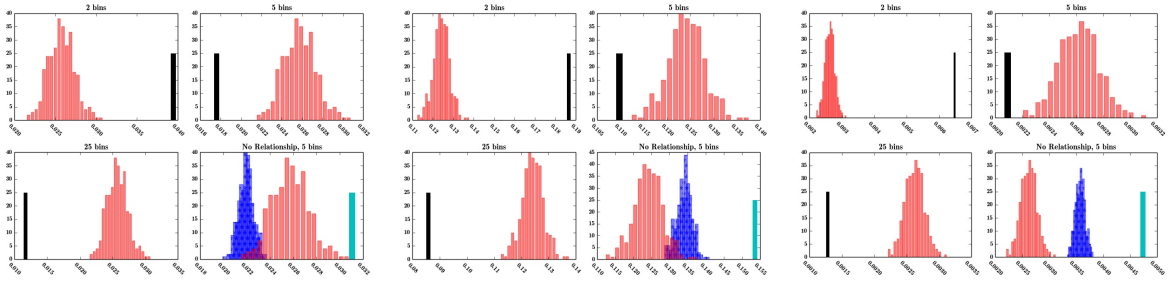
The proof for this follows in an identical manner to the proof for the non-partitioned case earlier. As above, the updated $\mathbf{z_t}$ can be obtained by re-permuting $\hat{\mathbf{z}}$

to preserve isotonicity with $(\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta})$. For a fully observed histogram, the update for $\mathbf{z}$ only involves a permutation at each step.

3.2.3 Blockwise Order Statistic Constraints

In the setup where the order statistics (or histograms) are computed over blockwise partitions of the sample, the permutation matrix is a blockwise permutation matrix and the isotonicity constraint is a blockwise isotonicity costraint.

Since the Bregman Divergence is identically separable, the update for $\mathbf{z}$ separates out into independent updates for every block which can be done in a manner identical to that given by Lemma 3.2.3. The update step for $\boldsymbol{\beta}$ remains unchanged.

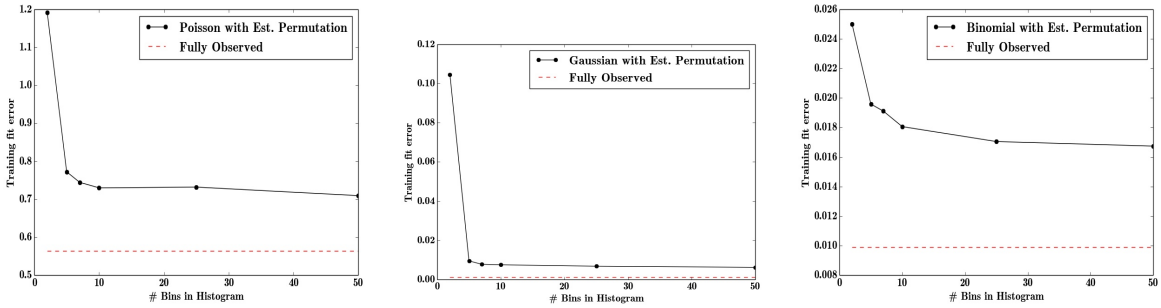


(a) Poisson Fit Error **(b)** Gaussian Fit Error **(c)** Binomial Fit Error

Figure 3.1: Permutation tests under Poisson, Gaussian and Binomial Estimation for 2, 5, 25 bins (top left, top right, bottom left) and "No Relationship" (bottom right)

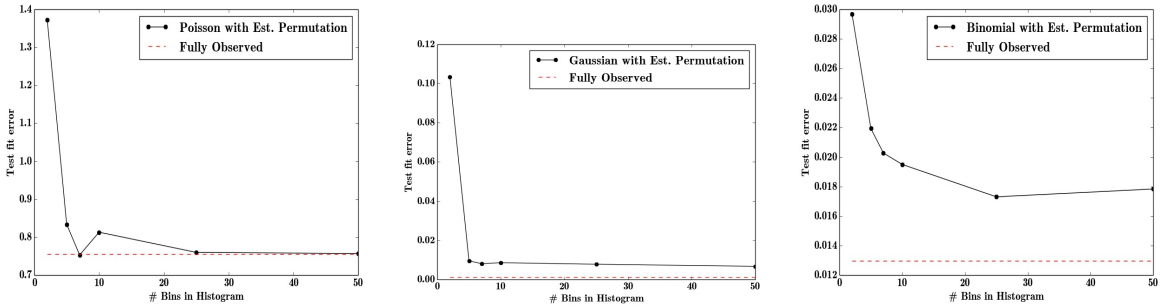
3.3 Experiments

We provide experimental results using both simulated data and real data. Error for each generalized linear model is defined as the corresponding Bregman divergence (square loss for Gaussian, generalized I-divergence for Poisson, etc. see



(a) Poisson Training Error **(b)** Gaussian Training Error **(c)** Binomial Training Error

Figure 3.2: Training Error under Poisson, Gaussian and Binomial Estimation



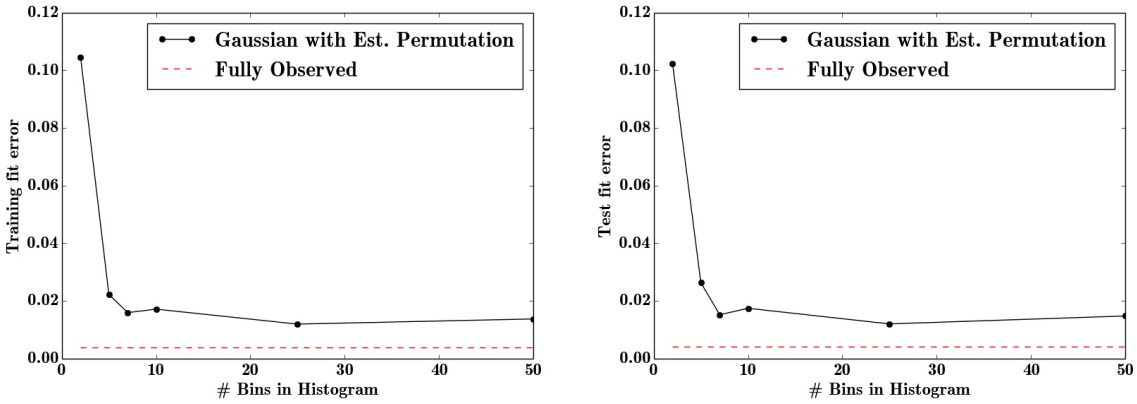
(a) Poisson Test Error **(b)** Gaussian Test Error **(c)** Binomial Test Error

Figure 3.3: Test Set Error under Poisson, Gaussian and Binomial Estimation

Banerjee et al. (2005)) between the true and recovered targets. The average errors for each model is shown separately.

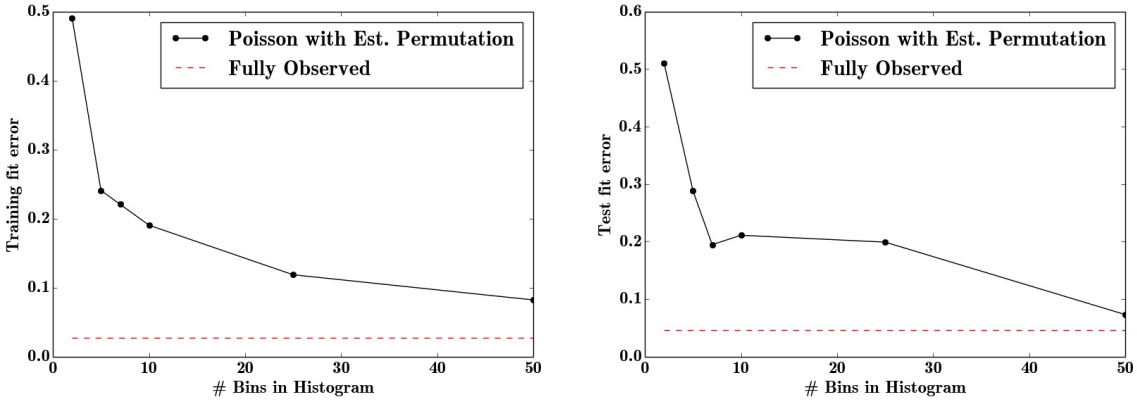
3.3.1 Simulated Data

We randomly generate different sets of real valued predictor variables and parameters, and use the corresponding exponential family to generate their respective response variables. We compute histograms for the response variables thus generated



(a) Training Set Error **(b)** Test Set Error

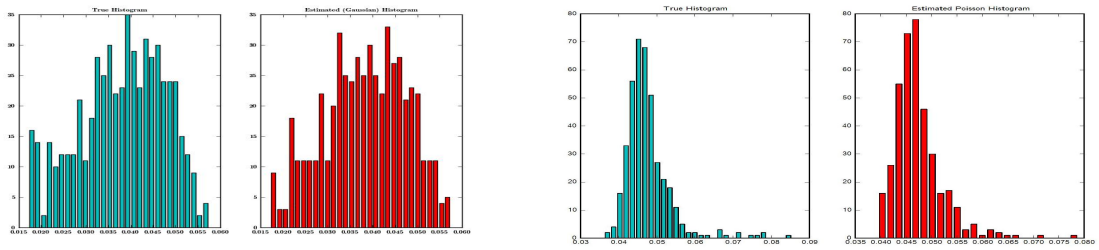
Figure 3.4: Performance on SynPUF dataset



(a) Training Set Error **(b)** Test Set Error

Figure 3.5: Performance on Texas Inpatient Discharge dataset

with varying number of bins and test our algorithm for each case. We perform the experiments for three different models - Gaussian, Poisson and Binomial.



(a) Recovered Histogram of DE-SynPUF Data **(b)** Recovered Histogram of Texas Discharged Data

Figure 3.6: Recovered Histograms of both datasets (true histograms on the left)

We perform a basic permutation test[2] to show how our algorithm performs with respect to the fit by a generalized linear model which knows the values of the target variables but permutes the target variables randomly for estimation. We perform the randomized permutations multiple times and plot a histogram of the fitting errors thus obtained and see how the results from our algorithm compares to the histogram (Figure 3.1). The black bar is the error obtained by our framework, the red histogram is the histogram of errors obtained by fitting after randomly permuting the targets. The blue histogram is the histogram of errors obtained by fitting a model where there is no relationship between the target variable and the covariate, the cyan bar is the result of our framework applied to this data with a histogram of 5 bins (histograms of other granularities perform similarly). Our test successfully rejects the null hypothesis of "no relationship when the the black bar is to the left of the red histogram. Figure 3.1 shows that as histogram becomes finer (i.e number of bins increase) error is lower i.e. black bar shifts towards left.

We plot the average fitting and predictive performance of our algorithm with

[2]Refer to Good (2005) for more details on permutation tests

increasing number of bins over five fold cross validation. We compare our results with the results obtained with the best possible GLM estimator which observes the full dataset (Figures 3.2 and 3.3)[3]. It can be seen in each case that as the histogram of targets becomes finer (i.e., more bins) the error decreases but with a diminishing returns property with respect to the coarseness of the histogram.

3.3.2 DE-SynPUF dataset

The CMS Beneficiary Summary DE-SynPUF dataset (DESynPUF, 2008) is a public use dataset created by the Centers for Medicare and Medicaid Services by applying different statistical disclosure limitation techniques to real beneficiary claims data in a way so as to very closely resemble real Medicare data. It is often used for testing different data mining or statistical inferential methods before getting access to real Medicare data. We use a subset of the DE-SynPUF dataset for a single state from the year 2008. With some trimming of datapoints (eg, we do not take into account deceased beneficiaries) we model outpatient institutional annual primary payer reimbursement amount ($PPPYMT\text{-}OP$) with a number of available predictor variables including age, race, sex, duration of coverage, presence/absence of a variety of chronic conditions, etc.

We perform a log transform and compute histograms of varying granularity on the target variables. We use a Gaussian model for our estimation and evaluate the average performance of our algorithm over five fold cross validation in fitting both

[3]training/test error in figures 3.2b and 3.3b for the Gaussian estimator for the fully observed case is ≈ 0

the training and test data sample points, comparing with the best possible Gaussian estimator which performs the estimation by observing the full dataset (Figures 3.4). As seen in the plot, the performance of our framework improves as the histogram of targets becomes finer in granularity and approaches the performance of the best Gaussian estimator. We also compare the histogram of target variables as recovered by our framework with the true histogram (Figure 3.6a).

3.3.3 Texas Inpatient Discharge dataset

We then test our algorithm on the Texas Inpatient Discharge dataset from the Texas Department of State Health Services (TxID, 2014), which is a healthcare dataset first used in Park & Ghosh (2014). As with the simulated data, we use histograms of varying granularity on the respective response variables and evaluate the average performance in fitting both the training and test data sample points over five fold cross validation. We use hospital billing records from the fourth quarter of 2006 in the Texas Inpatient Discharge dataset and regress it on the available individual level predictor variables including binary variables race and sex, categorical variables county and zipcode, and real valued variables like length of stay.

Following Park & Ghosh (2014), we perform a log transform on the hospital charges and length of stay before applying a Poisson regression model. We compare the performance of our algorithm over five-fold cross-validation with the best possible Poisson estimator which estimates in a fully observed scenario with an uncensored dataset (Figure 3.5). The plot shows that the performance of our framework improves with increasingly finer granularity of histograms and approaches the performance of

the best Poisson estimator. Finally, we compare the histogram recovered by our framework with the true histogram for the dataset (Figure 3.6b).

3.4 Conclusion

This chapter addresses the scenario where features are provided at the individual level, but the target variables are only available as histogram aggregates or order statistics. We proposed a simple algorithm to estimate the model parameters and individual level inferences via alternating imputation and standard generalized linear model fitting. We considered two limiting cases. In the first, the target variables are only known up to permutation. Our results suggest the effectiveness of the proposed approach when, in the original data, permutation testing accurately ascertains the veracity of the linear relationship. The framework was then extended to general histogram data with larger bins - with order statistics such as the median as a second limiting case. Experimental results on simulated data and real healthcare data show the effectiveness of the proposed approach which may have implications on using aggregation as a means of preserving privacy.

Chapter 4

Recovery of Sparse Parameter from Group-Wise Aggregated Data

4.1 Introduction

When sensitive or large-scale data is collected over natural geographic or demographic partitions of the population, individual samples are often aggregated over the corresponding sub-divisions and released as group-wise summaries (e.g., avergaes over race, gender, state, zip-code, etc.). For most applications in such domains, this aggregation paradigm is usually applied to both the feature or attribute variables, as well as target variables. Learning structured model parameters from such group-wise aggregated data is the focus of this chapter.

Group-wise aggregation is a common technique for sharing of privacy-sensitive healthcare data, where sensitive patient information is subject to various Statistical Disclosure Limitation (SDL) techniques (Armstrong et al., 1999) before public release. Similarly, large scale data collection programs like the General Social Survey

Portions of this chapter has been published as: A Bhowmik, J Ghosh, O Koyejo, "Sparse Parameter Recovery from Aggregated Data", International Conference on Machine Learning (ICML) 2016, New York, NY, USA

Co-authors have participated in discussions on research methods and contributed in the writing and review of the final manuscript.

(GSS) report data in aggregated form[1]. Data from IoTs and other distributed sensor networks are often collected in aggregated form to mitigate communication costs, and improve robustness to noise and malicious interference (Wagner, 2004; Zhao et al., 2003).

Building individual-level models given aggregates in the form of means, sample statistics, etc., constitutes a relatively unexplored semi-supervision framework. We note that even standard problems like regression and parameter recovery become very challenging in the context of aggregated data. Specifically, naïve application of standard techniques in the aggregated context is vulnerable to the ecological fallacy (Robinson, 2009; Goodman, 1953), wherein conclusions drawn from aggregated data can differ significantly from inferences at individual level, and are misleading to researchers/policy makers using the data.

As a first work on parameter recovery from aggregated data, we investigate the problem for regression in the case of linear models, where the mapping between input features and the output variable is defined by a vector parameter. We consider the scenario, very common in domains like healthcare, sociological studies, etc., where data is collected and aggregated within groups, e.g., patient records aggregated at county or hospital level, and empirical estimates of true group level moments for features and targets are the only available information.

While this problem is relatively easy to handle in the non-aggregated setup, parameter recovery becomes highly challenging when only aggregated data is avail-

[1]The General Social Survey, NORC, http://www3.norc.org/GSS+Website/

able and the resulting linear systems are under-determined. Well known works on compressed sensing (Donoho & Elad, 2003; Candes & Tao, 2005) have shown that recovery is still possible from such systems when the parameter is sparse (common in many applications of interest, e.g. in healthcare where interpretability is part of the desiderata), but existing analyses do not apply directly to the aggregated case.

Our work is motivated by the question: "Is it possible to infer the individual-level parameter of a linear model given aggregated data?" Surprisingly, we answer this question in the affirmative, and to our knowledge, ours is the first such work. We use techniques that exploit structural properties of the data aggregation procedure and show that under standard incoherence conditions on the matrix of true group level moments, the true parameter is recoverable with high probability.

The key contributions of this chapter are summarised below:

1. To our knowledge we are the first to investigate the problem of recovery of the sparse population parameter of a linear model when both target variables as well as features are aggregated as sample moments. We provide a theoretical analysis showing that under standard conditions, the parameter can be recovered exactly with high probability.

2. We extend the analysis to capture approximation effects such as sample estimates of the population moment, additive noise, and histogram aggregated targets, showing that the population parameter is recoverable in these scenarios.

3. In the bigger picture, our work extends existing results in the compressed sensing literature by providing guarantees for exact and approximate parameter recovery

for the case when the noise in the sensing matrix and measurement vector are linearly correlated, which may be of independent interest.

Experimental results on synthetic data are provided in support of these theoretical claims. We also show that the estimated parameter approaches the predictive accuracy of parameter estimation from non-aggregated or "individual-level" samples when applied to two real world healthcare applications - predictive modeling of reimbursement on CMS Medicare data, and estimation of healthcare charges using Texas State hospital billing records.

Note: Proofs for all results in this chapter are included in Appendix A.

4.2 Parameter Recovery from Exact Means

Let $\mathbf{x} \in \mathbb{R}^d$ represent features and $y \in \mathbb{R}$ represent the target variables, drawn independently from a joint distribution $(\mathbf{x}, \mathbf{y}) \sim \mathcal{P}$. We assume a linear model where each feature is related to the target y via some parameter $\boldsymbol{\beta}_0 \in \mathbb{R}^d$ with noise ϵ as

$$y = \mathbf{x}^\top \boldsymbol{\beta}_0 + \epsilon \tag{4.1}$$

where ϵ represents observation noise assumed zero mean $E[\epsilon] = 0$ without loss of generality. In the standard regression setting, data is observed at the individual level in the form of n pairs of targets and their corresponding features as $\mathbb{D}_{(x,y)} = \{(\mathbf{x}_i, y_i) : i = 1, 2, \ldots n\}$, so $\boldsymbol{\beta}_0$ may be estimated using standard techniques. Instead, we assume that the inputs $\mathbb{D}_x = \{\mathbf{x}_i : i = 1, 2, \cdots n\}$ and the targets $\mathbb{D}_y = \{y_l : l = 1, 2, \cdots n\}$ are subject to an aggregation process (not controlled by the learner)

that produces summaries. In particular, we focus on an aggregation procedure that produces means or first order moments of the data. A discussion on higher order moments is presented in Appendix A.

We consider the case when this aggregation procedure is applied separately to k subgroups of the population. This is common in many domains, e.g., in healthcare, such groups may refer to patient data aggregated by ward, or by hospitals, or based on administrative units like HRR's or HSA's. Similarly, the natural grouping could be demographic information for GSS data and topological clustering for sensor networks.

We assume that the grouping is fixed, and data associated with each group $j \in \{1, 2, \cdots k\}$ is drawn independently from a possibly group-dependent distribution $(\mathbf{x}, y)_j \sim \mathcal{P}_j$ with their own corresponding group-dependent means for covariates/features $\{\boldsymbol{\mu}_j = E_{\mathcal{P}_j}[\mathbf{x}],\ j = 1, \cdots, k\}$ and targets $\{\nu_j = E_{\mathcal{P}_j}[y],\ j = 1, \cdots, k\}$.

We also assume that the model parameter of interest $\boldsymbol{\beta}_0$ is shared by the entire population. By the distributive property of inner products and linearity of the expectation operator, any $\boldsymbol{\beta}_0$ consistent with the data satisfies the set of equations $\boldsymbol{\mu}_j^\top \boldsymbol{\beta}_0 = \nu_j \ \forall\ j = 1, 2, \cdots, k$. Let $\mathbf{M} = [\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \cdots \boldsymbol{\mu}_k]^\top \in \mathbb{R}^{k \times d}$ be the matrix of feature means, and $\mathbf{y} = [\nu_1, \nu_2, \cdots \nu_k]^\top \in \mathbb{R}^k$ is the vector of target means, it follows from eq. (4.1) that $\boldsymbol{\beta}_0$ satisfies

$$\mathbf{M}\boldsymbol{\beta}_0 = \mathbf{y}. \tag{4.2}$$

Clearly, if $k \geq d$ and the rank of $\mathbf{M}$ is greater than d, then (4.2) is sufficient to characterize $\boldsymbol{\beta}_0$. The more interesting case, and a more practical scenario, is when $k \ll d$, that is, the dimensionality of the problem is much larger than the number

of subgroups. We defer to compressed sensing approaches to estimate $\boldsymbol{\beta}_0$ from such systems.

4.2.1 Estimation from True Means using Compressed Sensing

The case of under-determined linear systems appears in a large number of practical scenarios and is a fairly well studied problem in compressed sensing. We adopt ideas from prior art in this field and demonstrate how they can be applied to our problem. An extended discussion on compressed sensing is available in Chapter 2, in this section we only summarise the principal results that we require for our analysis.

Definition 4.2.1. For a $k \times d$ matrix $\mathbf{M}$ and a set $T \subseteq \{1, 2, \cdots, d\}$, suppose $\mathbf{M}_T$ is the $k \times |T|$ matrix consisting of the columns of $\mathbf{M}$ corresponding to T. Then, the *s-restricted isometry constant* δ_s of the matrix $\mathbf{M}$ is defined as the smallest quantity δ_s such that the matrix $\mathbf{M}_T$ obeys

$$(1 - \delta_s)\|c\|_2^2 \leq \|\mathbf{M}_T c\|_2^2 \leq (1 + \delta_s)\|c\|_2^2$$

for every subset $T \subset \{1, 2, \cdots, d\}$ of size $|T| < s$ and all real $c \in \mathbb{R}^{|T|}$

Restricted isometry is a common and standard assumption in the sparse parameter recovery literature. Intuitively, this property means that when $\mathbf{M}$ satisfies Definition 4.2.1 with a small δ_s, every sub-matrix of small enough size constructed out of the columns of the matrix behaves approximately like an orthonormal system. In fact, a number of random matrices satisfy this property including the Gaussian ensemble and the Bernoulli ensemble (Donoho, 2006; Candès et al., 2006).

For the rest of the chapter we assume that the matrix of true means $\mathbf{M}$ satisfies the restricted isometry property. This is quite general as it is a direct corollary for many kinds of common and standard assumptions on the true mean matrix, for example the assumption that the true mean matrix is generated from a Gaussian distribution. Evidence from health care literature (Armstrong et al., 1999; Robinson, 2009) suggests that indeed, there is a significant geographical variation in demographics and health outcomes (due to variations in demographic make-up, average economic status, prevalent industries, etc.) which is often used as a predictive feature for healthcare models (Park & Ghosh, 2014; Bhowmik et al., 2015). All of this, together with our experiments on real datasets, suggest that there is sufficient inhomogeneity in mean healthcare attributes across groups to justify the matrix incoherence assumption for $\mathbf{M}$.

Suppose we had access to the true mean matrices $(\mathbf{M}, \mathbf{y})$. First, we consider the case when observations are noise-free, i.e. $\epsilon = 0$. Suppose $\boldsymbol{\beta}_0$ is known to be κ_0-sparse and $\mathbf{M}$ satisfies the restricted isometry hypothesis, then the following result applies:

Theorem 4.2.1 (Exact Recovery (Foucart, 2010))**.** *Let* $\Theta_0 = \frac{3}{4+\sqrt{6}} \approx 0.465$. *If there exists an* s_0 *such that* $\delta_{2s_0} < \Theta_0$ *for* $\mathbf{M}$*, then as long as* $\kappa_0 < s_0$*, the constraint* $\mathbf{M}\boldsymbol{\beta}_0 = \mathbf{y}$ *is sufficient to uniquely recover any* κ_0*-sparse* $\boldsymbol{\beta}_0$ *exactly as the solution of the following optimization problem:*

$$\min_{\boldsymbol{\beta}} \|\boldsymbol{\beta}\|_1 \quad s.t.\ \mathbf{M}\boldsymbol{\beta} = \mathbf{y}. \tag{4.3}$$

A similar result for approximate recovery holds for the case when the observations are corrupted with noise $\boldsymbol{\epsilon}$, i.e., instead of $\mathbf{y} = \mathbf{M}\boldsymbol{\beta}_0$, we are given $\mathbf{y}_\epsilon = \mathbf{M}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}$.

Theorem 4.2.2 (Approximate Recovery Candes (2008))**.** *Let* $\Theta_1 = \sqrt{2} - 1 \approx 0.414$. *If there exists an* s_0 *for* $\mathbf{M}$ *such that* $\delta_{2s_0} < \Theta_1$, *then as long as* $\kappa_0 < s_0$ *and the noise* $\boldsymbol{\epsilon}$ *in observations* $\mathbf{y}_\epsilon = \mathbf{M}\boldsymbol{\beta}_0 + \boldsymbol{\epsilon}$ *is bounded as* $\|\boldsymbol{\epsilon}\|_2 < \xi$, *any* κ_0*-sparse* $\boldsymbol{\beta}_0$ *can be recovered within an* ℓ_2 *distance of* $C_{s_0}\xi$ *from the true parameter* $\boldsymbol{\beta}_0$ *using the noisy measurements* $(\mathbf{M}, \mathbf{y}_\epsilon)$. *That is, the solution* $\widehat{\boldsymbol{\beta}}$ *to the following optimization problem:*

$$\min_{\boldsymbol{\beta}_0} \|\boldsymbol{\beta}\|_1 \ s.t. \ \|\mathbf{M}\boldsymbol{\beta} - \mathbf{y}_\epsilon\|_2 < \xi \tag{4.4}$$

satisfies $\|\widehat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\|_2 < C_{s_0}\xi$ *where the constant* C_{s_0} *depends only on* δ_{2s_0} *and is well-behaved (for example when* $\delta_{2s_0} = 0.2$, *the constant is less than 8.5).*

4.2.2 Empirical Mean Estimates and Aggregation Error

Clearly, if the matrix of true means $\mathbf{M}$ satisfies the restricted isometry hypothesis, and $\boldsymbol{\beta}_0$ is sufficiently sparse, Theorems 4.2.1 and 4.2.2 apply. Therefore, given the true population means $\mathbf{M}$ and $\mathbf{y}$, the parameter $\boldsymbol{\beta}_0$ can be recovered exactly from noiseless data $\mathbf{y}$ by solving (4.3) and approximately from noisy observations by solving (4.4).

Unfortunately, in many practical scenarios we do not have access to the true $\mathbf{M}$ or $\mathbf{y}$, but only to group level empirical estimates computed from a finite number of samples. Assume n samples for each group to simplify the analysis. Denote the corresponding empirically estimated means for the j^{th} group by $\hat{\boldsymbol{\mu}}_{j,n}$ and $\hat{\nu}_{j,n}$ for each $j = 1, \cdots k$. The corresponding sample mean matrices are given by $\widehat{\mathbf{M}}_n =$

$[\hat{\boldsymbol{\mu}}_{1,n}, \cdots \hat{\boldsymbol{\mu}}_{k,n}]^\top$ and $\widehat{Y}_n = [\hat{\nu}_{1,n}, \cdots \hat{\nu}_{k,n}]^\top$. The corresponding sample mean matrices are given by $\widehat{\mathbf{M}}_n = [\hat{\boldsymbol{\mu}}_{1,n}, \hat{\boldsymbol{\mu}}_{2,n}, \cdots \hat{\boldsymbol{\mu}}_{k,n}]^\top$ and $\widehat{Y}_n = [\hat{\nu}_{1,n}, \hat{\nu}_{2,n}, \cdots \hat{\nu}_{k,n}]^\top$.

The empirical mean estimation procedure introduces aggregation errors $\mathbf{e}_n$ and $\mathbf{s}_n$ to the setup. That is instead of the true group means $(\mathbf{M}, \mathbf{y})$, the data available for estimating $\boldsymbol{\beta}_0$ are restricted to empirical estimates $(\widehat{\mathbf{M}}_n, \widehat{Y}_n)$ where $\widehat{\mathbf{M}}_n = \mathbf{M} + \mathbf{e}_n$ and $\widehat{Y}_n = \mathbf{y} + \mathbf{s}_n$, and the results from section 4.2.1 no longer apply directly. For the rest of the chapter, we investigate parameter recovery for this scenario.

4.3 Parameter Recovery from Approximate Means

As mentioned earlier, the aggregation procedure for the estimation of true means introduces additive error terms $\mathbf{e}_n$ and $\mathbf{s}_n$ to the matrices $\mathbf{M}$ and $\mathbf{y}$. Note that for the models we study in this work, these two noise terms are not independent but are linearly correlated. Existing compressed sensing literature is restricted to the analysis of models where the additive error terms $\mathbf{e}_n$ and $\mathbf{s}_n$ are independent. Furthermore, any such existing analysis that deals with additive error terms are severely limited in the sense that they can only provide guarantees for approximate recovery rather than exact recovery (e.g. see Zhao & Yu (2006); Rosenbaum et al. (2013); Rudelson & Zhou (2015)).

Remarkably, as we show in the subsequent sections the true parameter is still exactly recoverable with high probability, even in the presence of linearly correlated aggregation error. This is because the aggregation procedure applied to linear models generates additional structure, which can then be exploited by the estimation procedure to get exact parameter recovery even from empirical estimates of the data

means from a finite number of samples.

We first analyse the case where the aggregation procedure has been applied to noise-free samples and then extend the analysis to the noisy case, and to the special case of data collected as histogram aggregates.

Throughout this chapter we shall make the standard assumption (Georgiou & Kyriakakis, 2006; Hsu et al., 2012) that the marginal distribution of each coordinate of the covariates is sub-Gaussian with parameter σ^2. Thus, for each covariate $x_j^{(i)} \in \mathbf{x}_j = [x_j^{(1)}, x_j^{(2)} \cdots x_j^{(d)}]$ and each group $j \in \{1, 2, \cdots k\}$, and for every $t \in \mathbb{R}$, the logarithm of the moment generating function is quadratically bounded

$$\ln E[e^{t(x_j^{(i)} - \mu_j^{(i)})}] < \frac{t^2\sigma^2}{2}.$$

Similarly, we assume that the marginal distribution for each noise term is zero-mean and sub-Gaussian with parameter ρ. Note that the assumptions on the covariates and the noise terms are only on the marginal distributions. In particular, we do not require either independence or identical distribution across groups or even across individual coordinates. As discussed in section 4.5.1, the analysis for alternative distributional assumptions follows along very similar lines by using other standard concentration inequalities. Proofs for all subsequent results are presented in Appendix A.

4.3.1 Noise-Free Observations

First we consider empirical means computed from noiseless observations. As mentioned earlier, the true parameter $\boldsymbol{\beta}_0$ can still be recovered exactly from empir-

ical estimates of group means $(\widehat{\mathbf{M}}_n, \widehat{Y}_n)$ despite the presence of linearly correlated aggregation error $(\mathbf{e}_n, \mathbf{s}_n)$.

Key observation: For a linear model, the relationship satisfied by the true group means $E[y] = E[\mathbf{x}]^\top \boldsymbol{\beta}_0$ is also exactly satisfied by the empirically estimated means $\frac{\sum y}{n} = \left(\frac{\sum \mathbf{x}}{n}\right)^\top \boldsymbol{\beta}_0$. Therefore, for aggregated noise-free observations, the equation

$$\widehat{\mathbf{M}}_n \boldsymbol{\beta}_0 = \widehat{Y}_n \tag{4.5}$$

still holds exactly. As long as the empirical moment matrix $\widehat{\mathbf{M}}_n$ satisfies the restricted isometry constraints, we may still guarantee exact recovery by solving the optimization problem:

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \|\boldsymbol{\beta}\|_1 \\ \text{s.t.} \quad & \widehat{\mathbf{M}}_n \boldsymbol{\beta} = \widehat{Y}_n. \end{aligned} \tag{4.6}$$

Our first main result is to show that this is indeed the case, and the true parameter $\boldsymbol{\beta}_0$ can be recovered with high probability if the number of samples n used to compute empirical moment estimates in each subgroup is sufficiently large.

Theorem 4.3.1 (Main result 1)**.** *Let* $\Theta_0 = \frac{3}{4+\sqrt{6}} \approx 0.465$*. Suppose there exists an* s_0 *such that the isometry constant* δ_{2s_0} *for the true mean matrix* $\mathbf{M}$ *satisfies* $\delta_{2s_0} < \Theta_0$*. Also suppose that the marginal distribution of the coordinates of each feature is sub-Gaussian with parameter* σ^2*. Then, given* $(\widehat{\mathbf{M}}_n, \widehat{Y}_n)$ *any* κ_0*-sparse* $\boldsymbol{\beta}_0$ *with* $\kappa_0 < s_0$ *can be recovered exactly with probability at least* $1 - e^{-C_0 n}$ *by solving (4.6). Here, the constant* C_0 *in the expression is such that* $C_0 \sim O\left(\frac{(\Theta_0 - \delta_{2s_0})^2}{kd\sigma^2(1+\delta_{2s_0})}\right)$*.*

We can unpack the result with respect to the constant C_0 which depends on the isometry parameter δ_{2s_0}, the size of the mean matrix (k, d) and the sub-Gaussian

parameter of the feature terms σ. The robustness of the isometry property of $\widehat{\mathbf{M}}_n$ depends on the strength of the isometry property in the true moment matrix $\mathbf{M}$. Fewer samples are required for estimating $\widehat{\mathbf{M}}_n$ if $\mathbf{M}$ satisfies the isometry hypothesis more robustly (that is, δ_{2s_0} small) and consequently, a larger value of $\frac{(\Theta_0-\delta_{2s_0})^2}{1+\delta_{2s_0}}$. Similarly, if the feature distributions have a thinner tail i.e. a smaller value of the sub-Gaussian parameter σ^2, empirically estimated means are more accurate with fewer samples.

4.3.2 Observations with Noise

We now consider the case when the observations are noisy and the equation (4.5) no longer holds exactly. In particular, we assume that the data used to compute the sample moments is observed with zero mean additive noise as $y^{\epsilon}_{i,j} = \mathbf{x}_{i,j}^{\top}\boldsymbol{\beta}_0 + \epsilon_{i,j}$ for each datapoint $i \in \{1, \cdots, n\}$ in population subgroup $j \in \{1, \cdots, k\}$. This leads to an error in the empirical target means over and above the aggregation error.

Let $\widehat{Y}_{n,\epsilon} = \widehat{Y}_n + \boldsymbol{\epsilon}_n$ where $\widehat{Y}_{n,\epsilon}$ (henceforth denoted $\widehat{Y}_{\epsilon}$) is the empirical target mean estimated from noisy samples and $\boldsymbol{\epsilon}_n$ is the cumulative estimation error due to noise in n samples. With the feature sample mean $\widehat{\mathbf{M}}_n$, eq. (4.5) becomes

$$\widehat{\mathbf{M}}_n\boldsymbol{\beta} = \widehat{Y}_n = \widehat{Y}_{\epsilon} - \boldsymbol{\epsilon}_n. \tag{4.7}$$

Similar to the results of Theorem 4.2.2, it can be expected that if the sample mean matrix $\widehat{\mathbf{M}}_n$ satisfies the isometry hypothesis for noisy measurements, and if the error term $\boldsymbol{\epsilon}_n$ is bounded as $\|\boldsymbol{\epsilon}_n\|_2 < \xi$ for some $\xi > 0$, then $\boldsymbol{\beta}_0$ can be recovered

to within an ℓ_2 distance of $O(\xi)$ by solving the following optimization problem

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \|\boldsymbol{\beta}\|_1 \\ \text{s.t.} \quad & \|\widehat{\mathbf{M}}_n\boldsymbol{\beta} - \widehat{Y}_\epsilon\|_2 < \xi. \end{aligned} \tag{4.8}$$

In fact, in our case we can show that the aggregation procedure smooths out the destabilising effects of noise in observations to allow arbitrarily accurate parameter recovery within any small degree ξ of ℓ_2 estimation error.

Theorem 4.3.2 (Main Result 2)**.** *Let $\Theta_1 = \sqrt{2} - 1 \approx 0.414$. Suppose there exists an s_0 such that the isometry constant δ_{2s_0} for the true mean matrix $\mathbf{M}$ satisfies $\delta_{2s_0} < \Theta_1$. Also suppose that the marginal distribution of the coordinates of each feature is sub-Gaussian with parameter σ^2, and noise in each observation is zero-mean and sub-Gaussian with parameter ρ^2. Let $\xi > 0$ be any small positive real value. Then, any κ_0-sparse $\boldsymbol{\beta}_0$ with $\kappa_0 < s_0$ can be recovered within an ℓ_2 distance of $O(\xi)$ with probability at least $1 - e^{-C_1 n} - e^{-C_2 n}$ by solving (4.8). Here, the constant C_1 is such that $C_1 \sim O\left(\frac{(\Theta_1 - \delta_{2s_0})^2}{kd\sigma^2(1+\delta_{2s_0})}\right)$ and the constant C_2 is such that $C_2 \sim O\left(\frac{\xi^2}{\rho^2 k}\right)$.*

The constant term in $O(\xi)$ is the same as that in Theorem 4.2.2 and it depends only on δ_{2s_0} and is well-behaved for small values of δ_{2s_0}. Note the similarity of the constant C_1 in the noisy case and the constant C_0 in the exact case. As for exact recovery, the probability of recovery depends on the tail properties of the feature distribution as well as the robustness of the isometry property for the true mean matrix $\mathbf{M}$. The constant $\frac{\xi^2}{\rho^2 k}$ in the additional term accounts for observational noise. As expected, more samples are required if the noise has heavy tails ρ^2 or if the degree of approximation ξ is small. In addition, the constant for $O(\xi)$ in the approximation factor may depend only δ_{2s_0} in a manner similar to Theorem 4.2.2.

4.3.3 Extension to Histogram Aggregation

For the preceding analysis, we have assumed that errors in the target moments is a result of the empirical aggregation or observational noise. It is worth noting that this analysis can be extended to cover any additional source of error which can be bounded deterministically or with high probability. An example of this is when the targets are available as histogram aggregates with bin size Δ and the mean is estimated from the histogram. Suppose h_Δ is the error in estimation of target mean from the histogram such that the estimated sample mean $\widehat{Y}_\Delta$ is related to the true sample mean for the targets as $\widehat{Y}_\Delta = \widehat{Y}_n + h_\Delta$.

Then, we can use the exact same procedure as for noisy observations to bound the ℓ_2 error in estimation of $\boldsymbol{\beta}_0$ to $O(\xi_\Delta)$ by solving the optimisation problem

$$\begin{aligned} \min_{\boldsymbol{\beta}} \quad & \|\boldsymbol{\beta}\|_1 \\ \text{s.t.} \quad & \|\widehat{\mathbf{M}}_n\boldsymbol{\beta} - \widehat{Y}_\Delta\|_2 < \xi_\Delta \end{aligned} \tag{4.9}$$

for some positive $\xi_\Delta > 0$.

The value of ξ_Δ and theoretical guarantees arising therefrom will depend on the manner in which the target mean in estimated from the histogram. Here, we analyse one such standard moment estimation approach.

Consider a single population subgroup. Suppose the range of the targets is bounded by some R, that is, $y_{\max} - y_{\min} < R$. We have a set of bins $\mathcal{B} = \{B_\tau = (b_\tau, b_{\tau+1}) : \tau = 1, 2, \cdots, \lfloor \frac{R}{\Delta} \rfloor\}$ such that $b_{\tau+1} - b_\tau = \Delta$ for each bin. We also have for each bin an integer n_τ which is the number of targets for that subgroup that fall in that particular bin. Suppose $\bar{b}_\tau = \frac{(b_\tau + b_{\tau+1})}{2}$ is the mid point of each bin. Then, the

target mean for that group is estimated as

$$\hat{\nu}_\Delta = \frac{\sum_\tau n_\tau \bar{b}_\tau}{\sum_\tau n_\tau} = \frac{\sum_\tau n_\tau \bar{b}_\tau}{n}.$$

For this mean imputation procedure, we get a very similar result to Theorem 3.2 for aggregated data that bounds the probability of recovery in terms of the isometry constants of the true mean matrix and the granularity of the histogram.

Theorem 4.3.3 (Main Result 3)**.** *Let* $\Theta_1 = \sqrt{2} - 1 \approx 0.414$. *Suppose there exists an* s_0 *such that the isometry constant* δ_{2s_0} *for the true mean matrix* $\mathbf{M}$ *satisfies* $\delta_{2s_0} < \Theta_1$.*Also suppose that each covariate has a sub-Gaussian distribution with parameter* σ^2. *Let the targets for each group be available as histogram aggregates with bin size bounded below by* Δ. *Then, any* κ_0*-sparse* $\boldsymbol{\beta}_0$ *with* $\kappa_0 < s_0$ *can be recovered within an* ℓ_2 *distance of* $O(\sqrt{k}\Delta)$ *with probability at least* $1 - e^{-C_1 n}$ *by solving (4.9) with* $\xi_\Delta = \sqrt{k}\frac{\Delta}{2}$. *Here, the constant* C_1 *is such that* $C_1 \sim O\left(\frac{(\Theta_1 - \delta_{2s_0})^2}{kd\sigma^2(1+\delta_{2s_0})}\right)$.

Note that the constants on $O(\sqrt{k}\Delta)$ are the same as in the case of noisy observations. Also, in the case of exact estimation, bin size $\Delta \to 0$, therefore $\boldsymbol{\beta}_0$ can be recovered exactly. Furthermore, the bin size does not have any effect on the sample complexity of recovery probability, only on the accuracy of estimation.

In particular, the recovery error is small for a histogram of fine enough granularity. In most cases of binned data, the bin size used for reporting the histogram decreases as a function of n. In fact for many real world scenarios (see Scott (1979a)) the bin size decreases at least as fast as $\Delta = O(\frac{1}{n^c})$ for some $0 < c < 1$. In any case, the worst case error in parameter estimation is limited solely by the bin size, and

tighter bounds can be obtained by making reasonable assumptions on the target distribution. Note that if instead of supplying a coarse histogram the data is released in full (without specifying the relationship between $\mathbf{x}$ and $\mathbf{y}$ in each group), the effective bin size is 0 and the parameter can be estimated exactly by Theorem 4.3.3.

Related Work

While there is a rich literature on sparse parameter recovery and predictive modeling in general, the aggregated data case is much more limited. To our knowledge, ours is the first analysis of sparse parameter recovery for aggregated data of *any* kind. Our main result while seemingly obvious after the fact, has not been shown in more than 60 years of ecological data analysis dating at least to Goodman (Goodman, 1953), with parallel work in the compressed sensing literature, and renewed interest in machine learning (Park & Ghosh, 2014; Bhowmik et al., 2015). Furthermore, as noted in section 4.5.1, our analysis can be easily extended to study sparse parameter recovery from aggregated data in various contexts using a wide variety of estimation techniques beyond what we present in this chapter.

The techniques used in our work follows a long line of research on compressed sensing as discussed in Section 4.2.1, where related analyses fall mainly under three categories:

1. error in the design matrix $\widehat{\mathbf{M}} = \mathbf{M} + \mathbf{e}$, without any error or noise in observation vector $\mathbf{y}$

2. noise in observations $\widehat{Y} = \mathbf{y} + \mathbf{s}$, with a fixed design matrix $\mathbf{M}$ without error

3. design matrix error $\mathbf{e}$ and observation noise $\mathbf{s}$, where $\mathbf{e}$ and $\mathbf{s}$ are independent

Prior work, eg. (Herman & Strohmer, 2010; Zhao & Yu, 2006; Rudelson & Zhou, 2015), deals only with case 1, or with cases 2 and 3 in a way to only provide *approximate* parameter recovery guarantees. We focus our investigation on the aggregated data case 4: where $\mathbf{E}$ and $\mathbf{s}$ are linearly correlated. Even ignoring the linear correlation in the noise model, the best existing analyses are still limited to using a naive error bounding technique to analyse the stability of the LASSO resulting in weak guarantees for only approximate parameter recovery.

In contrast, we propose non-trivial modifications to the analysis, and are able to exploit the additional structure generated by the data aggregation procedure to recover the sparse parameter *exactly* even with aggregation error, as in Theorem 4.3.1, and upto arbitrarily accurate degree of estimation from noisy data as we see in Theorems 4.3.2 and 4.3.3.

4.4 Experiments

We corroborate our theoretical results with experiments on synthetic data to show that probability of exact parameter recovery follows a pattern just as predicted by our main results. We also demonstrate the efficacy of our technique in two real world applications by applying it to predictive modeling of outpatient reimbursement claims in CMS Medicare data (DE-SynPUF), and to modeling healthcare costs using Texas Inpatient Discharge dataset (TxID) from the Texas Department of State Health Services.

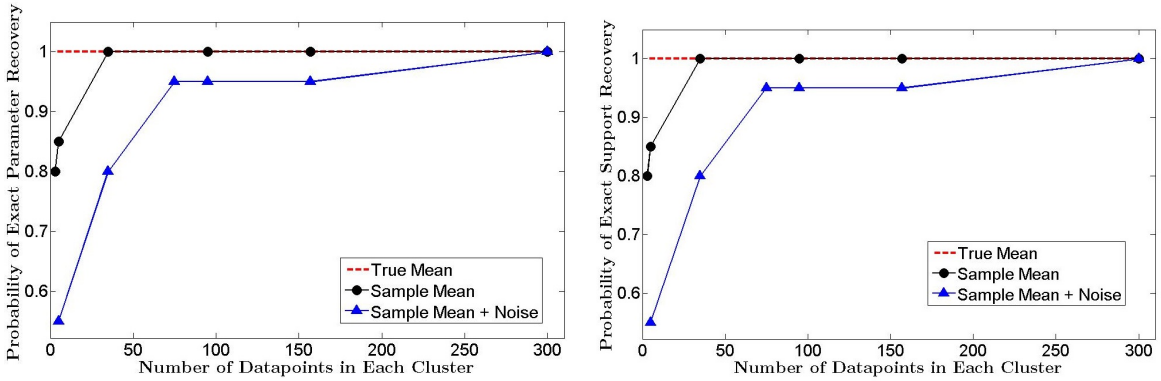
4.4.1 Synthetic Data

We first generate the true covariate mean matrix $\mathbf{M}$ using a Gaussian and a Bernoulli ensemble, and compute the respective true target means using a sparse $\boldsymbol{\beta}_0$. We then generate random covariates centred around the true mean matrix and compute the corresponding empirical mean matrix $\widehat{\mathbf{M}}_n$ from the covariates. The targets are then generated using the parameter $\boldsymbol{\beta}_0$. We consider two cases separately-noiseless targets $\mathbf{y}$ and targets $\mathbf{y}_\epsilon$ to which noise has been added. The corresponding empirical target means $\widehat{Y}_n$ and $\widehat{Y}_\epsilon$ are computed for both sets of targets and used together with the sample covariate means $\widehat{\mathbf{M}}_n$ to estimate $\boldsymbol{\beta}_0$.

This entire procedure is repeated multiple times and the proportion of instances in which the true parameter $\boldsymbol{\beta}_0$ is recovered exactly, both in magnitude and support, is plotted against the number of datapoints used to compute the empirical sample means. Figures 4.1 and 4.2 show the results for Gaussian and Bernoulli ensembles respectively. As can be seen in the figures, the probability of recovering the exact parameter increases as the number of data points used to compute the empirical sample means increases, in a manner exactly as predicted by our theoretical results.

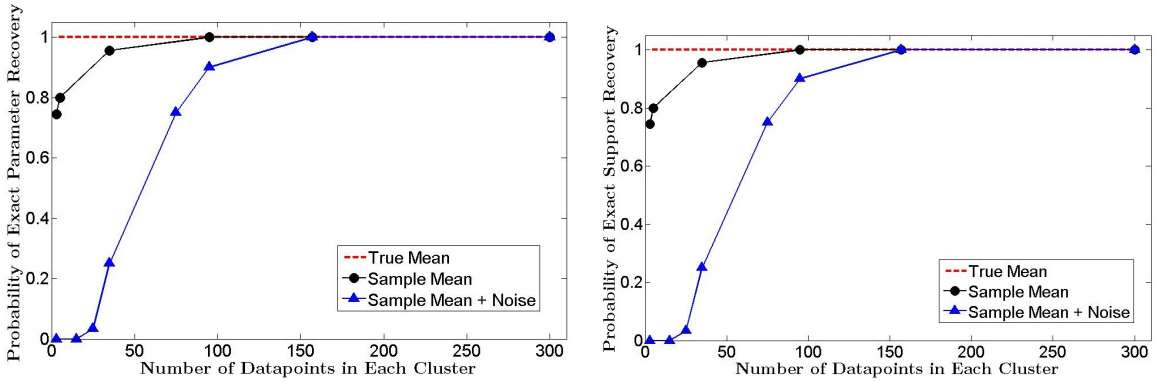
4.4.2 Real datasets - DE-SynPUF and TxID

We now apply our methods to two real datasets. Since ours is the first work on sparse recovery from aggregated data, we do not know of any competing algorithmic baselines. We evaluate our methods by comparing the parameter estimated from aggregated data to the performance upper bound of the "true" parameter that is



(a) Probability of recovering exact parameter **(b)** Probability of recovering exact signed support

Figure 4.1: Performance on Gaussian model with increasing number of datapoints in each group



(a) Probability of recovering exact parameter **(b)** Probability of recovering exact signed support

Figure 4.2: Performance on Bernoulli model with increasing number of datapoints in each group

estimated from the full non-aggregated dataset.

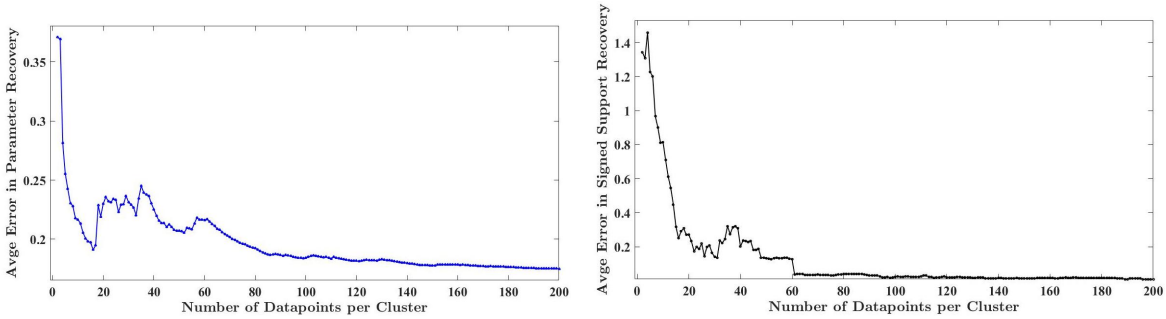
Our first dataset is the CMS Beneficiary Summary (DE-SynPUF) dataset DESynPUF (2008) which is a public use dataset created by the Centers for Medicare and Medicaid Services and is often used for testing different data mining or statistical inferential methods before getting access to full Medicare data. We use a subset of the DE-SynPUF dataset for Louisiana state from the year 2008 and model

outpatient institutional annual primary payer reimbursement ($PPPYMT\text{-}OP$) with all the available predictor variables that include age, race, sex, duration of coverage, presence/absence of a variety of chronic conditions, etc.

Our second dataset is the Texas Inpatient Discharge dataset (TxID) from the Texas Department of State Health Services ((TxID, 2014), see also (Park & Ghosh, 2014)). We model healthcare charges using hospital billing records from the fourth quarter of 2006 in the TxID dataset, and use all the available individual level predictor variables, which include demographic information like race, and real valued variables like length of hospital stay for each datapoint.

In both these datasets, we first use a LASSO estimator (with parameter chosen via cross-validation) on the full dataset to obtain a sparse regression parameter $\boldsymbol{\beta}_{full}$. We use a k-means algorithm to cluster the datapoints into groups and compute the sample means for each group with increasing number of datapoints. We then use only these empirical sample means to obtain an estimate $\boldsymbol{\beta}_{agg}$ for the parameter, and compare $\boldsymbol{\beta}_{agg}$ to the parameter $\boldsymbol{\beta}_{full}$ obtained from full non-aggregated dataset. Results averaged across multiple clusterings are shown in figures 4.3 and 4.4.

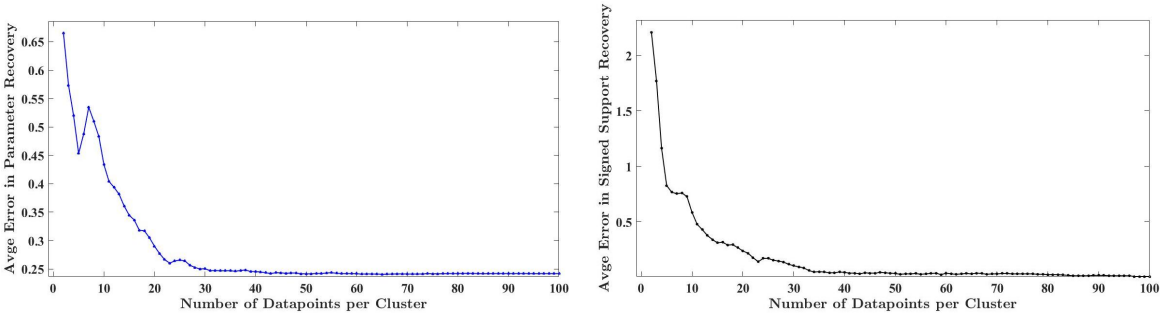
Figures 4.3a and 4.4a show the ℓ_2 norm of the distance between the parameter estimated from the full dataset $\boldsymbol{\beta}_{full}$ and the parameter estimated from the aggregated version $\boldsymbol{\beta}_{agg}$, for the DE-SynPUF dataset and TxID dataset respectively, plotted against the number of datapoints used to estimate the means. Figure 4.3b and 4.4b show the number of conflicts or discrepancies between the support (non-zero coordinates) of $\boldsymbol{\beta}_{agg}$ estimated from aggregated data and support of $\boldsymbol{\beta}_{full}$ estimated from the non-aggregated dataset, for the DE-SynPUF dataset and TxID dataset



(a) Avg. Error in Parameter Recovery $\|\boldsymbol{\beta}_{full} - \boldsymbol{\beta}_{agg}\|$

(b) Avg. Error in Support Recovery

Figure 4.3: Performance on DESynPUF dataset with increasing number of datapoints in each group



(a) Avg. Error in Parameter Recovery $\|\boldsymbol{\beta}_{full} - \boldsymbol{\beta}_{agg}\|$

(b) Avg. Error in Support Recovery

Figure 4.4: Performance on TxID dataset with increasing number of datapoints in each group

respectively. It can be seen that as the number of datapoints used to compute the sample means increases, the parameter recovered using aggregated data exactly identifies the support of the "true" parameter estimated from the full dataset, and also closely matches it in magnitude.

4.5 Discussion

4.5.1 Extensions

The techniques presented in this work can be applied to the parameter recovery problem in a much wider class of cases of interest by building on and extending existing results in the compressed sensing literature (see Candes et al. (2006); Candes & Tao (2007); Cai et al. (2010a, 2009), etc). In particular, we note that various alternative frameworks like non-sparse $\boldsymbol{\beta}_0$, alternative estimators to LASSO, beyond sub-gaussian assumptions on different marginals, etc. can be analysed in an identical manner, and our main results on parameter recovery would still continue to hold, albeit with slightly different sample complexity.

4.5.2 Higher Order Moments

The results in this chapter focused on estimation from first order moments. It may seem like including higher order moments might make estimation in this framework easier but it turns out that this is not the case in general. We include a discussion in Appendix A on the difficulties of using higher order moments for estimation. In particular, we prove a surprising and counter-intuitive negative result which shows that even with second order moments, in the general case the estimation cannot be guaranteed to be easier or more accurate than when we use only first order moments. Similar results may also hold for other higher order moments.

4.6 Conclusion

In this chapter we study the problem of parameter recovery for sparse linear models from data which has been aggregated in the form of empirical means computed from different subgroups of the population. We show that when the collection of true group moments is an incoherent matrix, the parameter can be recovered with high probability from the empirical moments alone provided the empirical moments are computed from a sufficiently large number of samples. We extend the framework to the case of moments computed from noisy or histogram aggregated data and show that the parameter can still be recovered within an arbitrarily small degree of error. We corroborate our theoretical results with experiments on synthetic data and also show results on two real world healthcare applications- predictive modeling of reimbursement claims from CMS Medicare data, and modeling healthcare charges using hospital billing records from the Texas Department of State Health Services.

Chapter 5

Frequency Domain Predictive Modelling with Spatio-Temporally Aggregated Data

5.1 Introduction

Analysis of spatio-temporally correlated data is an important and ever present problem in diverse and wide-ranging fields including econometrics (Davidson et al., 1978), climate science (Lozano et al., 2009; Liu et al., 2010), financial forecasting (Taylor, 2007) and Internet of Things (IoTs) (Da Xu et al., 2014; Li et al., 2013). Nearly all existing modelling techniques in literature assume access to datasets with individual level samples for each time and/or location index. However, in many real life cases (Burrell et al., 2004; Lozano et al., 2009; Davidson et al., 1978), for various reasons including measurement fidelity, robustness to random noise, cost of data collection, privacy preservation, scalability, etc., data is often collected and/or publicly reported as *aggregates* or *time averages*, collected over specific intervals and released periodically, e.g., data released by the Bureau of Labour Statistics

Portions of this chapter has been published as: A Bhowmik, J Ghosh, O Koyejo, "Frequency Domain Predictive Modelling with Aggregated Data", International Conference on Artificial Intelligence and Statistics (AISTATS) 2017, Fort Lauderdale, USA

Co-authors have participated in discussions on research methods and contributed in reviewing the final manuscript.

(US Department of Labour) and Bureau of Economic Analysis (US Department of Commerce), or by the General Social Survey (NORC) are often in this form.

The central question addressed in this chapter is whether one can provably learn individual level models given only aggregated spatio-temporal data– a challenging and relatively unexplored form of semi-supervision, which requires novel techniques and significant algorithmic innovation on the part of data analysts to perform modeling and inference. As a first work (to the best of our knowledge) on predictive modelling with spatio-temporally aggregated data, we tackle the problem in the context of predictive linear modelling where real valued targets are regressed on multivariate features via a vector parameter.

Even for this relatively simple setup, naive application of standard modelling techniques to aggregated data often fails due to ecological fallacy (Robinson, 2009; Freedman et al., 1991a; Goodman, 1953) wherein inferences drawn at the group level differ significantly from the ground truth at individual level. Learning is especially difficult if aggregation periods are not uniform or aligned across features and targets. For example, an econometric model may want to use as features metrics like GDP growth rate (reported quarterly), unemployment rate and inflation rate (reported monthly), interest rate and balance of trade (reported daily) and ratio of government debt to GDP (reported yearly) to predict, say, stock market indices and currency exchange rates (reported daily) (US Department of Commerce; US Department of Labour).

In such a scenario, it is extremely challenging even to formulate a cogent mathematical representation that captures the relationship among the available misaligned

aggregates. On the other hand, effective reconstruction of data at the individual-level is very difficult because aggregation fundamentally obfuscates local information.

5.1.1 Contributions

In this chapter, we demonstrate that by formulating the problem in the frequency domain, selected global properties of individual components of the model can be separately estimated with high fidelity even from aggregated data, which can then used for learning and inference without being affected by local-level information obfuscation caused by aggregation– all of this without any explicit data reconstruction. Our specific contributions are summarised below-

1. To our knowledge, we are the first to investigate the problem of predictive modelling from aggregated spatio-temporal data. We introduce a novel framework and new algorithmic mechanisms for learning from aggregated spatio-temporal data that leverages structural properties of frequency domain analysis techniques to perform predictive modelling with minimal data reconstruction.

2. We provide theoretical guarantees for our framework, and establish that under mild regularity conditions, the parameter vector learned from aggregated data suffers a generalisation error that is provably close to the optimal that can be obtained from any linear model in the non-aggregated setting, that is, with individual level samples

3. We extend our analysis to derive guarantees for our algorithm to capture real world approximation effects caused by aliasing and randomness in the data gen-

eration procedure, and show that our methods can still learn a parameter that closely matches the optimal generalisation error.

We empirically evaluate the efficacy of our methods on both synthetic data and three real datasets involving applications in ecological surveys, agricultural studies and climate science.

Note: The proofs for all results presented in this chapter are relegated to Appendix B

5.1.2 Related Work

There is a vast range of work on spatio-temporal data analysis (Lozano et al., 2009; Lambert et al., 2004; Ho et al., 2013) but very little existing literature applies to the aggregated case. The closest that come to our setup are interpolation techniques like Kriging (Stein, 2012; Oliver & Webster, 1990), which also typically assume that data is sampled at localised discrete positions on a grid, rather than as aggregates. Among frequency domain techniques, the closest line of work is spectral regression (Cai et al., 2007; Phillips et al., 1988; Corbae et al., 2002) which has been previously used in econometrics and financial modeling. However, existing work only deals with non-aggregated data in the discrete domain, and in particular, we have not come across an estimation framework nor analysis techniques, nor any guarantees for generalisation error as introduced in this chapter.

Note that while our work involves spatio-temporal data, the goal is nevertheless a general framework for predictive modelling rather than forecasting– in fact,

our methods can be used even outside spatio-temporal applications, e.g. in any domain wherein sampled measurements can be represented as tensors where a sense of ordering or structural chronology exists along each mode (for example, clinical measurements).

Existing literature as relates to estimation from aggregated data is discussed in chapter 2 and a detailed discussion on the same is, therefore, omitted from this chapter.

In chapter 2, we provided an in-depth discussion on frequency domain analysis and related topics. These results are examples of the well known duality properties of Fourier analysis, where global properties in the time domain are related to local properties in the frequency domain and vice versa. We shall use these properties extensively in our work.

Throughout this chapter, we assume that the power spectral density (and correspondingly, the autocovariance function) for every signal of interest exists finitely, and decays rapidly with lag for all processes involved. In particular, we assume that $\rho_{(\cdot)}(\cdot)$ is a Schwartz function (Terzioğlu, 1969), that is $\rho(\cdot)$ and all its derivatives decay at least as fast as any inverse polynomial. Therefore, most of the power for our signals will be concentrated around $\omega = 0$. An extended discussion on this is presented in section B.3 in Appendix B.

5.2 Problem Setup

In the interest of simplicity we delineate our setup for temporally aggregated data, where features $\mathbf{x}(t)$ and targets $y(t)$ are time series signals or processes. Discussion on higher dimensional aggregation frameworks are deferred to section 5.4.

Consider the task of predictive linear modelling, where real valued targets $y(t) \in \mathbb{R}$ are regressed on multivariate feature vectors $\mathbf{x}(t) \in \mathbb{R}^d$ via a parameter vector $\boldsymbol{\beta}^* \in \mathbb{R}^d$ in a linear model

$$y(t) = \mathbf{x}(t)^\top \boldsymbol{\beta}^* + \epsilon(t) \tag{5.1}$$

where $\epsilon(t)$ is a random noise process. For the rest of our chapter, we make the assumption that all our signals of interest $\mathbf{x}, y, \epsilon$ are centered and weakly stationary with finite variance.

Stationarity is a standard assumption in time series analysis and very common in many real life applications (see Granger & Newbold (2014); Dzhaparidze (2012); Feige & Pearce (1974)), and techniques like filtering out trend lines and differencing are often applied to the data to ensure stationarity before analysis (Hibon & Makridakis, 1997). Note that we do not assume any specific functional form for the generative processes (Gaussian, etc.) for the signals studied in this chapter.

Loss Function and Parameter Estimation

Standard statistical learning approaches estimate the optimal linear model given the data by minimising an appropriate loss function over the vector parameter

$\boldsymbol{\beta}$. Define the residue process at any particular $\boldsymbol{\beta}$ as

$$\varepsilon_\beta(t) = \mathbf{x}(t)^\top \boldsymbol{\beta} - y(t)$$

One potential option for a loss function might have been the total energy of the residue process $\int_{\mathbb{R}} |\varepsilon_\beta(t)|^2 dt$

However, the total energy in the noise process is often not finite (Koopmans, 1995; Tangirala, 2014), hence for weakly stationary processes, a better loss function to use is the variance of the noise process at time t, that is,

$$\mathcal{L}(\boldsymbol{\beta}) = E[|\varepsilon_\beta(t)|^2] = E[|\mathbf{x}(t)^\top \boldsymbol{\beta} - y(t)|^2]$$

By assumption our signals are weakly stationary, therefore the variance does not depend on t. Therefore, the "optimal" linear model parameter is given by

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \mathcal{L}(\boldsymbol{\beta}) = \arg\min_{\boldsymbol{\beta}} E[|\mathbf{x}(t)^\top \boldsymbol{\beta} - y(t)|^2] \tag{5.2}$$

Given access to the detailed, full-resolution dataset, the typical strategy for solving the estimation problem (5.2) is to replace the expectation by a sum over individual datapoints. This finite sum converges to the expectation given enough datapoints under certain conditions, for example, if the noise process is ergodic (Wiener, 1949). However, the story becomes more complicated if the data is available in aggregated form.

5.2.1 Data Aggregation in Time Series

Instead of the individual targets $y(t)$ at time t, we are given aggregates sampled with period T, which are of the form

$$\overline{y}[k] = \frac{1}{T} \int_{(k-1)T/2}^{kT/2} y(\tau) d\tau \tag{5.3}$$

for $k \in \mathbb{Z} = \{\cdots -1, 0, 1, \cdots\}$.

Features can also be aggregated in a more complicated manner with different periodicities, that is, each coordinate $\{\overline{x}_i(t) : \ i = 1, 2, \cdots d\}$ of the features $\mathbf{x}(t)$ can be aggregated periodically with period T_i as

$$\overline{x}_i[l] = \frac{1}{T_i} \int_{(l-1)T_i/2}^{lT_i/2} x_i(\tau) d\tau \tag{5.4}$$

Therefore, instead of the continuous time data $\{(\mathbf{x}(t), y(t)) : t \in \mathbb{R}\}$ specified across t, we are given access to discrete aggregates $\{\overline{y}[k] : k \in \mathbb{Z}\}$ and sets of aggregates $\{\{\overline{x}_i[l] : i = 1, 2, \cdots d\} : l \in \mathbb{Z}\}$.

5.3 Frequency Domain Parameter Estimation from Spatio-Temporally Aggregated Data

We show that an approximately equivalent frequency domain formulation of the problem allows us to sidestep the challenges inherent in a data aggregation setup without explicit reconstruction. Since local time-domain properties are captured by global frequency domain properties, a frequency domain analysis allows us to

individually extract high fidelity estimates of selected global properties of all the quantities involved to then use for inference and predictive modelling.

5.3.1 Frequency Domain Representation of Aggregated Time-Series Data

The first key insight that enables us to work with aggregated time series data is the fact that aggregation in the time domain corresponds to convolution and subsampling in the frequency domain.

Recall that in our setup, continuous signals of the form $z(t)$ get aggregated into samples of the form-

$$\overline{z}[k] = \frac{1}{T} \int_{(k-1)T/2}^{kT/2} z(\tau) d\tau \tag{5.5}$$

There are two steps here. First, the continuous process $z(t)$ is aggregated into the sliding-window averaged continuous process $\overline{z}(t)$ as

$$\overline{z}(t) = \frac{1}{T} \int_{t-T/2}^{t+T/2} z(\tau) d\tau \tag{5.6}$$

This is equivalent to a convolution operation $\overline{z}(t) = z(t) * u(t)$ with the square wave function $u(t) = \frac{1}{T}\mathbb{I}\{t \in (-T/2, T/2)\}$, where $\mathbb{I}\{\cdot\}$ is the indicator function. In the frequency domain, this is equivalent to multiplying with a sinc function $U_T(\omega) = \frac{sin(\omega T/2)}{\omega T/2}$.

The final observation sequence $\{z[k] : k \in \mathbb{Z}\}$ is obtained by sub-sampling at periodicity T the aggregated time series $\overline{z}(t)$; in the frequency domain this becomes

a $\frac{2\pi k}{T}$-periodicity sub-sampling operation, via a convolution with a delta train or a Dirac comb $\frac{1}{T}\sum_{k\in\mathbb{Z}}\delta(\omega-\frac{2\pi k}{T})$.

Therefore, putting it all together, we can write our observation signal in the frequency domain as

$$\bar{Z}(\omega) = \frac{1}{T}\sum_{k\in\mathbb{Z}} Z(\omega-\frac{2\pi k}{T})U_T(\omega-\frac{2\pi k}{T}) \tag{5.7}$$

$$= \frac{1}{T}Z(\omega)U_T(\omega)+\Delta_z(\omega|T) \tag{5.8}$$

where $\Delta_z(\omega|T)=\frac{1}{T}\sum_{k\in\mathbb{Z}\setminus\{0\}} Z(\omega-\frac{2\pi k}{T})U_T(\omega-\frac{2\pi k}{T})$ is the error due to aggregation and **aliasing**.

For succinctness of notation, we assume identical rates for aggregation and subsampling. Estimation is identical in the case where aggregation time period and reporting frequency are different for targets and features (e.g. in case of overlapping aggregation or sliding windows), but the analysis requires some additional bookkeeping - a brief discussion is included in section 5.4.

5.3.2 Formulation and Estimation Algorithm

We now proceed to formulate our parameter estimation framework in the frequency domain. First, we note that the Fourier[1] Transform $z \leftrightarrow \mathcal{F}z$ is a linear operation, therefore the linear relationship that holds in the time domain must also hold in the frequency domain. That is for any signal $\mathbf{x}(t), y(t)$ with noise $\epsilon(t)$, and

[1] as well as the Finite Fourier Transform

Algorithm 1 Fourier-domain Estimation from Aggregated Data

Input: $\overline{x}, \overline{y}, \omega_0, D, T_0$

Sample D frequencies uniformly in $(-\omega_0, \omega_0)$ to get

$$\Omega = \{\omega_1, \omega_2, \cdots \omega_D : \omega_i \in (-\omega_0, \omega_0)\}$$

for each $\omega \in \Omega$, and $i \in \{1, 2, \cdots d\}$ **do**

    compute the T_0-limited finite Fourier transforms

$$\overline{X}_{i,T_0}(\omega) = \mathcal{F}_{T_0}\overline{x}_i(\omega),\ \overline{Y}_{T_0}(\omega) = \mathcal{F}_{T_0}\overline{y}(\omega)$$

    reconstruct non-aggregated Fourier Transforms

$$\widehat{X}_{i,T_0}(\omega) = \frac{\overline{X}_{i,T_0}(\omega)}{U_{T_i}(\omega)},\ \widehat{Y}_{T_0}(\omega) = \frac{\overline{Y}_{T_0}(\omega)}{U_T(\omega)}$$

end for

Estimate the parameter as

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \frac{1}{|\Omega|} \sum_{\omega \in \Omega} \|\widehat{\mathbf{X}}_{T_0}(\omega)^\top \boldsymbol{\beta} - \widehat{Y}_{T_0}(\omega)\|^2$$

return $\widehat{\boldsymbol{\beta}}$

for any $\boldsymbol{\beta}$, we have

$$y(t) = \mathbf{x}(t)^\top \boldsymbol{\beta} + \epsilon(t) \Longleftrightarrow Y(\omega) = \mathbf{X}(\omega)^\top \boldsymbol{\beta} + \boldsymbol{\epsilon}(\omega)$$

Therefore, it stands to reason that if we have good estimates for $Y(\omega), \mathbf{X}(\omega)$ for specific values of ω, parameter estimation should be able to proceed in the frequency domain.

However, the preceding section makes it clear that unless our signals are band-limited, estimates for $\mathbf{X}(\omega)$ and $Y(\omega)$ will be affected by aliasing. Since in the real world we can only work with finite time signals, our signals will never be band-limited because they are time-limited.

Nevertheless, if we assume that the power spectral density for the original signal decays rapidly with ω and, for some ω_0, almost vanishes beyond $|\omega| > \omega_0$. Then, it is easy to see that the effect of aliasing from the sampling process will be minimum for all our signals around $\omega = 0$. Therefore, it makes sense to use only high fidelity estimates of $Y(\omega), \mathbf{X}(\omega)$ for estimation, by restricting ourselves to values of $\omega \in (-\omega_0, \omega_0)$.By doing so, we also bypass any necessity for reconstruction of the original values of our signals in the time domain.

These ideas are the crux of the intuition for our framework and algorithmic treatment of the problem. Section B.1 in Appendix B contains an extended expository discussion that motivates and outlines the steps involved in translating these intuitive ideas to specific algorithmic strategy in mathematical terms.

By formulating the estimation problem in the frequency domain in a way that exactly exploits these intuitive ideas, we can derive our first main result which

shows that under our assumptions, frequency domain parameter estimation leads to generalisation error that is close to the optimal.

Theorem 5.3.1. *Let $\boldsymbol{\beta}^*$ be the optimal parameter as in equation 5.2. Denote the parameter estimated from the T_0-restricted Fourier Transforms as*

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{\omega \in \Omega} E\left[\|\mathbf{X}_{T_0}(\omega)^\top \boldsymbol{\beta} - Y_{T_0}(\omega)\|^2\right] \tag{5.9}$$

Then, for every small $\xi_1, \xi_2 > 0$, there exist correspondingly T_0, ω_0, D such that for the set $\Omega = \{-\omega_0 < \omega_i < \omega_0 : i = 1, 2, \cdots |\Omega|\}$ with $|\Omega| = D$ sampled uniformly between $(-\omega_0, \omega_0)$, we have

$$\begin{aligned} E\left[|\mathbf{x}(t)^\top \widehat{\boldsymbol{\beta}} - y(t)|^2\right] <& (1+\xi_1)\left(E\left[|\mathbf{x}(t)^\top \boldsymbol{\beta}^* - y(t)|^2\right]\right) \\ &+ (1+\xi_1)\xi_2 \end{aligned}$$

with probability at least $1 - e^{-O(D^2\xi_2^2)}$

In essence, this result shows that given a long enough signal, with enough granularity in sampled frequencies, the estimated parameter $\widehat{\boldsymbol{\beta}}$ leads to a generalisation error that is arbitrarily close to the optimal generalisation error obtained by $\boldsymbol{\beta}^*$. Because of the multiple tunable parameters in our formulation, it allows for enough trade-offs that our algorithm can be applied to a wide range of applications (see for example, Wu (2005); Peligrad & Wu (2010); Robert & Casella (1999); Doucet et al. (2001)), and Theorem 5.3.1 can be used as a generic template to derive more precise and bespoke guarantees for each such case. The exact guarantees obtained will depend on the specifics of the application and the data setup– we provide a concrete example of a particular class of common cases in the subsequent section.

5.3.3 Aliasing and Approximation Effects

In real life cases, we have to deal with approximation effects arising from aliasing and randomness of the data that affect our algorithm and analysis procedure, especially in computing our objective function. However, we can show that in most cases the objective function in our estimator as defined in equation (5.9) can be closely approximated with mild regularity assumptions.

For instance, suppose we have data collected independently from N locations with corresponding T_0-restricted Fourier Transforms $\{(\mathbf{X}^j_{T_0}(\omega), Y^j_{T_0}(\omega)) : j = 1, 2, \cdots N\}$ (for example, these can be economic metrics from different states or counties, or meteorological measurements at different points in the atmosphere). We assume that the individual processes at each location is strictly sub-Gaussian (Buldygin & Kozachenko, 2000; Mendelson, 2011). We also assume that the power spectral density of all processes involved is finite for every $\omega \in (-\omega_0, \omega_0)$, and decays rapidly at a sub-Gaussian rate $e^{-O(\omega-\omega_0)^2}$ beyond $|\omega| > \omega_0$.

Then, the following result holds which shows that even for the case where the targets and features are aggregated at different rates, we can still estimate a parameter that leads to a generalisation error that is close to the optimal linear modelling error.

Theorem 5.3.2. *Let T_i be the sampling/aggregation period for the i^{th} coordinate $x_i(t)$ and T_y be the corresponding period for the target $y(t)$. Let $\omega_s = \frac{2\pi}{T_s}$ with $T_s = \max\{T_y, T_1, T_2, \cdots T_d\}$. Denote the parameter obtained by our estimator from N data*

sources as

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{j\in[N]} \sum_{\omega\in\Omega} \|\widehat{X}^j_{T_0}(\omega)^\top \boldsymbol{\beta} - \widehat{Y}^j_{T_0}(\omega)\|^2$$

Then, for every small $\xi_1, \xi_2, \xi_3 > 0$*, there exist correspondingly* T_0, ω_0, D *such that for the set* $\Omega = \{-\omega_0 < \omega_i < \omega_0 : i = 1, 2, \cdots |\Omega|\}$ *with* $|\Omega| = D$ *sampled uniformly between* $(-\omega_0, \omega_0)$*, we have, if the aggregation rate is high enough* $\omega_s > 2\omega_0$*,*

$$\begin{aligned} E\left[|\mathbf{x}(t)^\top \widehat{\boldsymbol{\beta}} - y(t)|^2\right] <& (1+\xi_1)\left(E\left[|\mathbf{x}(t)^\top \boldsymbol{\beta}^* - y(t)|^2\right]\right) \\ &+(1+\xi_1)(\xi_2 + \xi_3 + e^{-O((\omega_s - 2\omega_0)^2)}) \end{aligned}$$

with probability at least $1 - e^{-O(D^2\xi_2^2)} - e^{-O(N^2\xi_3^2)}$

Note that our estimation procedure requires no explicit reconstruction of the original time domain data, which would require spectral information about the signal over the entire spectrum, much of which is severely affected by aliasing effects. In contrast, our methods only use information about the specific parts of the spectrum which are robust and least impacted by aliasing, and are thus more accurate snapshots of the signal.

When the sampling and aggregation periodicity is uniform across all coordinates, an interesting effect can be observed wherein uniform aliasing effects in features and targets essentially cancel each other out. This is because the aliasing error Δ_x for features are related linearly to the error Δ_y for targets via the same parameter. Therefore, parameter estimation can proceed without explicit reconstruction of $\widehat{X}_i(\omega), \widehat{Y}(\omega)$ as a standard linear regression albeit with a slightly different noise

model. However, estimation can still be affected by aliasing in the noise in the signal, therefore, as our experiments on synthetic data shall show, it may preferable to perform estimation in the frequency domain nevertheless.

5.4 Discussion and Extensions

5.1. Multi-dimensional Aggregation:

So far our discussion has been limited to the case where d-dimensional feature vectors $\mathbf{x}$ and real valued targets y are obtained at (and aggregated along) points on a single dimension, i.e., time. We can extend our work very easily to the more general case, where features and targets are indexed by and averaged over points in the p-dimensional Euclidean space $\mathbb{R}^p$.

For example, in spatial climate models, we may use as features $\mathbf{x} \in \mathbb{R}^d$ and targets $y \in \mathbb{R}$ values of meteorological variables (CO_2 levels, temperature, etc.) at discrete points on the earth's surface, indexed by a 2-dimensional (latitude, longitude) vector (i.e., $p = 2$). But instead of $(\mathbf{x}, y)$ for every location, measurements may only be available aggregated averaged over regions on the earth's surface (e.g., averages over 1mi x 1mi spatial grids), which can then be used for learning climate models. Similarly, in 3-dimensional space, $p = 3$, measurements can be obtained aggregated over 3-d blocks. Note that the ambient dimension p is distinct from the dimensionality of the feature space d.

Suppose locations in $\mathbb{R}^p$ are indexed by points $\mathbf{v}$, and each such location is associated with its own d-dimensional feature vector $\mathbf{x}(\mathbf{v}) \in \mathbb{R}^d$ and real valued target

$y(\mathbf{v}) \in \mathbb{R}$, which are regressed on each other via a vector parameter $\boldsymbol{\beta}^* \in \mathbb{R}^d$ as

$$y(\mathbf{v}) = \mathbf{x}(\mathbf{v})^\top \boldsymbol{\beta}^* + \epsilon(\mathbf{v}) \tag{5.10}$$

Each signal here is again a random zero-mean, weakly stationary noise process with finite variance. Observations for any signal[2] $\mathbf{z}(\mathbf{v})$ are again obtained as aggregates over periodically translated bounded connected set $A \subset \mathbb{R}^p$ as

$$z[\mathbf{k}] = \frac{1}{Vol(A)} \int_{\mathbf{v} \in A + \mathbf{k}} z(\mathbf{v}) d\mathbf{v}$$

Given a signal $z(\mathbf{v})$, for any "frequency" vector $\boldsymbol{\theta} = [\theta_1, \theta_2, \cdots \theta_p] \in \mathbb{R}^p$, the **Multidimensional Fourier Transform** is defined in a way very similar to the one-dimensional case (Tangirala, 2014; Easton; Smith & Smith, 1995)

$$Z(\boldsymbol{\theta}) = \int_{\mathbb{R}^p} z(\mathbf{v}) e^{-\iota \langle \boldsymbol{\theta}, \mathbf{v} \rangle} d\mathbf{v} \tag{5.11}$$

where $\langle \cdot, \cdot \rangle$ represents the standard inner product.

All properties of Fourier Transforms required within the scope of this chapter follow exactly as in the unidimensional case (see Easton; Smith & Smith (1995)). For example, aggregation over regions defined by periodic translations of a set $A \subset \mathbb{R}^p$ becomes equivalent to multiplication in the frequency domain with the corresponding multidimensional Fourier Transform of the indicator function $g_A(\mathbf{v}) = \mathbb{I}(\mathbf{v} \in A)$. In particular, if A is the hypercube $A = \{\mathbf{v} : -a_i/2 \leq v_i \leq a_i/2\}$, then $\mathcal{F}g_A(\boldsymbol{\theta}) = \prod_{i=1}^p U_{a_i}(\theta_i)$, where $U_{(\cdot)}$ is the standard sinc function as in the unidimensional case.

The algorithm and results remain virtually identical with unidimensional quantities being replaced by their multidimensional equivalents. The only penalty

[2] where $\mathbf{z}(\mathbf{v})$ is a stand-in for either $\mathbf{x}(\mathbf{v})$ or $y(\mathbf{v})$

that we pay is the number of sampled frequencies required, that is $|\Omega|$, which can in some cases scale exponentially with p. However, we note that in most real life cases p is very small (limited to at most $p = 4$ for spatio-temporal applications), hence this is not a severe impediment on the application on our methods.

5.2. Sliding Windows:

The estimation protocol in this case remains unchanged, but the analysis involves a little extra book-keeping. Note that a sliding window basically means that the aggregation periodicity and sampling periodicity are different. Say T_a is the aggregation period, that is, the period over which averages are computed for the signal (as in equation (5.6)). Also let T_b be the sampling period, that is, the period with which the aggregated signal $\overline{z}(t)$ is sampled. Then, equation (5.7) can be rewritten as

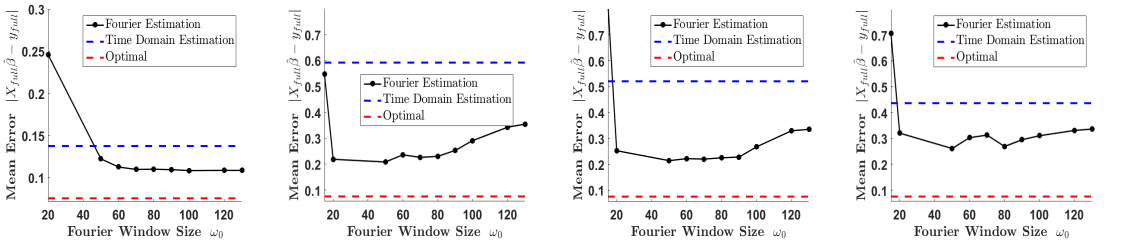
$$\bar{Z}(\omega) = \frac{1}{T_b} \sum_{k \in \mathbb{Z}} Z(\omega - \frac{2\pi k}{T_b}) U_{T_a}(\omega - \frac{2\pi k}{T_b}) \tag{5.12}$$

with a corresponding aliasing error term $\Delta_z(\omega | T_a; T_b) = \frac{1}{T_b} \sum_{k \in \mathbb{Z} \setminus \{0\}} Z(\omega - \frac{2\pi k}{T_b}) U_{T_a}(\omega - \frac{2\pi k}{T_b})$. Theorem 5.3.2 can then be extended to show that in general, if T_a is reasonably small relative to $\frac{2\pi}{\omega_0}$, the aliasing error is dominated by effects from T_b, the sampling period. However, if T_a becomes too large in comparison to $\frac{2\pi}{\omega_0}$, the sinc function $U_{T_a}(\omega)$ can become too sharp and peaky which may result in gaps in the spectrum covered by Ω (refer to the proofs in Appendix B for more details). This is intuitive since larger aggregation windows lead to higher loss of information.

5.3. Aggregation with Weighted Smoothing:

The analysis in the chapter has been presented in the context of a simple aggregation

schema that uses a square wave as a smoothing function for averaging. To cater to alternative aggregation schemata, one just needs to replace the sinc function $U(\cdot)$ with the Fourier Transform of the specific aggregation scheme being used– e.g., for Gaussian smoothing, the relevant Fourier Transform will be another Gaussian, etc. Our results remain unchanged for Schwartz smoothing functions, which includes most of the commonly used smoothing functions. In particular, note that the Gaussian function is a Schwartz function, and so is any smoothing function over a finite support (square wave, triangular wave, etc.), therefore their Fourier Transforms are Schwartz functions as well.

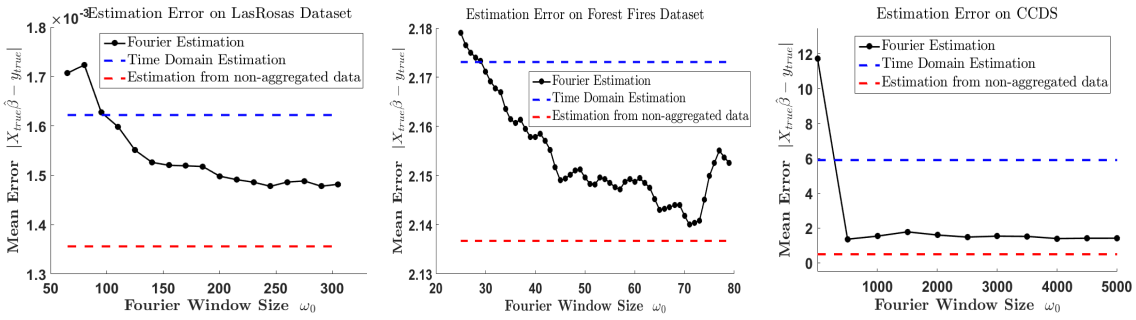


(a) Uniform Sampling **(b)** Low Discrepancy **(c)** Medium Discrepancy **(d)** High Discrepancy

Figure 5.1: Results on Synthetic Data – Mean Estimation Error with increasing Fourier Window ω_0 for uniform aggregation (5.1a), and non-uniform aggregation with increasing discrepancy among aggregation periodicities (5.1b through 5.1d). Frequency domain parameter estimation outperforms naive application of time domain methods

5.5 Experiments

We empirically evaluate the efficacy of our methods on both synthetic data and three real datasets. In each case, we use an aggregated version of the individual-level dataset for learning model parameters using the techniques in this chapter, and evaluate the results by computing the predictive error obtained by our parameter



(a) Las Rosas Dataset **(b)** Forest Fires Dataset **(c)** CCDS Dataset

Figure 5.2: Results on Forest Fires Dataset, Las Rosas Datasets show that frequency domain parameter estimation outperforms naive application of time domain methods and approaches the optimal for high enough ω_0. If ω_0 is too large, however, aliasing effects can lead to deteriorated performance as in Figure 5.2b

on the full non-aggregated dataset. Since this is a first work on this topic, we are unaware of any real algorithmic baselines. However, we do test our methods against two baselines- the "true" linear model which is learned with access to the full non-aggregated dataset, and a "time-domain" model that naively imputes individual-level measurements by substituting the corresponding average for the group.[3]

Our synthetic data experiments proceed as follows. We generate multivariate time series data as features $\mathbf{x}(t)$ and univariate time series data as targets $y(t)$ that obeys our assumptions in this chapter. We then aggregated this data– first using uniform sampling frequency, and second using non-uniform sampling frequency with increasing average discrepancy in the periodicity across features and targets. The aggregated data is then used for learning our model, and the results are compared against a time domain method that imputes the individual aggregates with group

[3]We also tried kriging for resampling i.e. reconstructing the non-aggregated data, then fitting a linear model on the resampled data. This approach performed poorly, hence we omit the results for clarity

level values.

Plots for mean estimation error $|\mathbf{x}(t)\widehat{\boldsymbol{\beta}} - y(t)|$ with increasing Fourier Window ω_0 are shown for the uniform sampling period in figure 5.1a, and for non-uniform sampling period with increasing discrepancy in periodicity in figures 5.1b through 5.1d. In each of these cases, the results show that beyond a certain value of ω_0, frequency domain learning significantly outperforms naive time domain modeling. As described in section 5.3.3, Figure 5.1a shows that for uniform sampling frequency, time domain methods can be still used but our framework is nevertheless preferable because aliasing from error signal can affect estimation accuracy in the time domain. Moreover, as we describe in the chapter, the performance of frequency domain estimation deteriorates if the value of ω_0 becomes too high because aliasing effects start distorting the results.

The first real spatio-temporal dataset involves an application from agricultural studies, wherein corn yield monitor data las from the Las Rosas agricultural plantation in Cordoba, Argentina is regressed against features including nitrogen levels, topographical properties, brightness value, etc. (see Bongiovanni & Lowenberg-DeBoer (2000); Lambert et al. (2004) for further details on the dataset).

The second real dataset is the Forest Fires Dataset from the UCI Machine Learning Repository University of California, Irvine which involves predictive modelling of burned acreage from forest fires in the northeast region of Portugal. by using as features meteorological and other data like relative humidity, ISI index, etc. (see Cortez & Morais (2007) for more details on the dataset).

In both these datasets, the data points are stamped with latitude-longitude positional indices, which are used to topographically order each observation. The ordered data is then aggregated based on positional indices and used for learning a linear model.

In our final experiment, we test our techniques on the Comprehensive Climate Dataset (CCDS) which is an extensive collection of climate modeling variables for North America compiled from various sources including NASA, National Oceanic and Atmospheric Administration (NOAA), National Climate Data Center (NCDC), etc. (see Lozano et al. (2009); Liu et al. (2010) for further details on the dataset). We use this dataset to model atmospheric vapour levels using various measurements, including carbon dioxide, methane, cloud cover, etc. and other extra-meteorological factors like rate of frost/rainy days, etc. over a grid that covers most of continental United States. This collection contains two datasets, one of which is aggregated and the other is observed at a much higher resolution. We use the aggregated dataset for learning $\widehat{\boldsymbol{\beta}}$ and test the predictive performance of our learned model on the higher resolution dataset.

Figures 5.2a, 5.2b and 5.2c show plots for mean estimation error $|\mathbf{x}(t)\widehat{\boldsymbol{\beta}} - y(t)|$ with increasing Fourier Window ω_0 for each of the three real datasets. Our results show that in all three datasets, for a large enough ω_0 our method significantly outperforms the corresponding time domain technique, and starts coming close to the performance of the optimal estimator.

5.6 Conclusion

In this chapter we investigated the problem of predictive modelling of linear models involving correllated spatio-temporal data when the data is available only in aggregated form rather than as individual-level measurements with localised estimates. In particular, we analysed the scenario where aggregation is non-uniform across targets and different coordinates of the features, leading to significant challenges in cogent mathematical representation of any relationship among available feature and target aggregates. We showed that by formulating the problem in the frequency domain and exploiting duality properties of Fourier analysis, many of the inherent structural challenges of this setting can be bypassed. We introduced a novel framework and new algorithmic techniques to perform frequency domain estimation and inference for this setup and provided both theoretical guarantees and empirical validation of our methods.

Chapter 6

Aggregation Paradigms and Learning with Sensitive Data

6.1 Introduction

So far in this dissertation, we have considered scenarios where data available for analysis was aggregated using a pre-defined schemata (histograms, group-wise moments, spatio-temporal averages, etc.) that was determined by the proprietor of the data, and was beyond the control of the practitioner who was actually analysing the data. In this chapter we consider the complementary problem, where the entity that is performing the aggregation has a common interest with the entity performing the data analysis, but they cannot store or share the data at the finest possible level of granularity for analysis without violating legal or ethical principles. However, being stakeholders in any results or inferences obtained by the data analysts, they need to perform the aggregation in such a manner as to ensure that the aggregates can still be used in a learning algorithm to train any machine learning models.

In all such cases, any machine learning solution has to operate under the constraint that data will only be available in small subsets for brief periods, after which they need to be aggregated by the system. In particular, most legal requirements like (Federal Trade Commission, 2005, 2018) stipulate that any stored information

compiled from sensitive data (e.g., patient records, user data, etc.) be in such a form that preserves privacy entirely, and guarantees non-identifiability.

Our setup is very different from standard privacy preserving data mining(Lindell & Pinkas, 2000). Most methods in this line of work are either based on cryptography or perturbation (Aggarwal & Philip, 2008), or it studies the setting where a common model is learnt with data stored across multiple sites but with restrictions on data sharing among these sites (Merugu & Ghosh, Nov, 2003, 2005). In each of these cases, individual datapoints are retained (up to noise) in at least one location– in our setup, they have to be deleted everywhere. Sketching methods that maintain summary statistics of the data often do not preserve privacy, or they tend to modify the relationship between covariates with targets, especially in non-linear models (Liberty, 2013). Stream data mining (Leskovec et al., 2014) are an extreme version of our setup where each data point is seen exactly once and then deleted immediately – most existing methods in that line of work can only be used to estimate very basic data summaries that are not informative enough for predictive modelling. SGD and other streaming techniques (Bottou, 2010; Recht et al., 2011) that store approximate gradients have a poor rate of convergence, and usually require repeated access to each data point (in the form of training epochs) to learn an effective model, which is disallowed by our problem setup. In fact, as we show in our experiments, SGD performs rather poorly in comparison to the methods we introduce in this work.

To summarise, our work is motivated by two complementary objectives:

1. design aggregation paradigms that protect data security and privacy

2. formulate learning algorithms that can use these aggregates or summaries to train predictive models that are effective at individual level predictions

This is a tall task, but we show that both these objectives can be achieved in several cases.

Motivating Example: Gaussian Regression

To illustrate our ideas, consider the case of Gaussian regression where covariates $\mathbf{X}$ are related to targets $\mathbf{y}$ via a linear parameter $\boldsymbol{\theta}$ as $\mathbf{y} = \mathbf{X}\boldsymbol{\theta} + \epsilon, \;\; \epsilon \sim \mathcal{N}(0, \sigma^2)$ It is well known that the MLE parameter $\boldsymbol{\theta}_{MLE}$ can be obtained in closed form from the data as

$$\boldsymbol{\theta}_{MLE} = (\mathbf{X}^\top \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{y}$$

Clearly, we do not require the entire dataset – the only relevant quantities that are required for learning the model are aggregates $\mathbf{X}^\top\mathbf{X}$ and $\mathbf{X}^\top\mathbf{y}$ which can be re-written as

$$\mathbf{X}^\top \mathbf{X} = \sum_{i=1}^{N} \mathbf{x}_i \mathbf{x}_i^\top, \quad \mathbf{X}^\top \mathbf{y} = \sum_{i=1}^{N} y_i \mathbf{x}_i^\top$$

Therefore, if we store only these aggregates, and delete the individual datapoints themselves, we can still recover the MLE parameter *error-free* without access to the raw dataset. For the Gaussian case, therefore, we have an exact solution.

Furthermore, this aggregation scheme also preserves data privacy since these aggregates cannot be used to reconstruct any individual data points within any tolerance level (see section 6.4).

This particular paradigm, of course, only applies to Gaussian regression, but we can use this as a basic modus operandi for other setups as well. In this chapter, we expand on these ideas and introduce a new framework of aggregation paradigms and learning algorithms that satisfy both our stated objectives for two other common models – binary classification and generalised linear models.

As an auxiliary contribution, we note that common privacy paradigms like differential privacy (Dwork, 2008) that are favoured in academia and other technical domains can often deemed too esoteric (Schneps & Colmez, 2013) to be satisfactory to certain regulatory bodies. Therefore, we introduce an alternative privacy criterion that is stringent yet easily comprehensible in lay terms, and we show how our methods satisfy this constraint – an extended discussion on this is deferred to section 6.4.

Contributions:

Our specific contributions are summarised below:

1. We design novel aggregation paradigms and learning algorithms that guarantee privacy while still allowing learning for a wide variety of models. We motivate our methods with Gaussian regression, and extend our methods to binary classification and generalised linear models. To our knowledge, we are the first to tackle this exact problem setup.

2. We provide a theoretical analysis as well as empirical evaluation for our methods with experiments on data from telecommunication and healthcare

3. Finally, we introduce the notion of Reconstructive Privacy and Total Recon-

structive Privacy as an alternative criterion for non-identifiability. We further show that each of our aggregation paradigms satisfy these constraints.

We call our framework SLAGG, or SLice and AGGregate, after the main steps involved in the procedure. While keeping the overall approach fairly simple and intuitive, we prove strong guarantees on its performance, and also show very favorable empirical results.

6.2 Problem Definition

In this work, we consider predictive models that are trained via supervised learning. Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots \mathbf{x}_N] \in \mathbb{R}^{N \times d}$ be a set of N data points in a d-dimensional feature space, and let $\mathbf{y} = [y_1, y_2, \cdots y_N] \in \mathcal{Y}^N \subseteq \mathbb{R}^N$ be their corresponding targets. We assume that their exists a function f such that for each $(\mathbf{x}, y)$ pair, we have $y = \mathrm{f}(\mathbf{x}) + \eta$, where η is random noise.

The standard machine learning setup estimates this function f using a training set of the form $\mathbb{D} = (\mathbf{X}, \mathbf{y}) \equiv \{(\mathbf{x}_i, y_i) : i = 1, 2, \cdots\}$ and a learning protocol that consists of solving the following optimisation problem

$$\mathrm{f}^* = \arg\min_{\mathrm{f} \in \mathcal{F}} \sum_{(\mathbf{x}, y) \in \mathbb{D}} \mathcal{L}(\mathrm{f}(x), y) \tag{6.1}$$

where $\mathcal{L} : \mathcal{Y} \times \mathcal{Y} \mapsto \mathbb{R}^+$ is a loss function that measures the discrepancy between predicted $\mathrm{f}(\mathbf{x})$ and measured y (e.g., negative log-likelihood).

In our setup, the full dataset is not available for training. Instead, the data is divided into M disjoint "chunks" or subsets as $\mathbb{D}_T = \{(\mathbf{x}_i, y_i) : i \in \mathcal{I}_T\}$, where

$\mathcal{I}_T \subset [N]$ are partitions of the index set, and $T = 1, 2, \cdots M$. For example, $\mathbb{D}_T$ may be customer data or patient records for the T^{th} month, which need to be compiled into non-identifiable aggregates and the individual data points are to be deleted at the beginning of the $(T+1)^{th}$ month for privacy reasons.

Therefore, instead of the full dataset $\mathbb{D}$, the learner is only allowed access to each chunk, one at a time, for a brief period of time. The learner's task is to use these chunks to learn a non-identifiable aggregates before the individual data points in each chunk are deleted. Finally, the learner will be required to devise a training algorithm for the final predictive model that only use these aggregates.

6.3 Aggregation Design Paradigms

The question now is how to use these chunks to learn an effective estimate of the function f. For this, we take inspiration from the concept of "sufficient statistics" in estimation theory that studies various methods to estimate a parameter for a distribution given data. Let $\boldsymbol{\theta}$ be a parameter to be estimated from a given dataset $\mathbb{D}$. A sufficient statistic for $\boldsymbol{\theta}$ is a quantity (or a set of quantities) $\mathcal{S}$ computed from the dataset $\mathbb{D}$ such that the posterior of the parameter given the statistic is independent of the individual datapoints themselves, that is, $P(\boldsymbol{\theta}|\mathcal{S}, \mathbb{D}) = P(\boldsymbol{\theta}|\mathcal{S})$. Basically, a sufficient statistic summarises the dataset by extracting from the individual datapoints all the information that is necessary for parameter estimation, and discards the rest.

Our task here is similar – given a data chunk, extract the useful information from the data chunk in the form of aggregates that can be subsequently used for

training a final predictive model. We now discuss specific instantiations of both an aggregation paradigm as well as a learning algorithm that only uses these aggregates. We have already seen this idea in action for the case of Gaussian linear regression. In the rest of the chapter, we extend these methods to the case of binary classification and generalised linear models.

6.3.1 Binary Classifiers

Unlike the Gaussian case, there is no nice closed form solution for most binary classification models. In fact, the model parameter itself may not always be unique and suffer from identifiability issues owing to rotational or scale invariance. Therefore, we study the case of binary classification not in formal model specification terms, but by treating a classifier as a black box with a specific probability of error over the population.

In particular, consider the case where one has access to multiple noisy classifiers. One can consider the output of each of these classifiers as noisy estimates for the "true" class label (defined as the mode of $P(y|x)$), and by taking the majority vote, one can estimate the true class label with high accuracy. Therefore, if we can "aggregate" each data chunk to learn a black box noisy binary classifier, we no longer need individual training datapoints themselves to get the final predictive model.

Hence, our protocol is the following:

1. For each data chunk $\mathbb{D}_T$, learn a classifier $\mathrm{f}_T : \mathcal{X} \mapsto \{0, 1\}$ from only the data points in $\mathbb{D}_T$

2. Given a new random sample $\mathbf{x}$, and the classifiers $\{\mathrm{f}_T : T = 1, 2, \cdots M\}$, obtain the corresponding predictions $\{\widehat{y}_T = \mathrm{f}_T(\mathbf{x}) : T = 1, 2, \cdot M\}$

3. Obtain the final estimate for the class label as

$$\widehat{\mathrm{f}(\mathbf{x})} = \text{median}\{y_T : T = 1, 2, \cdots M\} \tag{6.2}$$

We now analyse the predictive accuracy of our final classifier. To account for unavoidable noise and limitations of model class, we compare the performance of our method to the best possible model from the function class that can be learned from the individual non-aggregated data points. Let λ be the probability of misclassification on a randomly selected data point for the best possible model f^* from the function class. For any $\mathbf{x}$, let $z_T(\mathbf{x}, y) = \mathbb{I}\{\mathrm{f}_T(\mathbf{x}) \neq y\}$ where $\mathbb{I}$ is the indicator function. Note that since each data chunk $\mathbb{D}_T$ consists of i.i.d samples of the same size, z_T are independent and identically distributed random variables over the probability space for data chunks. For any $\mathbf{x}$, let $p = E[z_T]$ be an upper bound on the misclassification probability for the T^{th} classifier.We then have the following result:

Proposition 6.3.1. *Let $p < 0.5$ and $\widehat{f}$ be our final classifier from M data chunks as defined in equation 6.2. Then, the probability that $\widehat{f}$ does worse than f^* on any given datapoint is upper bounded by the quantity:*

$$\frac{1-p}{(1-\lambda)(1-2p)}\left[(1-p)p\ exp\left(2\kappa - \xi_M + \zeta_M\right)\right]^{M/2}$$

where $\kappa \approx 0.693$, $\xi_M \sim O(\frac{logM}{M})$, and $\zeta_M \sim O(\frac{1}{M^2})$

See **Appendix** C for the proof. It is easy to see that as M increases, the probability of error rapidly decreases. Recall from Ng & Jordan (2002) that the performance of the classifer learned from each chunk is close to model optimal if the size of each chunk is $N_T \sim \Omega(d)$ where d is the dimensionality of the data. Note that one corollary of this result is that **a learner that uses data chunks can potentially learn better than a single learner that uses the full non-aggregated dataset**. Indeed, this is exactly what happens with our experiments on real data as we show in section 6.5.

Multi-Class Case

Our analysis extends to the multi-class case by treating it as multiple 2-class classification, and then using union bound to get an upper bound on error. A similar result holds as above, with an additional multiplicative cost factor, which can be tuned by taking into consideration certain trade-offs. For example, if we use one-vs rest classification, the cost factor is L, where L is the number of classes. However, in this case, the number of data points in each chunk required to get an appropriate p is higher. To avoid this, one can use pair-wise classification models for each pair of class labels, but this latter step will introduce a cost-factor of $\binom{L}{2} \sim O(L^2)$.

6.3.2 Generalised Linear Models and Exponential Family Distributions

We now extend our techniques to generalized linear models or GLMs (McCullagh & Nelder, 1989; Nelder & Baker, 1972) which are generalizations of linear

regression that subsume various models like Poisson regression, logistic regression, etc. as special cases. They are the standard workhorse in a wide variety of domains, and can be used to model a wide variety of data types – Gaussian for real-valued, Poisson regression for integer valued, logistic for binary, log-Normal for non-negative reals, etc. Many such applications fall under our problem setup as they arise in sensitive domains with restricted access to data due to privacy and security constraints.

A GLM is usually parametrized by a convex function $\boldsymbol{\phi}$ (usually assumed or known, see (Banerjee et al., 2005; Acharyya & Ghosh, 2014)) and a parameter $\boldsymbol{\theta}$ (to be learned from data). Given a predictor $\mathbf{x}$ and a parameter $\boldsymbol{\theta}$, a generalised linear model generates the target y as follows. First, it computes the linear function of the predictor $\mathbf{x}^\top\boldsymbol{\theta}$ (also known as the canonical or natural parameter). Next, it transforms this scalar value using a monotonic link function $g_\phi(\cdot)$ (that depends on $\boldsymbol{\phi}$, see Banerjee et al. (2005)) in order to bring the real valued $\mathbf{x}^\top\boldsymbol{\theta}$ to the domain of y (e.g., logit if y is binary, exponential if y is non-negative, etc.).

The transformed scalar value $g_\phi(\mathbf{x}^\top\boldsymbol{\theta})$ is known as the mean parameter, and the target y is generated from it using a probability distribution $P_{\boldsymbol{\phi}}$ from the exponential family such that $E_{P_{\boldsymbol{\phi}}}(y|\mathbf{x}) = g_\phi(\mathbf{x}^\top\boldsymbol{\theta})$. The specific $P_{\boldsymbol{\phi}}$ depends on the GLM used (e.g. Poisson for Poisson regression, Bernoulli for logistic regression, etc.) but for each case, the probability distribution for $y|\mathbf{x}$ in a GLM takes the following form:

$$P_{\boldsymbol{\phi}}(y|\mathbf{x},\boldsymbol{\theta}) \propto \ exp\left(y \cdot \mathbf{x}^\top\boldsymbol{\theta} - \mathbf{G}_{\boldsymbol{\phi}}(\mathbf{x}^\top\boldsymbol{\theta})\right) \tag{6.3}$$

where $\mathbf{G}_{\boldsymbol{\phi}}$ is such that $g_\phi \equiv \nabla\mathbf{G}_{\boldsymbol{\phi}}$.The typical GLM parameter estimation algorithm proceeds by the standard log-likelihood minimisation approach that requires individual level training datapoints. The protocol in section 6.3.1 (maintaining independent

predictors and aggregating their outputs) cannot be used since for most GLM's (e.g., Poisson regression), the variance of the output can be very large and heteroskedastic, unlike for binary random variables. Hence, we need an alternative paradigm for estimating the GLM parameter.

Unbiased Estimators

Generally speaking, learning the MLE parameter $\boldsymbol{\theta}^*$ for a GLM from anything other than individual data points can be an intractable problem (Montanari et al., 2015). However, we can still approximate the model parameter using a learning protocol that is efficient and easily computable. For this, we use the idea of concentration in statistics – given a random variable $z \sim P_z$ such that $E[z] = \mu$, we can approximate μ by averaging samples $z_1, z_2, \cdots z_M$ where each z_i is drawn independent and identically distributed according to P_z. For example, using Hoeffding's inequality one can show that if z is a sub-Gaussian random variable and $\hat{\mu} = \frac{1}{M}\sum_i z_i$, the concentration can be exponentially fast in M – for any small $\epsilon > 0$, we have $P\left(|\hat{\mu} - \mu| > \epsilon\right) < exp\left(-O(M\epsilon)\right)$

Therefore, our data summarisation protocol basically consists of using each chunk of data $\mathbb{D}_T$ to try and compute noisy but unbiased estimates $\widehat{\boldsymbol{\theta}}_T$ of the "true" model parameter $\boldsymbol{\theta}$ as in equation 6.4. Let $\mathbb{P}$ be an unbiased estimator that takes any dataset $\mathbb{D}$ and outputs an estimate for the parameter $\mathbb{P}(\mathbb{D})$ such that $\forall \mathbb{D},\ E_{\mathbb{D}}[\mathbb{P}(\mathbb{D})] = \boldsymbol{\theta}^*$, the optimal model parameter. For any data chunk T, define $\widehat{\boldsymbol{\theta}}_T = \mathbb{P}(\mathbb{D}_T)$ as result of the estimator applied to the data chunk. We define our final parameter estimate

as

$$\widehat{\boldsymbol{\theta}} = \frac{1}{M}\sum_{T=1}^{M}\widehat{\boldsymbol{\theta}}_T$$

It is easy to see that if M is high enough, then with high probability, $\widehat{\boldsymbol{\theta}} \to \boldsymbol{\theta}^*$. Note that the quality of the individual $\widehat{\boldsymbol{\theta}}_T$ might be very poor because the data chunks used to learn $\widehat{\boldsymbol{\theta}}_T$ may be very small, but as long as they are unbiased and have finite variance, the average will converge to the optimal parameter.

Learning Protocol

We now give some intuition for our learning protocol. Recall that the probability distribution for $\mathbf{y}|\mathbf{X}$ in a GLM takes the following form:

$$P_{\boldsymbol{\phi}}(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}) \propto \prod_{(\mathbf{x},y)} exp\left(y \cdot \mathbf{x}^\top\boldsymbol{\theta} - \mathbf{G}_{\boldsymbol{\phi}}(\mathbf{x}^\top\boldsymbol{\theta})\right) \tag{6.4}$$

With $\nabla\mathbf{G}_{\boldsymbol{\phi}} \equiv g_\phi$, the gradient of the log likelihood with respect to $\boldsymbol{\theta}$ is

$$\nabla_{\boldsymbol{\theta}} LL(\mathbf{y}|\mathbf{X},\boldsymbol{\theta}) = \sum_{(\mathbf{x},y)}\left(y - g_\phi\left(\mathbf{x}^\top\boldsymbol{\theta}\right)\right)\mathbf{x} \tag{6.5}$$

Setting this to zero for the optimal $\boldsymbol{\theta}^*$ gives

$$\mathbf{X}g_\phi(\mathbf{X}^\top\boldsymbol{\theta}) = \mathbf{X}\mathbf{y} \tag{6.6}$$

where $g_\phi(\cdot)$ is applied elementwise. Clearly, eq. (6.6) does not have a closed form solution except when $\mathbf{X}, \mathbf{X}^\top$ are both invertible. Suppose we divided up $\mathbb{D}$ in chunks of d data samples each[1], where d is the dimensionality of the data. Then for each T,

[1]For reasons stabilitiy, in practice it is better to have a matrix that is slightly tall, that is, n slightly greater than d

we can obtain a parameter $\widehat{\boldsymbol{\theta}}_T$ that is locally optimal for the samples corresponding to the data chunk $\mathbb{D}_T$.

Therefore, our learning protocol can be summarised as follows –

1. For each data chunk $\mathbb{D}_T$, compute a locally optimal parameter as

$$\widehat{\boldsymbol{\theta}}_T = (\mathbf{X}_T\mathbf{X}_T^\top)^{-1}\mathbf{X}_T\mathbf{g}_{\boldsymbol{\phi}}^{-1}(\mathbf{y}_T)$$

2. Using the individual $\widehat{\boldsymbol{\theta}}_T$ for each data chunk $\mathbb{D}_T$, compute the final estimate for the global GLM parameter as

$$\begin{aligned}\widehat{\boldsymbol{\theta}} &= \frac{1}{M}\sum_{T=1}^{M}\widehat{\boldsymbol{\theta}}_T \\ &= \frac{1}{M}\sum_{T=1}^{M}(\mathbf{X}_T\mathbf{X}_T^\top)^{-1}\mathbf{X}_T\mathbf{g}_{\boldsymbol{\phi}}^{-1}(\mathbf{y}_T)\end{aligned}$$

Here, $\mathbf{g}_{\boldsymbol{\phi}}^{-1}$ is defined element-wise. In case y is outside the domain of $\mathbf{g}_{\boldsymbol{\phi}}^{-1}$, one can use any projection of y to the interior o the domain of $\mathbf{g}_{\boldsymbol{\phi}}^{-1}$ instead. We have the following result:

Proposition 6.3.2. *If $\mathbf{g}_{\boldsymbol{\phi}}^{-1}$ (equivalently g_ϕ) is a linear function, $\widehat{\boldsymbol{\theta}}$ is an unbiased estimator of $\boldsymbol{\theta}^*$*

See **Appendix** C for the proof. It follows that everything described in the previous section follows, and $\widehat{\boldsymbol{\theta}} \to \boldsymbol{\theta}^*$ if the number of data partitions M is large enough. Note that this covers a large class of GLM's because the link function is effectively linear for many exponential family distributions, like Gaussian, Exponential, Pareto, Chi-Squared, etc.

When g_ϕ is not linear

When the link function is not linear, there are a few possible directions one can take. The entire protocol remains the same, except $\mathbf{g}_{\boldsymbol{\phi}}^{-1}(y)$ now needs to be replaced by $E[\mathbf{g}_{\boldsymbol{\phi}}^{-1}(y)|\mathbf{x}]$. This quantity can be approximate using many methods, one such technique being sampling. Note that for a given $\mathbf{x}$, the target y is a sample from $P_{\boldsymbol{\phi}}(y|\mathbf{x}, \boldsymbol{\theta}^*)$ centred around $g_\phi(\mathbf{x}^\top \boldsymbol{\theta})$. Therefore, we take y as an estimate for $g_\phi(\mathbf{x}^\top \boldsymbol{\theta})$, use this y as the mean to sample $y_1, y_2, \cdots y_k$ i.i.d. from $P_{\boldsymbol{\phi}}$, and then compute $\frac{1}{k}\sum_i \mathbf{g}_{\boldsymbol{\phi}}^{-1}(y_i)$ as an estimate for $E[\mathbf{g}_{\boldsymbol{\phi}}^{-1}(y)|\mathbf{x}, \boldsymbol{\theta}^*]$. Another way is to use $\widehat{\boldsymbol{\theta}}_T$ as an estimate of θ^*, and use $g_\phi(\mathbf{x}^\top \widehat{\boldsymbol{\theta}}_T)$ as an estimate for for $g_\phi(\mathbf{x}^\top \boldsymbol{\theta})$, and then perform the sampling as in the previous case.

For both methods, we repeat the two steps, approximating $E[\mathbf{g}_{\boldsymbol{\phi}}^{-1}(y)|\mathbf{x}]$ and estimating $\widehat{\boldsymbol{\theta}}_T$, alternatingly until convergence. Note that this procedure is structurally similar to expectation-maximisation, which is a common technique for parameter estimation in latent variable models.

6.4 Discussion on Privacy

A key motivation for all our work so far has been data privacy, both in the context of ethics as well as new regulatory frameworks like the GDPR. While privacy has been studied in machine learning and statistics, the main lens for almost all prior art is differential privacy (Dwork, 2008; Dwork et al., 2014) which quantifies the amount of change to the output of an algorithm when you change the dataset by a single datapoint (usually exponential in terms of change to the dataset).

While differential privacy is fairly well accepted in academia and certain technical domains, it is still a rather abstract and esoteric criterion that is not easily comprehensible to lay persons. In certain scenarios like legal regulation, court cases, healthcare, etc. with high privacy sensitivity, or where the stakeholders may not have a strong mathematical background, differential privacy may be too technical to be sufficiently reassuring from either a human or a legal perspective (see (fivethirtyeight.com, 2017; Schneps & Colmez, 2013)). Out of an abundance of caution, practitioners often require privacy criteria that are stringent, yet also easily accessible to, for example, a judge or a regulator who may not possess the requisite mathematical foundation.

For this purpose, as an alternative to differential privacy, we introduce the notion of **Reconstructive Privacy** that is a strong guarantee yet easy to understand in non-mathematical terms.

Definition 6.4.1. Let $\mathbb{D} = \{\mathbf{z}_i : i = 1, 2, \cdots N\} \in \mathcal{Z}^N$ be a dataset, and $\mathbf{A} : \mathcal{Z}^N \mapsto \Lambda$ be an aggregating function. Let $\mathbb{D}' \in \mathcal{Z}^N$ be a candidate dataset such that $\mathbf{A}(\mathbb{D}) = \mathbf{A}(\mathbb{D}')$ Then, $\mathbf{A}$ is defined to preserve τ-**Reconstructive Privacy** over $\mathbb{D}$ if there exists at least one such $\mathbb{D}'$ such that the minimum distance between any two points in the two datasets is at least τ:

$$\sup_{\substack{\mathbb{D}' \in \mathcal{Z}^N \\ \mathbf{A}(\mathbb{D}')=\mathbf{A}(\mathbb{D})}} \min_{\mathbf{z} \in \mathbb{D}, \mathbf{z}' \in \mathbb{D}'} \|\mathbf{z} - \mathbf{z}'\| \qquad \geq \tau$$

If τ is unbounded, then $\mathbf{A}$ is said to preserve **Total Reconstructive Privacy**.

The basic idea is that aggregation keeps the data private since there are multiple candidate datasets (arbitrarily far apart) that lead to the same aggregates making it difficult to reconstruct the original from aggregates alone. Note that we define the "distance" between two datasets as the minimum distance between any two datapoints within the dataset, rather than the average distance between the datasets. In particular, for any $\tau > 0$, this forces the original dataset and any candidate dataset to have no overlap in their samples.

Note that unlike the idea of differential privacy, the notion of total privacy does not involve any consideration of the effectiveness or usability of an algorithm that satisfies the condition. The definition is purely a criterion that restricts the kind of data that can be stored – it is standalone and tuned exclusively to ensure privacy, and it is up to individual algorithms to satisfy performance benchmarks without violating total privacy constraints.

In the previous sections, we have already shown how our algorithms and learning paradigms satisfy performance requirements. Now we show that our aggregation frameworks also satisfy total reconstructive privacy constraints. For the Gaussian case, our aggregates satisfy total reconstructive privacy since the set of $(\mathbf{X}, \mathbf{y})$ that can lead to the summary statistics comes from the full $\mathbb{R}^{N \times d}$ as long as the datapoints are arranged in general position.

Privacy guarantees for binary classifiers depends a bit on the model. For models with a linear class boundary, total reconstructive privacy is maintained since the same boundary is still optimal for data points arbitrarily far away but balanced for each class. For non-linear models like kernelised SVM, the $y \sim sign\left(\mathrm{f}(\mathbf{x})\right)$ form

still allows scale invariance for most standard versions, hence total reconstructive privacy is still maintained.

Note that in case of SVM's, the dual formulation for kernelised SVM defines the classifier explicitly in terms of individual data points, i.e., support vectors. However, for any non-linear kernel we can use the idea of Randomised Kitchen Sinks (Rahimi & Recht, 2008, 2009) to re-parametrize these classifiers into the standard primal formulation so that the individual data points can still be deleted with only the parameter $\boldsymbol{\theta}$ being stored.

The GLM case is similar to the Gaussian case since it is easy to verify that the set of datapoints that lead to the same aggregate comes from the full vector space. Moreover, note that in equation 6.4, for any parameter $\boldsymbol{\theta}$, we get both rotational invariance (rotating each $\mathbf{x}$ around $\boldsymbol{\theta}$ keeps their product unchanged) as well as scale invariance since one can multiply each $\mathbf{x}, y$ by some $\gamma, \beta > 0$ such that $P_\phi(\gamma y|\beta\mathbf{x}, \boldsymbol{\theta}) \propto P_\phi(y|\mathbf{x}, \boldsymbol{\theta})$, thereby leading to the same optimal parameter.

6.5 Experiments

We demonstrate the efficacy of our methods with empirical evaluation on three real datasets from the healthcare and telecom domains where our problem setup is particularly relevant. We do not show experiments for Gaussian models since our algorithm is exact for that case, and hence will always arrive at the optimal solution. Due to the properties of the datasets, we use Logistic Regression for our experiments on binary classifiers, and Poisson regresssion for the experiments on data with real-valued (non-negative integer) targets (see Acharyya & Ghosh (2014) for a discussion

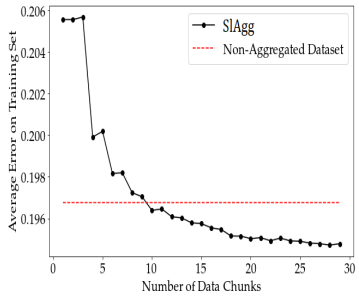
on GLM model selection).

Since this is a first work, we do not know of any alternative algortihmic competitors for our methods. Hence, in each case, we compare against three baselines. First, we use as a performance "upper-bound" the results obtained from learning from the full non-aggregated dataset with individual level samples. Second, we compare against an SGD learner. Third, we use an ecological regression (EcoReg) baseline (King et al., 2004; Brown & Payne, 1986) that treats aggregates as individual level samples and uses them for training.

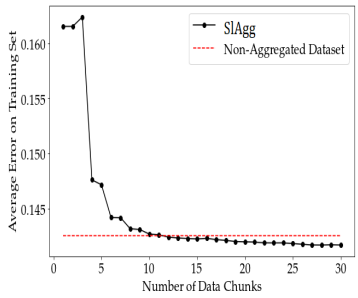
As performance metrics, we compare test error on unseen data as well as training error on the full training dataset (since our setup assumes that full training dataset is not available), averaged over 100-fold randomised cross validation (error bars were minuscule and are omitted for clarity). We show plots of performance metrics versus number of learners/data chunks seen by our method, as well as a final table of results. SGD and EcoReg are included only in the table and omitted from plots for clarity, since their performance is rather poor in comparison.

6.5.1 Binary Classification: Churn in Telecom:

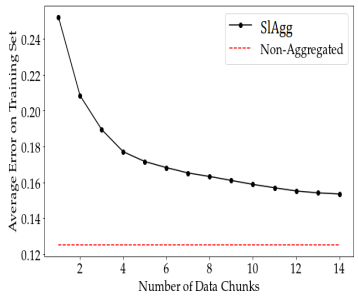
We use two datasets from the Telecom industry for our binary classification tasks. In both cases, the objective is to predict churn (Hung et al., 2006) from customer account and usage details. In the telecom industry, churn or attrition refers to the event where a customer terminates a service or contract with a particular company. Predicting churn in advance is critical for business, since dissatisfied cus-



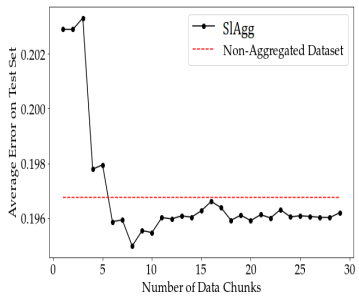
(a) Train Error on IBM (churn) dataset



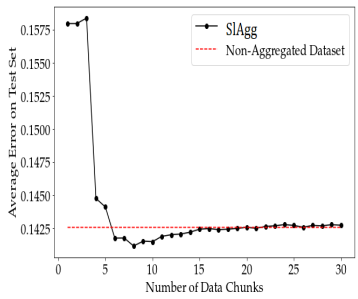
(b) Train Error on Kaggle (churn) dataset



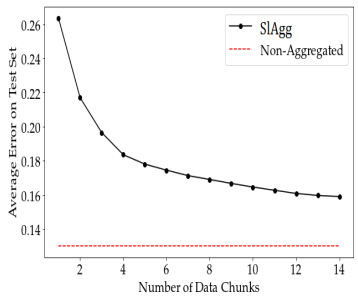
(c) Train Error on DESynPUF



(d) Train Error on IBM (churn) dataset



(e) Train Error on Kaggle (churn) dataset



(f) Train Error on DESynPUF

Figure 6.1: Training and Test error vs Number of Data Chunks on IBM, Kaggle and DESynPUF datasets: even with very few data chunks, our algorithm performs comparable to or even better than that obtained from a non-aggregated dataset
Note 1: Our algorithm does better than a binary classifier trained with non-aggregated data, exactly as predicted by Prop 6.3.1.
Note 2: Training error is shown here since our algorithm does not has full access to the training dataset

tomers can potentially be prevented from leaving by, for example, special offers and incentives, or by catering to specific complaints the customer might have.

Our first dataset is a Telecom dataset provided by IBM Watson Analytics. It consists of data from 7044 customers, including demographic information like gender, senior citizen and dependent status, as well as service information including monthly charges and tenure, plan details like data backup, online security and device

	# chunks		Non-Agg	Our Work	SGD	EcoReg
IBM (churn)	29	Train	0.1967	0.1947	0.285	0.356
		Test	0.197	0.196	0.287	0.357
Kaggle (churn)	30	Train	0.142	0.142	0.245	0.216
		Test	0.142	0.143	0.245	0.218
DESynPUF	14	Train	0.125	0.153	1.785	0.22
		Test	0.130	0.159	1.797	0.23

Table 6.1: Final Training and Test Error on all three datasets (with full data used) for learner with non-aggregate data, our method, SGD and naive averaging. Our method outperforms baseline and has performance very close to learner with full, non-aggregated dataset

protection, as well as usage details like streaming and internet (details of the dataset can be found in (IBM TJ Watson)).

Our second dataset is a very similar dataset from Kaggle (Kaggle Churn in Telecom) about customer churn. Similar to the first dataset, this one has information like state, phone plan details like international calling, as well as usage details like total calls and number of voicemail messages. In sum, the dataset has records for 3334 customers.

With minor processing (categorical variables converted to one-hot encoding, for example) we collect the data into chunks (of size 2.5 times the dimensionality of the data) and feed it into the various algorithms and note the results. The results (Fig 6.1 and Table 6.5) show that for both datasets, our algorithm needs only a few data chunks to achieve a performance better than learner with non-aggregated

dataset, and significantly outperforms SGD and EcoReg.

6.5.2 Real-valued data: Healthcare

We now apply our methods on a healthcare dataset where the objective is to estimate Medicare charges. This application falls under the purview of our problem setup since patient privacy and ethical considerations limit access to healthcare records for data analysis.

We use the CMS Beneficiary Summary DE-SynPUF dataset (DESynPUF, 2008) for our experiments. This a public use dataset created by the Centers for Medicare and Medicaid Services by applying different statistical disclosure limitation techniques to Medicare beneficiary claims data. We use a subset of the DE-SynPUF dataset for Louisiana state from the year 2008 and model outpatient institutional annual primary payer reimbursement (*PPPYMT-OP*) with available predictor variables that include age, race, sex, duration of coverage, presence of a variety of chronic conditions, etc.

Because of the nature of our target variables, we use a Poisson regression model for this application. The data is collected into chunks with size the same order as the dimensionality, with a few extra data points[2] in each chunk for stability. We feed the data chunks into each algorithm (and the full dataset, in case of the non-aggregated baseline) and compile the results. For EcoReg, there were not enough aggregates to learn with because of "perfect separation" issues, so we boosted its

[2]we used 15 extra data points, but results with other choices were very similar

performance by using twice the number of aggregates for model training.

The results (Fig 6.1 and Table 6.5) show that our techniques with only a few data chunks can perform very close to a learner with access to the full dataset. Note that there is a slight performance gap with respect to the optimal because the link function for Poisson is non-identity, and therefore we require an additional approximation step as detailed in section 6.3.2. SGD did not show promising performance. EcoReg seems to work better here, but recall that its performance was boosted, and it still failed to compare favourably with the our method.

6.6 Conclusion

In this chapter we tackle the problem of learning in the scenario when privacy, scalability, security, etc. concerns limit access to training data only in the form of chunks that need to be aggregated and deleted after a specific duration of time. We design aggregation techniques, as well as algorithms to learn models from these aggregates that can nevertheless work at the individual level. We motivate our techniques by using Gaussian regression, and subsequently extend them to the case of binary classification and GLMs. We provide both theoretical results as well as empirical evaluation for our work. Finally, we introduce a new alternative criterion for privacy preservation, as well as show that our methods satisfy that criterion.

Chapter 7

Predicting Cost-per-Click in Online Advertising from Aggregated Invoices

7.1 Introduction

One of the key areas in industry where aggregated data shows up quite frequently is when two or more different companies work on a common platform to provide a service or a product to consumers. A key coordination issue involves how the companies can work together without compromising the integrity, security and privacy of their own proprietary data. In these situations, it is very common for individual companies to protect proprietary ownership by only sharing their data with third parties in an aggregated form.

Online advertising (Goldfarb & Tucker, 2011; Perlich et al., 2014) is a domain where this scenario arises all the time. The advertising process (Zeff & Aronson, 1999; Yan et al., 2009) consists of two main players– the advertiser (e.g. Criteo or AdRoll) who provides the ads and bids for advertising space, and the publisher (e.g. Google or Facebook) who provides the platform on which to display such ads. The highest bidder is allowed to place an ad of their choice on the provided platform, and charged a specific fee whenever an end-user clicks on the ad. These transactional charges (called cost-per-click or CPC) are based on a pre-negotiated but only partially

declared auction mechanism that depends, among other things, on the advertiser's bid, the position of the ad, demand etc. Modelling and estimating CPC beforehand is, therefore, a key component of the complex bidding strategy (Ghosh et al., 2009; Yuan et al., 2014; Feldman et al., 2010), and also in the decision-making process for the specific advertisement to use which will optimally balance critical trade-offs between cost incurred and potential revenue earned from a given ad-space.

Unfortunately, for various reasons like transactional efficiency, protection of proprietary mechanism design, etc. publishers often submit the invoice of charges to advertisers on a cumulative basis (total charges over a day, etc.) rather than on a per-click basis. Therefore, the transaction data available to the advertiser for training their models only consists of aggregated CPC values, where the per-click charges have been obfuscated through averaging. A specific example of this type of data sharing happens on the Google Shopping product (NORC) where the platform shares only the aggregated and not per click cost information. Developing a framework to address this use case– training individual-level models with aggregated data– is thus a critical bottleneck in online advertising.

While we use online advertising as a motivating example, the specific setup we consider is ubiquitous across a much wider variety of domains. Aggregation is used as a statistical disclosure limitation technique in many privacy sensitive domains like healthcare (Park & Ghosh, 2012, 2014) where confidential information like hospital records are often aggregated to protect individual patients' privacy. In large scale data collection settings like census or population surveys (NORC) or meteorological studies (Lozano et al., 2009; Liu et al., 2010), individual level data is often collected

or stored as aggregates for scalability reasons. Sensor networks and IoTs (Wagner, 2004; Zhao et al., 2003) use data aggregation in the interest of robustness, when measurements by individual sensors tend to be corrupted with noise that gets canceled out when the measurements are averaged over space or time. Financial forecasting applications depend on economic metrics which are often released as aggregates (US Department of Labour; US Department of Commerce) by governmental agencies and independent think tanks.

The key problem we focus on in this work is this– how do we build models that work at the individual level but that can nevertheless be trained with data collected at the aggregate level? Unfortunately, despite its near universal presence, learning from aggregated data is still a relatively unexplored topic, and there are rarely any easy answers. This is a new and extremely challenging semi-supervised learning paradigm, and naive application of standard techniques almost always fail because of the ecological fallacy (Robinson, 2009; Kramer, 1983), wherein inferences drawn at the group level are significantly different from those drawn at the individual level.

In this chapter, we introduce a novel modelling and algorithmic framework to learn individual level models for the case when the target variables of interest are collected into group-wise aggregates. We emphasise that we use online advertising only as an example application for easier presentation of our learning framework. Our methods are extremely general and can be used for any application that involves learning from aggregated target variables. To this end, we use as our base modelling framework generalised linear models, which are a large class of models

that can handle diverse data types (real valued, binary, integer, etc.) and are the primary work-horse in a vast range of domains, from climate science to recommendation engines to healthcare. Since nearly all existing work on generalised linear models assume access to individual level data, we introduce significant modelling innovations on top of existing theory and algorithms in prior art that enables us to extend this large class of models to the aggregated data scenario.

Contributions: Our specific contributions are outlined below–

1. We introduce a novel framework that can learn individual level generalised linear models when the target data is available only as aggregates computed over sub-groups of the data space. To the best of our knowledge, we are the first to tackle this problem.

2. We design a new learning algorithm that uses alternating data imputation and estimation steps to train the generalised linear model with access only to aggregated target variables

3. We extend our analysis to cover cases where the data aggregation has been performed over arbitrary grouping paradigms to subsume cases like overlapping aggregation, sliding window, non-uniform aggregation, etc.

4. We empirically evaluate our methods on both synthetic and real datasets from the advertising domain to show the efficacy of our techniques. We further demonstrate the general applicability of our methods by evaluating the performance of our techniques on problems from the healthcare domain.

7.2 Problem Description

Let $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \cdots \mathbf{x}_n]$ be a set of feature vectors for n data samples and $Y = [y_1, y_2, \cdots y_n]$ be their corresponding values for the target variables of interest. In the interest of clarity and notational succinctness, we start off by describing our framework within the simplified context of linear models, and extend our analysis to the more general version in section 7.3. For a linear model, the corresponding target y for each covariate $\mathbf{x}$ is generated via a vector parameter $\boldsymbol{\beta}$ as

$$y = \mathbf{x}^\top \boldsymbol{\beta} + \epsilon$$

where ϵ is a zero mean noise term. Maximum likelihood parameter estimation involves solving an optimisation problem to minimise the regularised negative log-likelihood of the observed data,

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \sum_{(\mathbf{x},y)\in(\mathbf{X},Y)} \mathcal{L}(\mathbf{x}^\top \boldsymbol{\beta}, y) + \lambda \mathcal{R}(\boldsymbol{\beta}) \tag{7.1}$$

where $\mathcal{L}$ is the loss function ($\mathcal{L} = \|\cdot\|^2$ for Gaussian noise) and $\mathcal{R}(\cdot)$ is an appropriate regulariser (ℓ_2 for ridge regression, ℓ_1 for LASSO, etc.).

In the standard regression setting, the data used for training the model is

available at the individual level, in the form of n pairs of targets and their corresponding features as $\mathbb{D}_{(x,y)} = \{(\mathbf{x}_i, y_i) : i = 1, 2, \ldots n\}$, so $\boldsymbol{\beta}^*$ may be estimated using standard machine learning techniques. In our scenario, we do not have access to data at this level of granularity– instead, while $\mathbf{X}$ is fully observed, the target values $\{y_i : i = 1, 2, \cdots n\}$ are subjected to an aggregation process (partially specified, and not controlled by the learning agent) that produces a set of m summaries $\{z_k : k = 1, 2, \cdots m\}$ that are then made available to us. That is, instead of the individual y_i, we are provided with summaries $z_k = \sum_{i \in \mathcal{I}_k} y_i$, where each $\mathcal{I}_k \subset \{1, 2, \cdots m\}$ is an index set that defines which target variables contribute to a given aggregate.

For example, each y_i may be the individual CPC and z_k is the aggregate CPC value computed over all click activity in the k^{th} aggregation window, while $\mathbf{x}_i$ may be information like that is available in full to the advertiser like price or brand for products, or target country, campaign type, etc. for the ad itself. In privacy-sensitive applications, z_k might refer to aggregated information like health metrics or income aggregated at the zipcode level, while $\mathbf{x}_i$ may refer to data like demographic information (race, gender, etc.) available publicly at the individual level from, say, voter files. We assume for now that the aggregation indices are disjoint, that is, $\mathcal{I}_k \cap \mathcal{I}_{k'} = \varphi$ for $k \neq k'$, we extend this to the overlapping case in section 7.3.2.

We now introduce our learning framework for $\boldsymbol{\beta}$ that can predict individual level targets $\hat{y}$, but use only the aggregates z for training. Denote the set of n feature vectors stacked up into a matrix as $\mathbf{X} \in \mathbb{R}^{n \times d}$ and the set of m aggregates as $\mathbf{z} \in \mathbb{R}^m$. For notational succinctness, we use $\mathbf{y} \in \mathbb{R}^n$ to denote imputed or predicted targets,

and use $\mathbf{y}_{true} \in \mathbb{R}^n$ for the "ground truth". We call our framework EstImAgg or EStimate-IMpute for AGGregated Data after the main steps involved in the learning process.

Algorithm 2 EstImAgg-Simplified
Non-overlapping aggregation, Gaussian model

Input: $\mathbf{X}, \mathbf{z}$, Aggregation groupings $\mathcal{I}_k$
Initialise $\{y_i = z_k : i \in \mathcal{I}_k\}$
while not converged **do**
    Solve for $\boldsymbol{\beta}^+$ using standard methods given $\mathbf{y}$

$$\boldsymbol{\beta}^+ = \arg\min_{\boldsymbol{\beta}} \sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{y})} \|\mathbf{x}^\top \boldsymbol{\beta} - y\|^2 + \lambda \mathcal{R}(\boldsymbol{\beta})$$

    for each $k \in \{1, 2, \cdots m\}$ **do**
        Compute imputed aggregate error given new $\boldsymbol{\beta}^+$

$$\gamma_k = \frac{1}{|\mathcal{I}_k|}\left(z_k - \sum_{i\in\mathcal{I}_k} \mathbf{x}_i^\top \boldsymbol{\beta}^+\right)$$

        Impute each target based on the true aggregate

$$\forall \;\; i \in \mathcal{I}_k \;\; : \;\; y_i^+ = \mathbf{x}_i^\top \boldsymbol{\beta}^+ - \gamma_k$$

    end for
    Update variables $(\mathbf{y}, \boldsymbol{\beta}) = (\mathbf{y}^+, \boldsymbol{\beta}^+)$
end while
return $\mathbf{y}, \boldsymbol{\beta}$

In the standard case, learning the model effectively implies minimising the loss function only over the parameter $\boldsymbol{\beta}$, as in Equation 7.1. In our case, we not only have to estimate the parameter $\boldsymbol{\beta}$, but also the non-aggregated targets $\mathbf{y}$ subject to the constraints that the imputed estimates agree with the aggregates $\mathbf{z}$, which adds an extra set of constraints $\sum_{i\in\mathcal{I}_k} y_i = z_k \;\; \forall \; k = \{1, 2, \cdots m\}$ to the optimisation

problem 7.1.

We solve this using alternating minimisation. The first step, solving for $\boldsymbol{\beta}$ given a particular value of $\mathbf{y}$, is a simple regression parameter estimation problem. The second step, solving for $\mathbf{y}$ given a particular estimate for $\boldsymbol{\beta}$, is more interesting. For a Gaussian model ($\mathcal{L} = \|\cdot\|^2$), the optimisation problem is as follows

$$\mathbf{y}^+ = \underset{\mathbf{y}}{\operatorname{argmin}} \sum_{(\mathbf{x},y)\in(\mathbf{X},\mathbf{y})} \|\mathbf{x}^\top\boldsymbol{\beta} - y\|^2 \;\text{ s.t. } \sum_{i\in\mathcal{I}_k} y_i = z_k \;\; \forall\; k \tag{7.2}$$

By using standard optimisation theory it can be proved that the optimal value for $\mathbf{y}$ for this problem can actually be obtained in closed form. As shown in Algorithm 2, the solution involves applying an additive update to each estimated target to make it compatible with the aggregates.

The learning steps described in plain English above is summarised as Algorithm 2. Note that while we do not explicitly assign a mathematical form for regularisation function $\mathcal{R}$, it only appears in our algorithm in the estimation of the parameter $\boldsymbol{\beta}$. While the specific form depends on the structure imposed on $\boldsymbol{\beta}$ (ℓ_2 for ridge regression, LASSO for sparsity, etc.), this nevertheless remains a standard regularised regression problem and off-the-shelf estimators are available for almost any such choice of $\mathcal{R}$ that is commonly used in practice.

7.3 General Formulation

In this section, we generalise our framework to handle a wider class of problems by extending our methods to incorporate generalised linear modelling, and

modifying our algorithm to handle arbitrary aggregation paradigms.

7.3.1 Generalised Linear Models

While least squares regression is useful for modeling continuous real valued data generated from a Gaussian distribution, this is not always a valid assumption. In many cases, the data of interest may be binary valued or count valued, and generalised linear models are more appropriate for such scenarios. A detailed note on GLM's is provided in Chapter 2, here we only summarise the main concepts. In a GLM, the response variables, y are generated from a distribution in the exponential family centred around a mean parameter that is related to a linear function of the predictor $\mathbf{x}$ via a monotonic link function often denoted as $(\nabla\phi)^{-1}(\cdot)$. Here, $\boldsymbol{\phi}$ is a convex function that depends on the specific exponential family distribution used (Banerjee et al., 2005). Specifically, given a predictor $\mathbf{x}$, a parameter $\boldsymbol{\beta}$ and a probability distribution $P_{\boldsymbol{\phi}}$ from the exponential family, the target y is obtained according to the distribution $P_{\boldsymbol{\phi}}$ such that

$$y|\mathbf{x} \sim P_{\boldsymbol{\phi}}(\eta_{\mathbf{x}}), \quad \text{where} \quad \eta_{\mathbf{x}} = E_{P_{\boldsymbol{\phi}}}(y|\mathbf{x}) = (\nabla\phi)^{-1}(\mathbf{x}^{\top}\boldsymbol{\beta})$$

7.3.1.1 Loss Function: Bregman Divergences

As noted in Chapter 2, the matching loss functions associated with learning GLM parameters are distance-like functions called Bregman divergences, which are generalisations of square loss. Let $\boldsymbol{\phi} : \Theta \mapsto \mathbb{R}$ be a strictly convex, closed function on a convex domain $\Theta \subseteq \mathbb{R}^n$, that is differentiable on $\text{int}(\Theta)$. Then, for any $\mathbf{a}, \mathbf{b} \in \Theta$, the Bregman divergence $D_{\boldsymbol{\phi}}(\cdot\|\cdot)$ between $\mathbf{a}$ and $\mathbf{b}$ corresponding to the function $\boldsymbol{\phi}$

is defined as

$$D_{\boldsymbol{\phi}}(\mathbf{a}\|\mathbf{b}) \triangleq \boldsymbol{\phi}(\mathbf{a}) - \boldsymbol{\phi}(\mathbf{b}) - \langle \nabla\boldsymbol{\phi}(\mathbf{b}), \mathbf{a} - \mathbf{b}\rangle$$

where g_ϕ is the gradient of the function $\boldsymbol{\phi}$, applied elementwise. Bregman divergences are convex in their first argument. Although strictly speaking they are not a distance metric, they satisfy many properties of metrics, for example $D_{\boldsymbol{\phi}}(\mathbf{a}\|\mathbf{b}) \geq 0$ for any $\mathbf{a}, \mathbf{b}$, and $D_{\boldsymbol{\phi}}(\mathbf{a}\|\mathbf{b}) = 0$ if and only if $\mathbf{a} = \mathbf{b}$. Many standard distance-like functions like Square loss and KL-divergence are members of this family (see Table 2.1 in Chapter 2).

Bregman Divergences have a very close relationship with generalized linear models. In particular, there is a one-to-one correspondence between each GLM and each Bregman divergence via the convex function $\boldsymbol{\phi}(\cdot)$ that is also closely related to the specific exponential family distribution associated with the GLM.

Specifically, for our work we use the fact that MLE parameter estimation in a GLM with given object features $\mathbf{X}$ and target variable $\mathbf{y}$ is equivalent to minimising $D_{\boldsymbol{\phi}}\left(\mathbf{y}\|(\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta})\right)$ over $\boldsymbol{\beta}$, that is, the optimal parameter is the minimiser $\widehat{\boldsymbol{\beta}}$ for the following optimisation problem,

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{(\mathbf{x},y)} D_{\boldsymbol{\phi}}\left(y\|(\nabla\phi)^{-1}(\mathbf{x}^\top\boldsymbol{\beta})\right) + \lambda\mathcal{R}(\boldsymbol{\beta})$$

where $\boldsymbol{\phi}(\cdot)$ is the convex function associated with the particular GLM used. For example, maximum likelihood for a Gaussian model or standard linear regression corresponds to square loss, for Poisson the corresponding divergence is generalized I-divergence (GI-divergence) and for Binomial, the corresponding divergence is the

Kullback-Leibler or KL divergence. We refer the reader to Banerjee et al. (2005) for a detailed exposition on the relationship between Bregman Divergences and GLM's.

In particular, we note that the only aspect of our framework that is affected by generalising linear regression to GLMs is in the loss function, which now changes from a square loss to a general Bregman Divergence.

7.3.2 Extension to Overlapping Aggregation

In section 7.2, we studied and derived an algorithm for the case when the aggregation is non-overlapping. We now generalise this to a wider class of aggregation paradigms. The basic formulation remains unchanged– the aggregation paradigm only shows up in the constraint set of our optimisation framework as a linear constraint.

Consider the matrix $\mathbf{\Gamma} \in \mathbb{R}^{m \times n}$ such that for each $k \in \{1, 2, \cdots m\}$ and each $i \in \{1, 2, \cdots n\}$, we have $\mathbf{\Gamma}_{ki} = 1$ if and only if $i \in \mathcal{I}_k$, where $\mathcal{I}_k$ are the aggregation groupings as defined in section 7.2. Then, it is clear that $\mathbf{\Gamma y} = \mathbf{z}$ exactly captures the aggregate summaries $z_k = \sum_{i \in \mathcal{I}_k} y_i$.

Extensions to arbitrary aggregation paradigms is now obvious– we just design the appropriate $\mathbf{\Gamma}$ matrix. For example, overlapping aggregation can be represented by a $\mathbf{\Gamma}$ where multiple rows have 1's in the same column. A sliding window aggregation framework with window size τ can be represented with a $\mathbf{\Gamma}$ such that $\mathbf{\Gamma}_{ik} = 1$ for $i \in \{k\tau, k\tau + 1, \cdots (k+1)\tau\}$, and 0 otherwise.

We can extend this further to the case where the aggregation is weighted– in this case, $\mathbf{\Gamma}$ is a matrix with general real-valued entries, rather than a binary matrix.

One example where this occurs is when the grouping is done over sub-populations sampled non-uniformly, and the aggregates are computed from weighted summaries that represent unbiased estimates of the mean parameter for the target value over the sub-populations. Another example of weighted aggregation is when the target variables are multiplied with a random matrix in the interest of statistical disclosure limitation. All these cases can be handled in the same manner as for a binary matrix by using an appropriate $\mathbf{\Gamma}$.

7.3.3 Learning Algorithm for the General Case

We are now ready to describe the modelling framework and the solution algorithm for the general case. Let $\boldsymbol{\phi}$ be the convex function on which the Bregman divergence corresponding to the GLM used is defined. We overload our notation and use $g_\phi(\cdot)$ and $(\nabla\phi)^{-1}(\cdot)$ to denote functions applied to the individual elements of the vector, that is for a vector $\mathbf{a}$, we have $g_\phi(\mathbf{a}) = [g_\phi(a_1), g_\phi(a_2) \cdots g_\phi(a_n)]$ and $(\nabla\phi)^{-1}(\mathbf{a}) = [(\nabla\phi)^{-1}(a_1), (\nabla\phi)^{-1}(a_2) \cdots (\nabla\phi)^{-1}(a_n)]$ whenever this is well defined. Let $\mathbf{\Gamma}$ be the aggregation matrix as defined in section 7.3.2. Then, the optimisation problem for the general version of the problem is

$$\begin{aligned} \min_{\mathbf{y},\boldsymbol{\beta}} \quad & D_{\boldsymbol{\phi}}\left(\mathbf{y} \| (\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta})\right) + \lambda\mathcal{R}(\boldsymbol{\beta}) \\ \text{s.t.} \quad & \mathbf{\Gamma}\mathbf{y} = \mathbf{z} \end{aligned} \tag{7.3}$$

As earlier, we use alternating minimisation to solve this optimisation problem for the imputed targets $\mathbf{y}$ and parameter $\boldsymbol{\beta}$ respectively. Given a particular set of

values for the imputed targets, estimating the parameter is equivalent to solving

$$\boldsymbol{\beta}^{+} = \arg\min_{\boldsymbol{\beta}} D_{\boldsymbol{\phi}}\left(\mathbf{y} \| (\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta})\right) + \lambda\mathcal{R}(\boldsymbol{\beta})$$

This is the standard formulation for estimating the parameter for a generalised linear model, and off-the-shelf packages are available for most programming platforms to solve this. The more interesting problem here is the imputation of targets $\mathbf{y}$ given a particular value of $\boldsymbol{\beta}$, which can be summarised as the following optimisation problem

$$\begin{aligned} \mathbf{y}^{+} = \quad & \underset{\mathbf{y}}{\text{argmin}} \quad D_{\boldsymbol{\phi}}\left(\mathbf{y} \| (\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta})\right) \\ & \text{s.t.} \qquad\quad \boldsymbol{\Gamma}\mathbf{y} = \mathbf{z} \end{aligned} \tag{7.4}$$

Because Bregman Divergences are convex in their first argument, and because the constraint set is linear, this optimisation problem is convex in terms of $\mathbf{y}$. In fact, not only is the optimisation problem convex, it can be shown that we can actually estimate the optimal value of $\mathbf{y}$ in closed form

Lemma 7.3.1. ***Target Imputation:*** *Given $\boldsymbol{\beta}$, the optimality conditions for the optimisation problem as described in equation 7.4 lead to the parameter $\mathbf{y}^{+}$, where $\mathbf{y}^{+}$ is defined as*

$$\mathbf{y}^{+} = (\nabla\phi)^{-1}\left[\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\Gamma}^{\top}\left(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^{\top}\right)^{-1}\boldsymbol{\Gamma}\left(\mathbf{X}\boldsymbol{\beta} - g_{\phi}\left((\boldsymbol{\Gamma}^{\top}\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}\mathbf{z}\right)\right)\right] \tag{7.5}$$

where the operations g_{ϕ} and $(\nabla\phi)^{-1}$ are applied elementwise to their corresponding vector arguments.

The proof of this follows directly by using either optimality conditions on the Lagrangian, or by using the Karush-Kuhn Tucker conditions for the constrained optimisation problem.

Algorithm 3 EstImAgg

Arbitrary aggregation, Generalised Linear Models

Input: $\mathbf{X}, \mathbf{z}$, GLM $\sim \boldsymbol{\phi}$, Aggregation matrix $\boldsymbol{\Gamma}$

Initialise $\{y_i = \sum_k \frac{z_k}{|\mathcal{I}_k|} : i \in \mathcal{I}_k\}$

while not converged **do**

Solve for $\boldsymbol{\beta}^+$ using standard GLM estimation

$$\boldsymbol{\beta}^+ = \arg\min_{\boldsymbol{\beta}} D_{\boldsymbol{\phi}} \left(\mathbf{y} \| (\nabla\phi)^{-1}(\mathbf{X}\boldsymbol{\beta})\right) + \lambda\mathcal{R}(\boldsymbol{\beta})$$

Compute imputed aggregate error given new $\boldsymbol{\beta}^+$

$$\boldsymbol{\vartheta} = \left(\mathbf{X}\boldsymbol{\beta} - g_\phi\left((\boldsymbol{\Gamma}^\top\boldsymbol{\Gamma})^{-1}\boldsymbol{\Gamma}\mathbf{z}\right)\right)$$

Transform aggregate error

$$\boldsymbol{\xi} = \boldsymbol{\Gamma}^\top \left(\boldsymbol{\Gamma}\boldsymbol{\Gamma}^\top\right)^{-1} \boldsymbol{\Gamma}\boldsymbol{\vartheta}$$

Impute each target applying a monotonic transform

$$\mathbf{y}^+ = (\nabla\phi)^{-1}\left(\mathbf{X}\boldsymbol{\beta} - \boldsymbol{\xi}\right)$$

Update variables $(\mathbf{y}, \boldsymbol{\beta}) = (\mathbf{y}^+, \boldsymbol{\beta}^+)$

end while

return $\mathbf{y}, \boldsymbol{\beta}$

The steps involved in the overall algorithm for the general case (GLMs with arbitrary aggregation paradigms) is summarised as Algorithm 3. For better presentation and intuitive clarity, we separate out the target imputation step into three parts– they can, of course, be combined in any implementation as in equation 7.5. We use the generalised inverse in the matrix inversion steps whenever the matrices involved are not full rank.

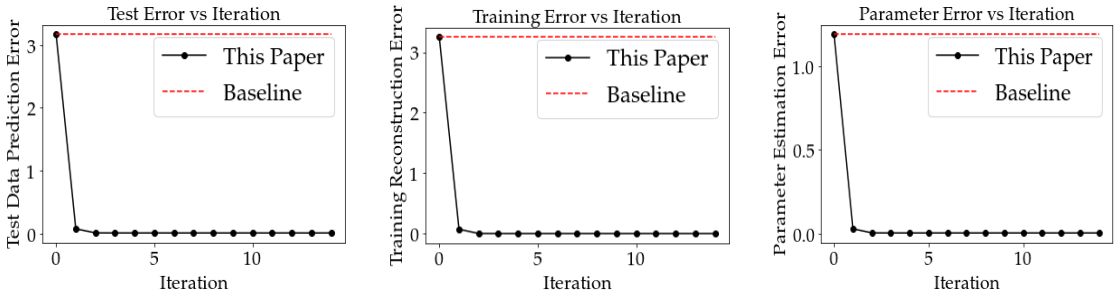
Convergence: While the optimisation problem is not jointly convex, and the solution methodology in algorithms 2 and 3 involve alternating estimation and imputation steps, it can be shown that both the algorithms always converge.

Lemma 7.3.2. ***Convergence:*** *For any choice of initialisation and set of inputs, both algorithms 2 and 3 converge to a local minimum.*

This fact can be proved using the observation that every step in the algorithm reduces the value of the objective function, and the fact that the objective function is bounded below by 0. The objective function is not jointly convex except for specific types of Bregman Divergences (Acharyya et al., 2012; Acharyya & Ghosh, 2014), hence convergence is local.

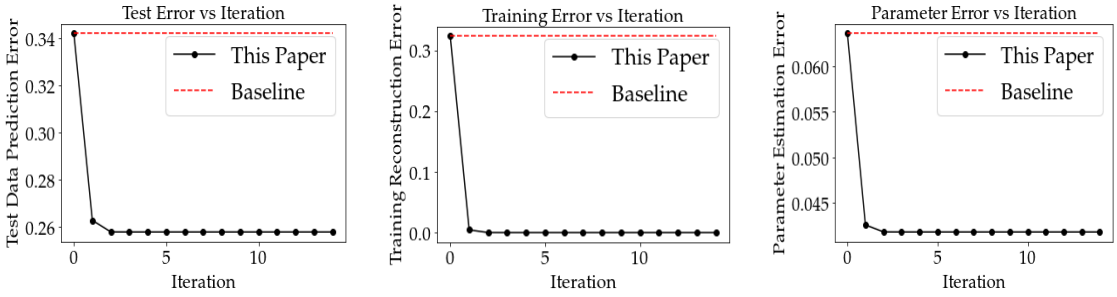
7.4 Experiments

We evaluate our methods on synthetic data as well as three real datasets with applications in online advertising and healthcare. As mentioned earlier, because of a lack of existing literature on this topic we are not aware of any algorithmic baselines



(a) Test Prediction Error **(b)** Train Reconstruction Error **(c)** Parameter Estimation Error

Figure 7.1: Synthetic Data (Gaussian): Error on predictions for test data, error in reconstructed training data and estimation error for parameter recovery plotted vs iteration for Gaussian Model. Our model outperforms the baseline in all three metrics and converges within very few iterations



(a) Test Prediction Error **(b)** Train Reconstruction Error **(c)** Parameter Estimation Error

Figure 7.2: Synthetic Data (Poisson): Error on predictions for test data, error in reconstructed training data and estimation error for parameter recovery plotted vs iteration for a Poisson Regression Model. Our model outperforms the baseline in all three metrics and converges within very few iterations

for our work. Therefore, the performance of our method is compared against a **straightforward baseline** which plugs in the aggregates z as individual level labels into standard machine learning estimators and learns a predictive model.

The evaluation metrics that we use are threefold. First, following standard practice we examine the performance of our algorithms in predicting target values on an unseen test set– to this end, we compute the ℓ_2 prediction error $\|\mathbf{y}_{true}^{test} - (\nabla\phi)^{-1}(\mathbf{X}^{test}\boldsymbol{\beta}_{estim})\|$ between the true value of the test target and the estimated

value obtained by our algorithm. All of this is done at the individual-sample level of granularity. Next, note that while reconstruction is not an explicit objective, our algorithm nevertheless involves a data imputation step. Therefore, whenever we have access to ground truth data for the training set, we also evaluate the ℓ_2 reconstruction error $\|\mathbf{y}_{true}^{train} - \mathbf{y}_{recons}^{train}\|$ between the true value of the training target and the imputed values. This is compared against the ℓ_2 error as obtained by replacing every target variable by the aggregate. Finally, whenever we have access to the "true" parameter of the GLM, we also compare the parameter recovery error $\|\boldsymbol{\beta}_{true} - \boldsymbol{\beta}_{estim}\|$ of the estimated parameter for both our algorithm as well as the baseline.

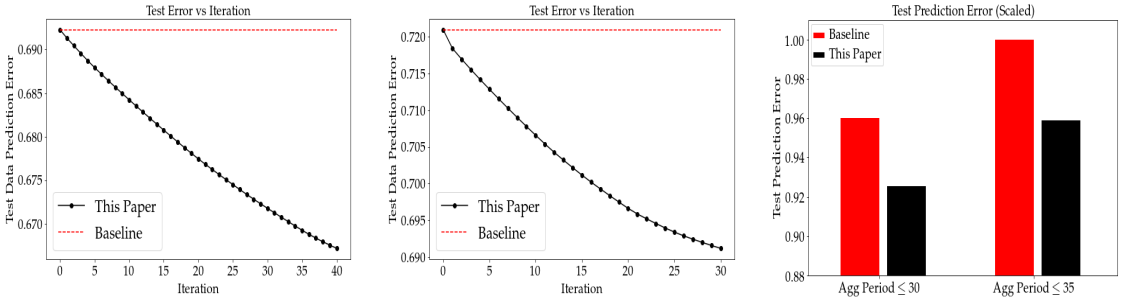
7.4.1 Synthetic Data

We run two different experiments on synthetic data– one for a Gaussian regression model and one for a Poisson regression model. In both cases, we generate covariates $\mathbf{X}$ independently and identically distributed according to the standard Normal. We also generate the true parameter $\boldsymbol{\beta}^*$ by sampling from the multivariate standard normal distribution in the same manner. We then use the parameter and the covariates to generate the corresponding targets $\mathbf{y}$. For the Gaussian model, the targets are generated as the linear function of each $\mathbf{x}$ with $\boldsymbol{\beta}^*$, that is, $y|\mathbf{x} \sim \mathcal{N}(\mathbf{x}^\top \boldsymbol{\beta}^*, \sigma^2)$. For the Poisson model, following standard practice (McCullagh & Nelder, 1989; Banerjee et al., 2005) we further apply the canonical exponential link function to $\mathbf{x}^\top \boldsymbol{\beta}^*$ to bring this linear function to the domain of a Possion distribution. That is, we generate each y from their covariate $\mathbf{x}$ as $y|\mathbf{x} \sim \text{Poiss}\,(\eta_{\mathbf{x}})$, where $\eta_{\mathbf{x}} = exp(\mathbf{x}^\top \boldsymbol{\beta}^*)$ is the mean around which y is distributed.

In each case we generate the aggregation set $\mathbf{\Gamma}$ in the following manner– for each aggregate, we sample a Bernoulli variable for every data point independently with probability $\rho = 5\%$. If we sample 1, the datapoint is used for computing the aggregate, otherwise not. This is equivalent to setting every element of $\mathbf{\Gamma}$ independently as Bernoulli(ρ). Initialisation for $\mathbf{y}$ is deterministic, and uses the aggregates as the "true" individual labels.

We then feed these aggregates to our algorithm together with the grouping information $\mathbf{\Gamma}$. The evaluation metrics are threefold. First, we compute prediction error at the individual level for each method on samples in the test. Next, we calculate error in reconstructing the training data as imputed by our method. Finally, since we have the "true" parameter for this set of experiments, we also plot parameter recovery error.

Figures 7.1 and 7.2 show the results for the Gaussian regression model and the Poisson regression model respectively. In both cases, it can be seen that our algorithm significantly outperforms the baseline and converges within a very small number of iterations. The same experiments repeated with other values of ρ and for varying problem size showed similar results. Note that for the Poisson regression case, even though training reconstruction error is close to 0 the other metrics are not. This is an artefact of how GLM solvers work for Poisson regression where the targets are always integer valued even when the mean parameter $exp(\mathbf{x}^\top \boldsymbol{\beta})$ is not, and where the noise is directly proportional to the mean parameter. Reconstruction error nevertheless goes close to 0 because we have additional information in the form of aggregates to estimate the training $\mathbf{y}$.



(a) Test Error for Aggregation over $\leq$ 30 clicks **(b)** Test Error for Aggregation over $\leq$ 35 clicks **(c)** (Scaled) Test Error Comparison

Figure 7.3: Real Data: Estimating CPC for Online Advertising: Error on predictions for test data plotted vs iteration for a Log-Normal Model, for aggregation period limited to 30 clicks (figure 7.3a) and 35 clicks (figure 7.3b). Errors for both cases shown in figure 7.3c, scaled for ease of comparison. Our model outperforms the baseline, leading to nearly 4-5% improvement in predictive performance

7.4.2 Real Data: CPC in Online Advertising

Recall that the digital advertising (Zeff & Aronson, 1999; Yan et al., 2009) process comprises of an advertiser who bids for ad-space provided by a publisher for a fee called cost-per-click or CPC (Hu et al., 2010) computed based on an auction-mechanism that is only partially revealed. Designing an effective and profitable bidding strategy depends on models which require the CPC value at the click-level. However, in many advertising products (e.g., the Google Shopping advertising product (NORC)) the CPC data is only available to the advertiser in aggregated form because the publishers provide only daily or hourly invoices with the charges computed over the entire set of clicks. Our task here is to design effective predictive models for estimating CPC that can work at the per-click level but use only aggregates for training.

We evaluate our methods on a real-world proprietary aggregated-CPC dataset from Criteo (criteo.com), an online advertising company that provides personalised

behaviourally retargeted advertising services for Internet retailers. As mentioned earlier, our task here is to design predictive models for estimating CPC at the per-click level, but use only the available aggregated CPC data for training. We use a subset of Criteo's advertising data collected over a period of one week in February 2017. The dataset contains 25691 instances of ad-click data for different products, and each sample consists of the the CPC aggregate corresponding to its aggregate group, as well as a feature-set of size 1551 containing information about product-country, product-price, timestamp, campaign-type, etc. The groupings over which the aggregates have been computed are known. We also use taxonomic category information for each product in the dataset, based on the Google Product Taxonomy (Google Product Taxonomy) which is widely used in the online ad-tech space.

Since the number of times an advertisement gets clicked per day can be arbitrary, the aggregation period for each aggregate has been computed over varying numbers of clicks. Ideally, we would prefer to test the performance of our methods in predicting per-click CPC, but by design this information is not available for all data points. However, ground truth itemised CPC is nevertheless available for display ads that have only been clicked once during the entire aggregation period, hence in this case the aggregate is equal to the per-click CPC. Since the number of such samples are extremely limited, we use the entire set of single-click data points as test data, so that we can evaluate the performance of our predictive model at the granularity that is required for the real world application. The remaining data, which have only aggregated CPC information, is used for training.

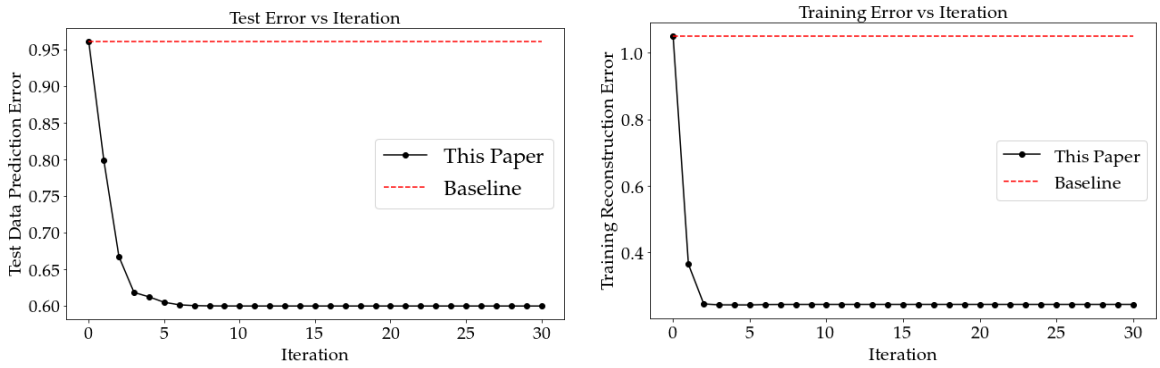
Based on common industry practice in the ad-tech domain, we use a log-

normal model (Johnson et al., 1994) as our base GLM framework for predicting CPC. We perform our experiments on two different scenarios– first where the data has been aggregated over at most 30 clicks, and similarly for data aggregated over 35 clicks. Since ground truth information is not available for training data, we only show predictive performance on test data.

Figures 7.3 shows the plots of test error versus iteration for the two sets of experiments. For reasons of stability, we use a validation set to define a maximum number of iterations for the algorithm. In both cases, we can see that our algorithm results in a improvement in predictive performance over the baseline. For ease of understanding, we provide the scaled bar chart of the final average error values in figure 7.3c. It is clear that EstImAgg results in a nearly 4-5% improvement in estimation error. To put that in context, note that the online advertising industry sees billions of dollars in transactions per year (AdWeek, March 14, 2017) and even an improvement a few percentage points can indicate significant difference in revenue.

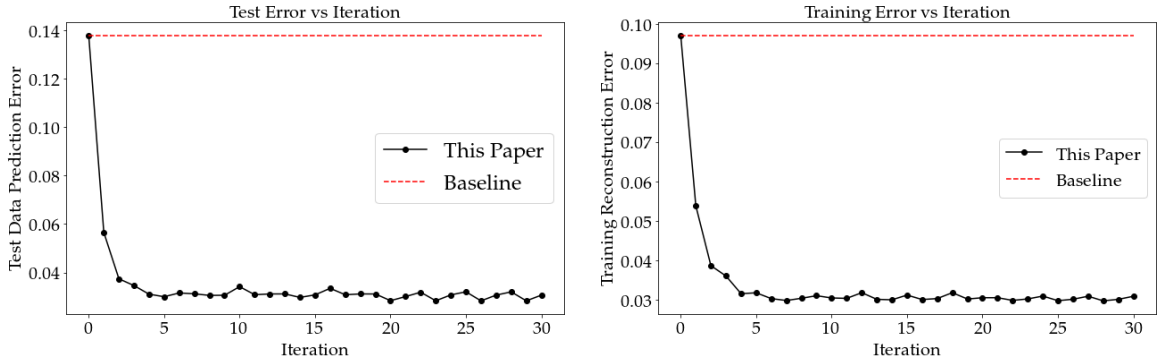
7.4.3 Real Data: Healthcare

We presented our work so far within the context of predicting click-level CPC in the online advertising domain. However, aggregated data is common in many other fields, and our framework can be applied in an identical manner to domains beyond online advertising. Healthcare is one such domain where data aggregation arises naturally– privacy concerns regarding the confidentiality of patient information limits the kinds of data that can be released to the public, and statistical disclosure techniques like aggregation is one of the most popular techniques for this purpose.



(a) Test Prediction Error **(b)** Train Reconstruction Error

Figure 7.4: Real Data: Estimation of Medicare Reimbursement Using CMS Data: Error on predictions for test data and error in reconstructed training data plotted vs iteration, as estimated using a Gaussian Model. Our model outperforms the baseline and converges within very few iterations, with a reasonably faithful reconstruction of the training data



(a) Test Prediction Error **(b)** Train Reconstruction Error

Figure 7.5: Real Data: Estimation of Texas State Hospital Charges: Error on predictions for test data and error in reconstructed training data plotted vs iteration, as estimated using a Poisson Regression Model. Our model outperforms the baseline and converges within very few iterations, with a reasonably faithful reconstruction of the training data

Patient information like healthcare charges, prevalence of pathological conditions, etc. are often released at the county or zip-code level, which can then be combined with publicly available census data to perform individual level predictive modelling.

We use two healthcare datasets to demonstrate the general applicability of our

framework– the application involved is predicting Medicare reimbursements for the first dataset, and hospital charges for the second dataset. Since we need individual-level ground truth data for testing, we only use public datasets where the target variable is available in non-aggregated form, and aggregate the training data before applying our techniques.

Our first dataset is the CMS Beneficiary Summary DE-SynPUF dataset (DESyn-PUF, 2008), which is a public use dataset created by the Centers for Medicare and Medicaid Services by applying different statistical disclosure limitation techniques to real beneficiary claims data in a way so as to very closely resemble real Medicare data. We use a subset of the DE-SynPUF dataset for Louisiana state from the year 2008 and model outpatient institutional annual primary payer reimbursement (*PPPYMT-OP*) with all the available predictor variables that include age, race, sex, duration of coverage, presence of a variety of chronic conditions, etc. Because of the nature of our target variable, we use a Poisson regression model for this problem.

Our second dataset is the Texas Inpatient Discharge dataset from the TX Department of State Health Services (TxID, 2014) that contains information about total medical charges that were paid by patients in various hospitals around Texas (see Park & Ghosh (2014) for more details on the dataset). We use hospital billing records from the fourth quarter of 2006 in the Texas Inpatient Discharge dataset and regress it on the available individual level predictor variables including binary variables race and sex, categorical variables county and zipcode, and real valued variables like length of stay. Following (Park & Ghosh, 2014), we perform a log transform on the hospital charges and length of stay before applying a Poisson regression model

(Banerjee et al., 2005).

For both the datasets, we have ground truth information for our targets, which we use to form aggregates. Just like in the synthetic data experiments, we generate aggregates by sampling the targets independently and adding them up, which is equivalent to sampling the individual entries of $\mathbf{\Gamma}$ independently according to a Bernoulli distribution. We then use these aggregates as training data and feed them to our framework along with the corresponding variables.

Similar to the synthetic data experiments, we evaluate the performance of our algorithm by computing the predictive error on the test set as well as the reconstruction error on the training data. Since we do not have access to ground truth information about the true parameter, we skip parameter recovery error in this case.

Figures 7.4 and 7.5 show the results for test data estimation error and training data reconstruction error for the DESynPUF and TxID datasets respectively. For both these datasets, the plots show that our algorithm significantly outperforms the baseline with respect to either metric. Furthermore, our framework reaches a reasonably steady-state solution fairly rapidly within a few iterations of the algorithm. Results for other similar values of ρ were similar.

7.5 Conclusion

In this chapter, we introduced a novel learning framework that can learn generalised linear models when the targets are only available as aggregates computed over arbitrary groupings of the data samples. This is an important learning paradigm

in domains ranging from online advertising to healthcare where privacy and proprietary concerns limit the release of data at a granular level. We developed a new algorithm and empirically demonstrated its efficacy in learning under aggregation constraints with experiments on both synthetic data as well as data from healthcare and from the online advertising industry.

Chapter 8

Conclusion and Future Work

8.1 Summary of the Dissertation

While aggregated data occurs in a large number of modern contexts and use-cases, there is very limited prior art on training machine learning models using data that is only available in aggregated form. In particular, problems arising from in-built structural properties like ecological fallacy make it very challenging to adapt traditional methods and techniques to the aggregated data framework. This dissertation addresses these lacunae in existing literature on the subject of using machine learning models when data is only available in aggregated form. To this end, we develop algorithmic techniques and learning frameworks for several different contexts and problem settings.

In Chapter 3, we considered the setup where covariates or $\mathbf{x}$-variables are known in full, non-aggregated form, but targets or y-variables are known only in the form of histograms or order statistics. We tackled this problem in the context of generalised linear models, and introduced a novel learning algorithm that uses alternating target imputation and parameter estimation steps to learn a predictive model given only the histogram aggregated data. As a sanity check, we performed permutation testing to assess the fidelity of our data imputation procedure, and

showed that with histograms of sufficient granularity, our reconstructed targets are close to the true values with high statistical significance. Empirical evaluation of our methods on both synthetic and real healthcare data were used to demonstrate the efficacy of our techniques. Results showed that when the histograms are constructed with fine grained binning, our performance is competitive with that of a learner who has access to the full non-aggregated data.

We next turned our attention to the case when both the covariates or features, as well as targets, were available in aggregated form. Specifically, Chapter 4 studied this setup in the case where the features and targets were available as group-wise moments. There are two sources of error in this scenario– aggregation error that is the result of using only a finite number of samples in computing the aggregates, and measurement error that arises due to noise in the data collection procedure. We considered this problem in the context of linear models and showed that under standard isometry conditions on the data matrix, and structural assumptions on the model parameter, with high probability the true model parameter can be learned exactly from just the group-wise means alone, provided the means have been computed from a sufficiently large number of data points. We extended our methods to the case of noisy measurements and showed that sparse parameter recovery is still possible up to an arbitrarily low tolerance given that aggregates have been obtained from a large enough sample size. Finally, we studied the case when the data has been aggregated into histograms, and proved that the recovery of a sparse model parameter up to a tolerance that depends on the granularity of the histograms. Experiments on synthetic data were used to corroborate our theoretical results, and empirical evaluation

on healthcare data showed the relevance of our techniques in real world applications.

Chapter 5 investigated predictive modelling in the context of spatio-temporal aggregation– a common occurrence in many areas like econometrics, finance, climate science, IoTs, etc. where data is often released as, for example, monthly or yearly averages. The chapter studied this problem in the context of linear models, where each coordinate of the feature variables, as well as the target variables, are only available as aggregates computed over non-uniform sliding windows that are not necessarily aligned with each other. We showed that by modelling the problem in the frequency domain, training becomes much more tractable. We showed that by formulating the problem in the frequency domain and exploiting duality properties of Fourier analysis, many of the inherent structural challenges of this setting can be bypassed. We introduced a novel framework and new algorithmic techniques to perform frequency domain estimation and inference for this setup and provided a theoretical analysis that showed the competitiveness of our learning algorithm with the best possible linear algorithm in terms of agnostic generalisation error. Empirical validation of our methods on synthetic data was used to reinforce our theoretical guarantees, while experiments on real data from ecological studies and climate sciences demonstrated the efficacy of our methods in practical contexts.

We then moved on to study the complementary problem in chapter 6, where the task was to learn aggregation techniques that would still allow model learning while preserving privacy and data integrity, and corresponding learning algorithms that can use these techniques. We motivate our techniques by using Gaussian regression, and subsequently extend them to the case of binary classification and gen-

eralised linear modelling. We provided both theoretical results as well as empirical evaluation for our work on real heatlthcare data and data from the telecommunications industry. Furthermore, to bridge the gap between mathematical theory and lay stakeholders in the context of privacy preservation literature, we introduces a new alternative criterion for measuring privacy, as well as showed that all our methods satisfied that criterion.

Finally. in Chapter 7 we saw a concrete example of our methods being used in a real life application. Specifically, we considered an application in the domain of online advertising, where a complex bidding process determines auctions and monetary transactions between advertisers and ad publishers, but where advertisers are handicapped in terms of training their bidding models due to not being granted access by publishers to per-click costs interred when a customer clicks on their ad. To bypass this issue, we developed an algorithm for estimating cost-per-click or CPC given only aggregated invoives collected on a daily or hourly basis. We tested our techniques on real life data from the online advertising company Criteo where they showed significant improvement over basline methods. Finally, we also tested our methods on healthcare data to show the wider applicability of our framework on domains beyond online advertising.

8.2 Future Directions of Research

The problem of learning from aggregated data arises in a wide range of contexts and application spaces, and it is not possible to encapsulate all possible settings within a single dissertation. Here we provide a few pointers towards challenges that

remain to be tackled and application areas that require further investigation into developing techniques and frameworks that allow training of individual level models given only aggregates.

First, we note that a wide range of model families can benefit from being studied in this context. In particular, kernel methods that model targets to be predicted using the relationship of their corresponding feature vectors vis-a-vis specifically selected samples in the training dataset via a kernel which acts as a proxy for similarity metrics. This dependence of kernel methods on access to individual data points makes it challenging to adapt to the aggregated data setup, since the kernel computation (or calculating the Gram matrix) explicitly requires knowledge of non-aggregated covariates. Nevertheless, there are techniques that transform the kernel training methods back into the primal domain (Rahimi & Recht, 2008, 2009), which can potentially provide a roadmap for adapting this wide class of models into this data setup.

Another interesting application involves matrix completion given averages over randomly sampled entries. The standard recommendation system setup assumes that data is available as localised estimates, that is, user ratings are available individually for every item in a subset of items relevant to the user. In many cases, however, such localised estimates are unavailable- instead, the available data is average user ratings for different groups of items. For example, the Google Now or Yahoo homepage may offer a curated set of items to a user at the beginning of a browsing session, and at the end of the activity the user may be asked to rate his experience- this user feedback would encapsulate the average rating for the entire

set of curated items rather than individually. Under low rank assumptions on the data matrix, this can be formulated as a convex optimisation problem with data constraints. However, it remains to be seen whether the existing results on matrix completion from randomly selected entries like (Candès & Recht, 2009; Recht, 2011; Recht et al., 2010) can apply to the case when measurements are only available as averages over individual entries.

More generally, a wide range of non-linear modelling paradigms offer possibilities of extension wherein they are brought into the fold of learning under aggregated data constraints. We have already provided a subset of solutions for the case of generalised linear models, but a vast range of alternative tools exist in the arsenal of existing machine learning research. The solutions to each method will be application specific and would be required to exploit structural properties of the modelling paradigm– an example is the possibility of exploring generative properties of various deep learning models (Kingma & Welling, 2013; Goodfellow et al., 2014) for data reconstruction– but regardless of approach, this would nonetheless be a compelling direction for future research.

At the other end of the spectrum, non-linear aggregation and non-uniform aggregation in general is yet another compelling application. In particular, as noted in Chapter 7, our methods for learning for linearly aggregated data led to an estimation algorithm that had an elegant structural formalism, where the data imputation step involved piece-wise linear shift to estimates obtained from the linear model at the current iteration, which mirrors the linearity in the aggregation procedure. This suggests a form of "structural regularisation", wherein the data constraints arising

from various aggregation paradigms might lead to their specific dual formalisms in the data imputation step.

There is ample scope for future research into the practical aspects of modelling with aggregated data like non-idependent data generation processes, concept drift, and wider application across problems that arise in various domains. We have already covered a selection of application areas in this dissertation, like healthcare, avertising technology, telecommunications, climate science, etc. by using data from these domains for our empirical evaluation. However, aggregated data arises in many other domains like e-commerce, econometrics, finance, political science, etc. and there exists many avenues for testing out techniques in these domains.

Finally, machine learning is used in many areas like personal finance, criminal justice system, etc. with significant potential impact on underprivileged communites and has a lot of social justice implications– techniques involved in the aggregated data context can potentially be of use in such domains. Fairness in machine learning is a particularly compelling motivation, for example in the design aggregation schemata that can either result in fair impact or be used to detect unfairness in the results of algorithms. Voting rights and gerrymandering are yet other areas where aggregation is a naturally arising structural feature. There is significant debate in political science and legal jurisprudence on defining a "good" geographic scheme for aggregating voters together into a electoral district, subject to legal, demographic and ethical constraints, and the line of research presented in this dissertation can potentially be of significant utility in making headway into this important problem.

Appendices

Appendix A

Recovery of Sparse Parameter from Group-Wise Aggregated Data: Appendix

A.1 General Remarks on the Results

As mentioned in Chapter 4, existing analyses in the sparse sensing literature are inadequate for analysing the aggregated data case, and our guarantees are much stronger than what could be achieved by a naive analysis.

The most general setup of the problem under study can be written in the following form:

$$\begin{aligned} &\text{Estimate:} && \boldsymbol{\beta}_0 \\ &\text{Given:} && \widehat{\mathbf{M}}, \widehat{Y} \\ &\text{where:} && \widehat{\mathbf{M}} = \mathbf{M} + \mathbf{e} \\ & && \widehat{Y} = \mathbf{y} + \mathbf{s} \\ & && \mathbf{y} = \mathbf{M}\boldsymbol{\beta}_0 \end{aligned} \tag{A.1}$$

There are four variations of this problem that are of interest in our setup:

1. error in design matrix $\widehat{\mathbf{M}} = \mathbf{M} + \mathbf{e}$, without noise in observation vector $\mathbf{y}$ (that is, $\mathbf{s} = 0$)

2. noise in observations $\widehat{Y} = \mathbf{y} + \mathbf{s}$, with exact design matrix $\mathbf{M}$ (that is, $\mathbf{e} = 0$)

3. design matrix error $\mathbf{e}$ and observation noise $\mathbf{s}$, where $\mathbf{e}$ and $\mathbf{s}$ are independent, $\mathbf{e} \perp\!\!\!\perp \mathbf{s}$

4. the aggregated data case (as we study in this work) which contains both design matrix error $\mathbf{e}$ and observation noise $\mathbf{s}$, and where $\mathbf{e}$ and $\mathbf{s}$ are linearly correlated

To our knowledge, all prior work in the literature (eg. Herman & Strohmer (2010); Chi et al. (2011); Rosenbaum et al. (2013); Rudelson & Zhou (2015) among others) only concern themselves with cases 1, 2 and 3. Moreover, for papers that do deal with case 2 and 3, unless $\mathbf{s} = 0$ the existing analysis will be restricted to providing only *approximate* recovery guarantees. Thus, these methods do not apply directly to case 4, a setup that almost always arises in the context of data aggregation.

We focus our investigation on the aggregated data case, that is, case 4: where $\mathbf{E}$ and $\mathbf{s}$ are linearly correlated. First of all, the existing literature does not make it clear how linearly correlated noise affects sparse parameter recovery from standard methods (like the LASSO or basis pursuit), and if the parameter can be recoverable in such cases. Even ignoring the linear correlation in the noise model, naive application of existing techniques that involve bounding error magnitudes will only be able to provide approximate recovery guarantees (where the degree of ℓ_2-approximation would depend on $\|\boldsymbol{\beta}_0\|$).

The key observation that allows us to bypass all these limitations is the fact that while $\mathbf{E}$ and $\mathbf{s}$ are correlated, we have one more piece of the puzzle that can be used to augment the information in equation A.1: the fact that not only are $\mathbf{E}$ and $\mathbf{s}$ linearly correlated, they are tied together via the true parameter $\boldsymbol{\beta}_0$ in the form of

the expression $\mathbf{s} = \mathbf{E}\boldsymbol{\beta}_0$. This is an artefact of the natural structure that is generated by data aggregation in linear models.

This observation is key to bypassing the problems in parameter recovery outlined earlier. Indeed, we show that not only can we guarantee parameter recovery using standard compressed sensing algorithms, we can also guarantee *exact* parameter recovery, as we see in Theorem 4.3.1, and recovery upto arbitrarily accurate degree of estimation as we see in 4.3.2 and 4.3.3. These results, while seemingly intuitive after the fact, have not been shown in either the compressed sensing literature, or in the literature on ecological estimation dating back at least 60 years to Goodman (1953), and to our knowledge, ours is the first work that examines and gives guarantees for the structured parameter recovery problem in the context of aggregated data.

Furthermore, as we mention in Chapter 4, our analysis techniques generalise beyond the exact problem setup and estimation procedure that we present in this work, and can be easily extended to analyse sparse or approximately sparse parameter recovery from aggregated data in a wide variety of contexts (non-sparse $\boldsymbol{\beta}_0$, beyond sub-Gaussian assumptions, etc. (see for example Candes et al. (2006); Cai et al. (2009)) and using various kinds of estimators beyond the LASSO or basis pursuit (for example the Dantzig selector, Matrix Uncertainty-selector, etc., (see Candes & Tao (2007); Rosenbaum et al. (2013)). While the sample complexity required may vary a little from case to case, our main results, on exact parameter recovery or recovery to within any arbitrary degree of approximation, would remain the same.

A.2 Proofs of Main Results

Note that the analysis presented below is one out of many possible approaches. Slightly different bounds can be achieved using different methods of analysis, for example using the Bauer-Fike Theorem, Weyl's Inequality, Wielandt Hoffman theorem, etc. and the bounds derived below can be made tighter by making further assumptions on the distributions of covariates or noise terms, etc.

The main property that enables recovery of sparse parameters from an underdetermined linear system is the restricted isometry condition, also sometimes known as the Uniform Uncertainty Principle.

For the matrix $\boldsymbol{M} \in \mathbb{R}^{k \times d}$ and any set $T \subseteq \{1, 2, \cdots, d\}$, suppose $\boldsymbol{M}_T$ is the $k \times |T|$ matrix consisting of the columns of $\boldsymbol{M}$ corresponding to T. Then, the *s-restricted isometry constant* δ_s of the matrix $\boldsymbol{M}$ is defined as the smallest quantity such that the matrix $\boldsymbol{M}_T$ obeys

$$(1 - \delta_s)\|c\|_2^2 \leq \|\boldsymbol{M}_T c\|_2^2 \leq (1 + \delta_s)\|c\|_2^2$$

for every subset $T \subset \{1, 2, \cdots, d\}$ of size $|T| < s$ and all real $c \in \mathbb{R}^{|T|}$

As in Chapter 4, we assume that $\mathbf{M}$ satisfies the restricted isometry hypotheses for both exact recovery and noisy recovery. That is, there exists an s_0 such that the following conditions are satisfied with respect to the $2s_0$-restricted isometry constants δ_{2s_0} for $\mathbf{M}$ in the manner as defined below:

1. For exact recovery from noise-free measurements, we assume $\delta_{2s_0} < \Theta_0 = \frac{3}{4+\sqrt{6}} \approx 0.465$

2. For approximate recovery from noisy measurements, we assume $\delta_{2s_0} < \Theta_1 = \sqrt{2} - 1 \approx 0.414$

However, we do not know the true mean matrix $\mathbf{M}$, only the sample mean matrix $\widehat{\mathbf{M}}_n = \mathbf{M} + \mathbf{E}_n$, where $\mathbf{E}_n$ is the matrix of aggregation error owing to empirical estimation from a finite number of samples. We now show that when the true mean matrix $\mathbf{M}$ satisfies the restricted isometry conditions, given enough samples n so will the sample mean matrix $\widehat{\mathbf{M}}_n$ with high probability.

We first show the following result for the isometry constants for $\widehat{\mathbf{M}}_n = \mathbf{M} + \mathbf{E}_n$ in terms of the eigenvalues of $\mathbf{E}_n$.

Lemma A.2.1. *Let δ_s be the s-restricted isometry constant for $\mathbf{M}$. Let $\sqrt{\lambda_n}$ denote the absolute value of the largest (in absolute value) singular value of $E_{n,T}$ for all subsets $T \subset \{1, 2, \cdots d\}$. Then, $\zeta_s = (\delta_s + \lambda_n + 2\sqrt{\lambda_n(1-\delta_s)})$ is such that for every subset $T \subset \{1, 2, \cdots, d\}$ of size $|T| < s$ and all real $c \in \mathbb{R}^{|T|}$*

$$(1 - \zeta_s)\|c\|_2^2 \leq \|(\mathbf{M}_T + E_{n,T})c\|_2^2 \leq (1 + \zeta_s)\|c\|_2^2 \tag{A.2}$$

Proof. For every subset $T \subset \{1, 2, \cdots, d\}$ and all real $c \in \mathbb{R}^{|T|}$ we have by triangle inequality,

$$\|(\mathbf{M}_T + E_{n,T})c\| \leq \|\mathbf{M}_T c\| + \|E_{n,T}c\| \leq (\sqrt{1+\delta_s} + \sqrt{\lambda_n})\|c\|$$

Also,

$$\begin{aligned}\sqrt{(1-\delta_s)}\|c\| &\leq \|\mathbf{M}_T c\| \\ &= \|\left((\mathbf{M}_T + E_{n,T}) - E_{n,T}\right) c\| \\ &\leq \|(\mathbf{M}_T + E_{n,T})c\| + \|E_{n,T}c\| \\ &\leq \|(\mathbf{M}_T + E_{n,T})c\| + \sqrt{\lambda_n}\|c\|\end{aligned}$$

Therefore, we have

$$\begin{aligned}(\sqrt{1-\delta_s} - \sqrt{\lambda_n})\|c\| &\leq \|(\mathbf{M}_T + E_{n,T})c\| \\ &\leq (\sqrt{1+\delta_s} + \sqrt{\lambda_n})\|c\|\end{aligned}$$

Assume[1] $\lambda_n < (1+\delta_s)$, and $\zeta_s = (\delta_s + \lambda_n + 2\sqrt{\lambda_n(1+\delta_s)}) < 1$, then we have

$$\sqrt{(1-\zeta_s)} \leq \sqrt{1-\delta_s} - \sqrt{\lambda_n}$$

and

$$\sqrt{(1+\zeta_s)} = \sqrt{1+\delta_s} + \sqrt{\lambda_n}$$

This completes the proof. □

We now bound the singular values of $E_{n,T}$.

Lemma A.2.2. *Let $\sqrt{\lambda_n}$ denote the absolute value of the largest (in absolute value) singular value of $E_{n,T}$ for any $T \subset \{1, 2, 3, \cdots d\}$. Then*

$$\lambda_n \leq \|\mathbf{E}_n\|_F^2$$

where $\|\cdot\|_F$ denotes the Frobenius norm.

[1] We shall prove later that with overwhelmingly high probability $\lambda_n < \left(\frac{(\Theta-\delta_s)^2}{9(1+\delta_s)}\right)$ where $\Theta < 1$. This subsumes both the assumptions stated here.

Proof. Let $\sqrt{\lambda_\tau^{(n,T)}}$ for $\tau = 1, 2, \cdots$ be absolute values of the non-zero singular values of $E_{n,T}$. Consider the singular value decomposition of $E_{n,T} = U\Lambda V^\top$. Then

$$\begin{aligned} \|E_{n,T}\|_F^2 &= Trace(E_{n,T}^\top E_{n,T}) = Trace(\Lambda^\top \Lambda) \\ &= \sum_\tau \lambda_\tau \geq \max_\tau \lambda_\tau^{(n,T)} \end{aligned}$$

Therefore, for every T we have

$$\max_\tau \lambda_\tau^{(n,T)} \leq \|E_{n,T}\|_F^2 \leq \|\mathbf{E}_n\|_F^2$$

Since $\lambda_n = \max_T \max_\tau \lambda_\tau^{(n,T)}$, we have the result. □

This is just one approach, similar results can also be obtained, for example, by bounding the eigenvalues using the Gershgorin Circle Theorem.

Finally we show that with high probability λ_n can be bounded.

Lemma A.2.3. *Suppose each covariate has a sub-Gaussian distribution with parameter σ^2, that is, for each covariate $x_{j,i} \in \mathbf{x}_j = [x_{j,1}, x_{j,2} \cdots x_{j,d}]$ and each group $j \in \{1, 2, \cdots k\}$, we have for every $t \in \mathbb{R}$, the logarithm of the moment generating function is quadratically bounded*

$$lnE[e^{t(x_{j,i}-\mu_{j,i})}] < \frac{t^2\sigma^2}{2}$$

Then, for any positive $\theta > 0$, the probability $P(\lambda_n > \theta) < 2kd\ e^{-n\theta/2kd\sigma^2}$

Proof. Note that the $(j,i)^{th}$ element of the matrix $\mathbf{E}_n$ is the zero random variable $E_{n,(j,i)} = \frac{\sum_{m=1}^n (x_{j,i}^{(m)} - \mu_{ji})}{n}$, where $x_{j,i}^{(m)}$ is the m^{th} observation of the i^{th} covariate in the j^{th} group, and μ_{ji} is the mean of the i^{th} covariate in the j^{th} group.

Since each covariate has a sub-Gaussian distribution with parameter σ^2, we have, by Hoeffding's inequality for sub-Gaussian random variables, for any $\theta > 0$

$$\begin{aligned} P\left(|E_{n,(ij)}| > \sqrt{\theta}\right) &= P\left(|\frac{\sum_{m=1}^{n}(x_{j,i}^{(m)} - \mu_{ji})}{n}| > \sqrt{\theta}\right) \\ &< 2e^{-n\theta/2\sigma^2} \end{aligned}$$

Therefore, using Lemma A.2.2, we have

$$\begin{aligned} P(\lambda_n > \theta) &\leq P(\|\mathbf{E}_n\|_F^2 > \theta) \\ &\leq \sum_{ij} P(E_{n,(ij)}^2 > \frac{\theta}{kd}) \\ &\leq \sum_{ij} 2e^{-n\theta/2kd\sigma^2} \\ &= 2kd\ e^{-n\theta/2kd\sigma^2} \end{aligned}$$

where the second inequality is by union bound and the third is due to Hoeffding's inequality. □

We are now in a position to prove the main results.

A.2.1 Proof of Theorem 4.3.1

Proof. We saw in Lemma A.2.1 that is the s-restricted isometry constants for $\mathbf{M}$ are δ_s, then the corresponding s-restricted isometry constants for $\widehat{\mathbf{M}}_n$ are

$$\zeta_s = \delta_s + \lambda_n + 2\sqrt{\lambda_n(1+\delta_s)} < \delta_s + 3\sqrt{\lambda_n(1+\delta_s)}$$

for small enough λ_n

Let $\Theta_0 = \frac{3}{4+\sqrt{6}} \approx 0.465$. Suppose there exists an s_0 such that the isometry constant δ_{2s_0} for the true mean matrix $\mathbf{M}$ satisfy $\delta_{2s_0} < \Theta_0$. Using Theorem 4.2.1 (Foucart, 2010), we can see that any κ_0 sparse $\boldsymbol{\beta}_0$ can be recovered from $\widehat{\mathbf{M}}_n$ if the corresponding isometry constants for $\widehat{\mathbf{M}}_n$ satisfy $\zeta_{2s_0} < \Theta_0$, that is

$$\begin{aligned} & \zeta_{2s_0} && < \Theta_0 \\ \equiv\ & \zeta_{2s_0} - \delta_{2s0} && < \Theta_0 - \delta_{2s_0} \\ \Leftarrow\ & 3\sqrt{\lambda_n(1+\delta_{2s_0})} && < \Theta_0 - \delta_{2s_0} \\ \equiv\ & \lambda_n && < \vartheta_{s_0} \end{aligned} \tag{A.3}$$

where

$$\vartheta_{s_0} = \left(\frac{(\Theta_0 - \delta_{2s_0})^2}{9(1+\delta_{2s_0})} \right)$$

All that is left to show is that the condition $\zeta_{2s_0} < \Theta_0$ is true with high probability. This is straightforward by using Lemma A.2.3 and the results in equations (2) above. We have,

$$\begin{aligned} P(\zeta_{2s_0} < \Theta_0) &> P(\lambda_n < \vartheta_{s_0}) \\ &= 1 - P(\lambda_n > \vartheta_{s_0}) \\ &\geq 1 - e^{-C_0 n} \qquad \text{by Lemma A.2.3} \end{aligned}$$

where the constant C_0 is such that

$$C_0 = O\left(\frac{\vartheta_0}{kd\sigma^2} \right) = O\left(\frac{(\Theta_0 - \delta_{2s_0})^2}{kd\sigma^2(1+\delta_{2s_0})} \right)$$

□

A.2.2 Proof of Theorem 4.3.2

Proof. Using Theorem 4.2.2 (Candes, 2008), recovery of $\boldsymbol{\beta}_0$ within an $O(\xi)$ distance is possible if the restricted isometry constants for $\widehat{\mathbf{M}}_n$ satisfy $\zeta_{2s_0} < \Theta_1$ where $\Theta_1 = \sqrt{2} - 1 \approx 0.414$, and the error term $\boldsymbol{\epsilon}_n$ is bounded as $\|\boldsymbol{\epsilon}_n\|_2 < \xi$. For succinctness, we drop the subscript from the error term and denote $\boldsymbol{\epsilon}_n$ simply as $\boldsymbol{\epsilon}$.

The probability of the restricted isometry condition being violated for the sample means can be bounded in a manner similar to the proof of theorem 4.3.1 as

$$P(\zeta_{2s_0} > \Theta_1) \leq e^{-C_1 n}$$

where $C_1 \sim O\left(\frac{(\Theta_1 - \delta_{2s_0})^2}{kd\sigma^2(1+\delta_{2s_0})}\right)$. The probability of the error being too large can be bounded in a similar fashion by using Hoeffding's inequality as

$$\begin{aligned}
P(\|\boldsymbol{\epsilon}\|_2 > \xi) &= P(\sum_{j=1}^{k} \epsilon_j^2 > \xi^2) \\
&\leq \sum_{j=1}^{k} P(\epsilon_j^2 > \frac{\xi^2}{k}) \\
&= \sum_{j=1}^{k} P(|\epsilon_j| > \frac{\xi}{\sqrt{k}}) \\
&\leq \sum_{j=1}^{k} 2\ e^{-n\xi^2/2\rho^2 k} \\
&= 2k\ e^{-n\xi^2/2\rho^2 k}
\end{aligned}$$

where the first inequality is by union bound and the second inequality is due to Hoeffding's inequality.

Therefore the probability of recovery within $O(\xi)$ is bounded below by

$$1 - P(\zeta_{2s_0} > \Theta_1) - P(\|\boldsymbol{\epsilon}\|_2 > \xi) = 1 - e^{-C_1 n} - e^{-C_2 n}$$

where $C_1 \sim O\left(\frac{(\Theta_1 - \delta_{2s_0})^2}{kd\sigma^2(1+\delta_{2s_0})}\right)$ and $C_2 \sim O\left(\frac{\xi^2}{\rho^2 k}\right)$ □

As mentioned earlier, there are multiple other approaches for special cases and using alternative conditions for successful recovery of sparse or nearly sparse vectors from under-determined linear systems, see for instance Candes & Tao (2007), Candes & Plan (2011), Cai et al. (2010b), Cai et al. (2010a), Cai et al. (2009), etc. The analysis with alternative assumptions follows along the same lines as that presented in our work.

A.2.3 Proof of Theorem 4.3.3

Proof. Note that the observations where the target mean is estimated from aggregated data as $\widehat{Y}_\Delta = \widehat{Y}_n + h_\Delta$ can be considered noisy observations of the type $\widehat{\mathbf{M}}_n \boldsymbol{\beta}_0 = \boldsymbol{\upsilon}_\Delta - h_\Delta$. Therefore, using Theorem 2.2, recovery of $\boldsymbol{\beta}_0$ within an $O(\xi_\Delta)$ distance is possible if the restricted isometry constants for $\widehat{\mathbf{M}}_n$ satisfy $\zeta_{2s_0} < \Theta_1$ and the the error term h_Δ is bounded as $\|h_\Delta\|_2 < \xi_\Delta$. The probability of the restricted isometry hypothesis being violated is

$$P(\zeta_{2s_0} > \Theta_1) \leq e^{-C_1 n}$$

where $C_1 \sim O\left(\frac{(\Theta_1 - \delta_{2s_0})^2}{kd\sigma^2(1+\delta_{2s_0})}\right)$. This part is exactly identical to the proof of Theorem 4.3.2.

The bound on the error in estimation of target means can be done in a deterministic manner as follows.

The mean estimation procedure from the histogram is exact if the targets in each bin are distributed symmetrically around the mid point of each bin. Note that since each target is at a maximum distance of $\frac{\Delta}{2}$ from the mid point of their corresponding bin, by setting every target to the mid point of the bin we incur at most an error of $\frac{\Delta}{2}$ for each target. Therefore, the maximum possible error in estimating the sample mean in each group is

$$|\hat{\nu}_n - \hat{\nu}_\Delta| < \frac{\Delta}{2}$$

And hence, the error term h_Δ is bounded in ℓ_2 as

$$\|h_\Delta\|_2 < \sqrt{k}\frac{\Delta}{2}$$

This is of course a loose bound which assumed a worst-case pathological condition. Better bounds on the recovery error can be obtained by appropriate regularity assumptions on the distribution of the targets. □

A.3 Higher Order Moments

Consider the τ^{th} order moments under a linear function

$$\rho_\tau = E[y^\tau] = E[(\mathbf{x}^\top \boldsymbol{\beta})^\tau], \quad \tau = 1, 2, 3, \cdots \tag{A.4}$$

If all moments of the covariates are known, that is, $\{E[\Pi_j x_j^{a_j}] : a_j \in \mathbb{Z}_+, \sum_j a_j = \tau\}$ is known, then the right hand side of (A.4) is a scalar valued (shifted) homogeneous polynomial function in $\boldsymbol{\beta}$ of degree τ. Therefore, (A.4) is essentially a set of

multivariate polynomial equations in $\boldsymbol{\beta} = [\beta_1, \beta_2, \cdots, \beta_d]$. First consider whether the problem is well-defined, that is, whether the system of equations (A.4) has a unique solution. There is a considerable amount of literature in computational algebraic geometry that deals with the determination of whether a system of multivariate polynomial equations has at least one solution or is inconsistent (using, for instance, techniques and results in Adams & Loustaunau (1994); Ruiz (1985)). In our case, this question is moot since we assume that the data is generated according to a linear model and therefore, there exists at least one solution. Unfortunately, testing for uniqueness of solution is a much harder problem.

As a base case, consider only using the first two moments. This is a widely applicable case since for many commonly used distribution choices for $\mathcal{P}_x$ like Multivariate Gaussian, Poisson, etc. the first two moments completely characterise the entire distribution.

The equations (A.4) can now be written comprising of a set of linear and a set of quadratic equations. The linear system of equations involving first order moments from each population sub-group $j \in \{1, 2, \cdots k\}$ is as follows:

$$E[\mathbf{x}_{(j)}^\top \boldsymbol{\beta}] = E[y_{(j)}] \Leftrightarrow \boldsymbol{\mu}_j^\top \boldsymbol{\beta} = \nu_j \quad j = 1, 2, 3, \cdots k \tag{A.5}$$

Similarly, the set of quadratic equations involving second order moments from each population-subgroup $j \in \{1, 2, \cdots k\}$ can be written as follows:

$$E[\boldsymbol{\beta}^\top (\mathbf{x}\mathbf{x}_{(j)}^\top) \boldsymbol{\beta}] = E[y_{(j)}^2] \Leftrightarrow \boldsymbol{\beta}^\top \Sigma_j \boldsymbol{\beta} = \sigma_j^2 \quad j = 1, \cdots, k \tag{A.6}$$

where Σ_j and σ_j^2 are the covariance of $\mathbf{x}$ and variance of y corresponding to the j^{th} population subgroup.

Geometrically, (A.5) and (A.6) represent in terms of $\boldsymbol{\beta}$ a set of k hyperplanes and a set of k ellipsoids centred at the origin in $\mathbb{R}^d$ space. The problem has a unique solution if the set of hyperplanes and the set of ellipsoids have a single point of intersection.

Counting the number of points of intersection of polynomials in real space is a difficult problem in the general case. It is usually studied for the complex space $\mathbb{C}^d$ under the umbrella of enumerative geometry (Katz, 2006). As earlier, if $k \geq d$ and under the assumption that at least one solution exists (the system is consistent), the set of hyperplanes is sufficient to recover the true $\boldsymbol{\beta}_0$. We would ideally like to see if knowledge of second order moments can reduce the number of population subgroups k required for a unique solution, or aids the estimation process in any other way.

Let Σ be some covariance matrix and $U\Delta_S U^\top$ be its singular value decomposition, where U is an orthonormal matrix and $\Delta_S = diag(S)$ is a diagonal matrix of loadings $S = [s_1, s_2, \cdots s_d] \succeq \mathbf{0}$. Let $\sigma^2 \in \mathbb{R}_+$ be any positive real value. Then for a given $\boldsymbol{\beta}$ to satisfy the second order moment constraint

$$\boldsymbol{\beta}^\top \Sigma \boldsymbol{\beta} = \sigma^2 \tag{A.7}$$

means that the ellipsoid Σ in $\mathbb{R}^d$ centred at the origin with axes defined by U and of size (S, σ^2) passes through $\boldsymbol{\beta}$.

We now show that in the general case, knowledge about second order moments do not help.

Proposition A.3.1. *Suppose $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are two points in $\mathbb{R}^d$ such that the origin, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are not collinear. For any arbitrary $\sigma^2 > 0$ and any arbitrary choice of*

axes U, the set of loadings S for which both $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ satisfy equation (A.7) with $\Sigma = U\Delta_S U^\top$ and $\Delta_S = diag(S)$ is given by the intersection of a $(d-2)$-dimensional vector space with the positive orthant.

Before we prove this, let us unpack this result. The essential idea is that, barring non-degenerate cases like $S = \mathbf{0}$, and for $d > 2$, a $(d-2)$ dimensional vector space intersects the positive orthant in an infinite number of points, assuming they do intersect. Therefore, for any two points in $\mathbb{R}^d$, there exist an infinite number of ellipsoids for every given size σ^2 and axes U which passes through both the points.

The implications of the above result are the following. Suppose we place constraints on $\boldsymbol{\beta}$ to constrain it to some set $\mathcal{C}$. Then if $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are any two points in $\mathcal{C}$, we can easily find any number of arbitrary second order moment conditions that are satisfied by both $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$. Therefore, estimation with information about second order moments from k groups for any $k < \infty$ cannot be guaranteed to be any better than estimation without second order moments in the general case.

Furthermore, since the result holds for arbitrary values of σ^2 and U, it also implies that many types of common assumptions like sparsity or norm constraints on $\boldsymbol{\beta}$, rank constraints on the covariances Σ_k, etc. are insufficient in general to make the parameter recovery problem well defined with second order moments alone. Similar results can potentially be obtained for higher order moments by noting that a set of higher order polynomial equations can be converted into polynomial equations of degree $\tau \leq 2$ by introducing auxiliary variables.

Proof. Let $\Sigma = U\Delta_S U^\top$ where U is a unitary matrix and $\Delta_S = diag(S) = diag(s_1, s_2, \cdots, s_d)$

is a diagonal matrix. Let $\boldsymbol{\beta}_1, \boldsymbol{\beta}_2 \in \mathbb{R}^d$ be any two arbitrary points. Take the projections of each $\boldsymbol{\beta}_i$ on the axes defined by the j^{th} column $\boldsymbol{u}_j$ of U for each j. Let $\lambda_{j,2} = (\boldsymbol{\beta}_1^\top \boldsymbol{u}_j)^2$ and $\lambda_{j,2} = (\boldsymbol{\beta}_2^\top \boldsymbol{u}_j)^2$ be the corresponding squared projections of the two points $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ on each axis $\boldsymbol{u}_j$ for $j = 1, 2, 3, \cdots d$.

Concatenate the projections into the matrix $\boldsymbol{\Lambda} = [\boldsymbol{\Lambda}_1; \boldsymbol{\Lambda}_2]^\top \in \mathbb{R}^{2\times d}$ where $\boldsymbol{\Lambda_1} = [\lambda_{1,1}, \lambda_{2,1}, \cdots \lambda_{d,1}]^\top$ and $\boldsymbol{\Lambda_2} = [\lambda_{1,2}, \lambda_{2,2}, \cdots \lambda_{d,2}]^\top$.

It is easy to verify that

$$\begin{aligned} \boldsymbol{\beta}_1^\top \Sigma \boldsymbol{\beta}_1 &= \boldsymbol{\Lambda}_1^\top S \\ \boldsymbol{\beta}_2^\top \Sigma \boldsymbol{\beta}_2 &= \boldsymbol{\Lambda}_2^\top S \end{aligned}$$

Therefore, $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ will both satisfy the second moment equation (A.7) for any ellipsoid defined by $(\Sigma = U\Delta_S U^\top, \sigma^2)$ if

$$\boldsymbol{\Lambda}^\top S = [\sigma^2; \sigma^2] \tag{A.8}$$

$$S \succeq \mathbf{0} \tag{A.9}$$

In terms of S, this represents an intersection of a $d-2$ dimensional vector space $\boldsymbol{\Lambda}^\top S = [\sigma^2; \sigma^2]$ with the positive orthant $S \succeq \mathbf{0}$ which is satisfied by an infinite number of solutions in terms of S. □

Note that $\boldsymbol{\Lambda}^\top S = [\sigma^2; \sigma^2]$ is inconsistent if $\boldsymbol{\beta}_1$ and $\boldsymbol{\beta}_2$ are collinear with the origin, that is, $\boldsymbol{\beta}_1 = \eta\boldsymbol{\beta}_2$ for some η with $|\eta| \neq 1$. If $\boldsymbol{\beta}_1 = \pm\boldsymbol{\beta}_2$, then if one satisfies the ellipsoid constraint, the other trivially satisfies it as well.

Appendix B

Frequency Domain Predictive Modelling with Spatio-Temporally Aggregated Data: Appendix

B.1 Frequency Domain Formulation

Consider our original loss function

$$\mathcal{L}(\boldsymbol{\beta}) = E[|\mathbf{x}(t)^\top \boldsymbol{\beta} - y(t)|^2]$$

As earlier, denote the residue term[1] at $\boldsymbol{\beta}$ as $\varepsilon_\beta(t) = \mathbf{x}(t)^\top \boldsymbol{\beta} - y(t)$, therefore our loss function can be written as

$$\mathcal{L}(\boldsymbol{\beta}) = E[\varepsilon_\beta(t)^2]$$

Suppose $P_{\varepsilon_\beta}(\omega)$ is the power spectral density of the residue term $\varepsilon_\beta(t)$. Then, we have

$$\mathcal{L}(\boldsymbol{\beta}) = E[\varepsilon_\beta(t)^2] = \int_{-\infty}^{\infty} P_{\varepsilon_\beta}(\omega) d\omega \tag{B.1}$$

As mentioned previously, we assume that $P_{\varepsilon_\beta}(\omega)$ decays rapidly with ω and almost vanishes beyond a certain $|\omega| > \omega_0$ (see section B.3 for an extended discussion on this). Therefore, the integral on the right hand side can be approximated by a

[1]Note that the residue process $\varepsilon_\beta(t)$ is equal to the error process $\epsilon(t)$ at $\boldsymbol{\beta} = \boldsymbol{\beta}^*$, where $\boldsymbol{\beta}^*$ is the true parameter

finite integral as

$$\int_{-\infty}^{\infty} P_{\varepsilon_\beta}(\omega)d\omega \approx \int_{-\omega_0}^{\omega_0} P_{\varepsilon_\beta}(\omega)d\omega$$

for a suitable ω_0.

Next, because we assume that $P_{\varepsilon_\beta}(\omega)$ exists finitely for every ω, the integral on the right hand side above can be approximated by averaging the readings of $P_{\varepsilon_\beta}(\omega)$ over a finite set of frequencies $\Omega = \{\omega_1, \omega_2, \cdots, \omega_M\}$ as

$$\int_{-\omega_0}^{\omega_0} P_{\varepsilon_\beta}(\omega)d\omega \approx \frac{1}{|\Omega|}\sum_{\omega\in\Omega} P_{\varepsilon_\beta}(\omega)$$

for a suitable Ω.

Finally, recall the definition of power spectral density

$$P_{\varepsilon_\beta}(\omega) = \lim_{T\uparrow\infty} E\left[\frac{\|\int_{-T}^{T} \varepsilon_\beta(t)e^{-\iota\omega t}dt\|^2}{2T}\right]$$

Again, because $P_{\varepsilon_\beta}(\omega)$ is assumed to exist finitely for every $\omega \in \Omega$, for a high enough T_0, the limit on the right hand side can be replaced by the value of the function at $T = T_0$

$$\lim_{T\uparrow\infty} E\left[\frac{\|\int_{-T}^{T} \varepsilon_\beta(t)e^{-\iota\omega t}dt\|^2}{2T}\right] \approx \frac{1}{2T_0}E\left[\|\boldsymbol{\epsilon}_{\beta,T_0}(\omega)\|^2\right]$$

where $\boldsymbol{\epsilon}_{\beta,T_0}(\omega) = \mathbf{X}_{T_0}(\omega)\boldsymbol{\beta} - Y_{T_0}(\omega)$ is the T_0 restricted finite Fourier Transform of the residue at $\boldsymbol{\beta}$.

To summarize, the preceding discussion outlines the path by which our original loss function

$$\mathcal{L}(\boldsymbol{\beta}) = E[|\mathbf{x}(t)^\top \boldsymbol{\beta} - y(t)|^2]$$

can be substituted by an approximate frequency domain equivalent

$$\hat{\mathcal{L}}(\boldsymbol{\beta}) = \frac{1}{2T_0|\Omega|} \sum_{\omega \in \Omega} E\left[\|\mathbf{X}_{T_0}(\omega)\boldsymbol{\beta} - Y_{T_0}(\omega)\|^2\right]$$

Since, the minimizer of an optimisation problem is invariant to positive scalar multiplication of the objective function, we use as our estimator

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{\omega \in \Omega} E\left[\|\mathbf{X}_{T_0}(\omega)\boldsymbol{\beta} - Y_{T_0}(\omega)\|^2\right]$$

B.2 Proofs of Main Results

We shall now make these ideas more concrete. We recall the main aspects of our setup below.

(i). We work with a parametric linear model, where the target variable $y(t)$ is regressod on predictor variables $\mathbf{x}(t)$ via a fixed parameter vector $\boldsymbol{\beta}^*$ as

$$y(t) = \mathbf{x}(t)^\top \boldsymbol{\beta}^* + \epsilon(t)$$

for each t.

The partial Fourier Transforms for our signals are

$$\mathbf{X}_T(\omega) = \int_{-T}^{T} \mathbf{x}(t) e^{-\iota\omega t} dt$$
$$Y_T(\omega) = \int_{-T}^{T} y(t) e^{-\iota\omega t} dt$$

(ii). We assume that each of our signals are weakly stationary stochastic processes with mean zero, and rapidly decaying autocorrelation function $\rho_{(\cdot)}(\tau)$ and finite variance $\rho_{(\cdot)}(0)$. In particular, this implies that $\varepsilon_\beta(t)$ is also centered and weakly stationary with rapidly decaying autocovariance function

We also assume finite power spectral density for all our signals, that is, we assume that

$$P_z(\omega) = \lim_{T\uparrow\infty} E\left[\frac{\|Z_T(\omega)\|^2}{2T}\right] \tag{B.2}$$

$$= \lim_{T\uparrow\infty} E\left[\frac{\|\int_{-T}^{T} z(t)e^{-\iota\omega t}dt\|^2}{2T}\right] \tag{B.3}$$

$$= \int_{-\infty}^{\infty} \rho_z(\tau)e^{-\iota\omega(\tau)}d\tau \tag{B.4}$$

is finite for every ω, and finitely integrable over $\omega \in (-\infty, \infty)$. It follows from these assumptions that the PSD will also be finite for the residue process $\varepsilon_\beta(t)$.

(iii). By linearity of Fourier transform, we have

$$Y_T(\omega) = \mathbf{X}_T(\omega)^\top \boldsymbol{\beta} + \boldsymbol{\epsilon}_T(\omega)$$

for any $T, \omega, \boldsymbol{\beta}$.

(iv). We define the optimal parameter $\boldsymbol{\beta}^*$ as the one that minimises the generalisation error, that is,

$$\boldsymbol{\beta}^* = \arg\min_{\boldsymbol{\beta}} \quad E\left[|\mathbf{x}(t)^\top \boldsymbol{\beta} - y(t)|^2\right] \tag{B.5}$$

We estimate our parameter in the frequency domain instead, as

$$\widehat{\boldsymbol{\beta}} = \arg\min_{\boldsymbol{\beta}} \sum_{\omega \in \Omega} E\left[\|\widehat{\mathbf{X}}_{T_0}(\omega)^\top \boldsymbol{\beta} - \widehat{Y}_{T_0}(\omega)\|^2\right]$$

for fixed parameters ω_0, T_0 and a set $\Omega = \{-\omega_0 < \omega_i < \omega_0 : i = 1, 2, \cdots |\Omega|\}$ of real valued "frequencies" sampled uniformly between $\omega \in (-\omega_0, \omega_0)$. Let $|\Omega| = D$. Also, define

$$\hat{\mathcal{L}}(\boldsymbol{\beta}) = \frac{1}{T_0|\Omega|} \sum_{\omega \in \Omega} E\left[\|\widehat{\mathbf{X}}_{T_0}(\omega)^\top \boldsymbol{\beta} - \widehat{Y}_{T_0}(\omega)\|^2\right]$$

We now prove some results that will be necessary in deriving our main theorems.

Lemma B.2.1. *There exists an $0 < \xi_{\omega_0} < 1$ for every ω_0 (conversely, for every $\xi_{\omega_0} \in (0, 1)$, there exists ω_0) such that*

$$\begin{aligned}(1 - \xi_{\omega_0}) E\left[|\mathbf{x}(t)^\top \boldsymbol{\beta} - y(t)|^2\right] &\leq \int_{-\omega_0}^{\omega_0} P_{\varepsilon_\beta}(\omega) d\omega \\ &\leq E\left[|\mathbf{x}(t)^\top \boldsymbol{\beta} - y(t)|^2\right]\end{aligned} \tag{B.6}$$

Proof. We use the following standard result. For any weakly stationary signal $\mathbf{z}$, we have

$$E\left[|z(t)|^2\right] = \int_{-\infty}^{\infty} P_z(\omega) d\omega \tag{B.7}$$

By equation (B.7), we have

$$E\left[|\mathbf{x}(t)^\top \boldsymbol{\beta} - y(t)|^2\right] = E[|\varepsilon_\beta(t)|^2] = \int_{-\infty}^{\infty} P_{\varepsilon_\beta}(\omega) d\omega$$

Also, by assumption, $P_{\varepsilon_\beta}(\omega)$ is finite for every ω, and finitely integrable over $(-\infty, \infty)$. Moreover, by definition of power spectral density, $P_{\varepsilon_\beta}(\omega) \geq 0$ for each ω. Hence, the result. □

The constant ξ_{ω_0} depends on the exact functional form of P_{ε_β}, or equivalently, of ρ. Standard rates can be obtained by using the fact that $\frac{P_{\varepsilon_\beta}(\omega)}{\int_{-\infty}^{\infty} P_{\varepsilon_\beta}(\omega)d\omega}$ is a valid probability density function, and using the tail probability results for the corresponding probability distribution.

For example, if ρ exhibits a Gaussian decay (analogous to normal distribution), that is, $\rho(\tau) \sim exp(-O(\tau^2))$, then P_{ε_β} also exhibits a Gaussian decay, that is $P_{\varepsilon_\beta}(\omega) \sim exp(-O(\omega^2))$, and therefore, $\xi_{\omega_0} \sim exp(-O(\omega_0^2))$. Similarly, if ρ exhibits power law/ Lorentzian decay (analogous to Cauchy distribution), that is, $\rho(\tau) \sim \frac{1}{O(\tau^2)}$, then P_{ε_β} exhibits exponential decay (Laplace distribution), that is $P_{\varepsilon_\beta}(\omega) \sim exp(-O(|\omega|))$, and therefore $\xi_{\omega_0} \sim exp(-O(|\omega_0|)$. Similar arguments can be made for other decay rates using Fourier duality.

This makes intuitive sense because the more spread out $\rho(\tau)$ is, the more peaky $P_{\varepsilon_\beta}(\omega)$ is and the smaller the value of ω_0 required. This means that if the error terms are well-correlated, most of the instantaneous power will be concentrated within a very small range of frequencies.

Lemma B.2.2. *Suppose* $\Omega = \{-\omega_0 < \omega_i < \omega_0 : i = 1, 2, \cdots |\Omega|\}$ *is a set of real valued frequencies sampled uniformly between* $[-\omega_0, \omega_0]$. *Then, for any* $\xi_D \in (0, 1)$,

with probability at least $1 - exp(-O(|\Omega|^2 \xi_m^2))$ *we have*

$$\int_{-\omega_0}^{\omega_0} P_{\varepsilon_\beta}(\omega) d\omega - \xi_D \leq \frac{1}{|\Omega|} \sum_{\omega \in \Omega} P_{\varepsilon_\beta}(\omega) \leq \int_{-\omega_0}^{\omega_0} P_{\varepsilon_\beta}(\omega) d\omega + \xi_D \tag{B.8}$$

Proof. This is standard Monte Carlo approximation. In particular, consider ω to be a random variable distributed uniformly in $(-\omega_0, \omega_0)$. Now consider the random variable $\zeta(\omega) = P_{\varepsilon_\beta}(\omega)$. Then, we have for this random variable, $\int_{-\omega_0}^{\omega_0} P_{\varepsilon_\beta}(\omega) d\omega = E_{U(-\omega_0,\omega_0)}[\zeta(\omega)] = E[\zeta]$.

Since P_{ε_β} is finite by our assumption, and ω has a finite support $(-\omega_0, \omega_0)$, we also have that $\zeta(\omega) = P_{\varepsilon_\beta}(\omega)$ has a finite support, and we have our result using Hoeffding's inequality (Hoeffding, 1963). □

Lemma B.2.3. *Let* $\xi_{T_0} \in (0, 1)$*. Then, for every* ω*, there exists a* $T_0(\omega)$ *such that*

$$\begin{aligned} -\xi_{T_0} + P_{\varepsilon_\beta}(\omega) < \frac{1}{2T_0(\omega)} E\left[\|\mathbf{X}_{T_0}(\omega)^\top \boldsymbol{\beta} - Y_{T_0}(\omega)\|^2\right] \\ < P_{\varepsilon_\beta}(\omega) + \xi_{T_0} \end{aligned} \tag{B.9}$$

Proof. Define the partial power spectral density of $\varepsilon_\beta(t)$ as

$$g_{\varepsilon_\beta}(T; \omega) = E\left[\frac{\| \int_{-T}^{T} \varepsilon_\beta(t) e^{-\iota\omega t} dt\|^2}{2T}\right]$$

By definition of power spectral density, we have

$$P_{\varepsilon_\beta}(\omega) = \lim_{T \uparrow \infty} E\left[\frac{\| \int_{-T}^{T} \varepsilon_\beta(t) e^{-\iota\omega t} dt\|^2}{2T}\right] = \lim_{T \uparrow \infty} g_{\varepsilon_\beta}(T; \omega)$$

By assumption, the power spectral density is finite and converges for each ω to $P_{\varepsilon_\beta}(\omega)$. Therefore, for every ω, we have that $g_{\varepsilon_\beta}(T;\omega)$ must be a Cauchy sequence with respect to T. That is, for every $\omega \in (-\omega_0, \omega_0)$, and every $\xi_{T_0} \in (0,1)$, $\exists T_0(\omega)$ such that for all $T > T_0(\omega)$,

$$-\xi_{T_0} + P_{\varepsilon_\beta}(\omega) < g_{\varepsilon_\beta}(T;\omega) < P_{\varepsilon_\beta}(\omega) + \xi_{T_0}$$

□

Remark: The exact value of T_0 does not contribute to computation time or space complexity, etc. beyond the computation of the respective Fourier Transforms, and can be chosen as large as required without any additional expenditure in the algorithm. In fact, the optimisation step itself does not depend on T_0, therefore by taking a large enough T_0, we can push ξ_{T_0} to as small as required.

B.2.1 Proof of Theorem 5.3.1

We are now in a position to prove our first main result.

Proof: For any ξ_{T_0}, there always exists a T_0 that is the maximum $T_0(\omega)$ over all $\omega \in (-\omega_0, \omega_0)$ such that Lemma 3 is satisfied, i.e.,

$$\begin{aligned} T_0 = &\min\ T \\ s.t.\quad &|g_{\varepsilon_\beta}(T';\omega) - P_{\varepsilon_\beta}(\omega)| < \xi_{T_0} \\ &\forall T' > T, \forall \omega \in (-\omega_0, \omega_0) \end{aligned}$$

Combining Lemmata B.2.1, B.2.2 and B.2.3 we have, for every $\xi_{T_0}, \xi_D, \xi_{\omega_0} \in (0,1)$, there exist T_0, ω_0 such that for some set $\Omega = \{-\omega_0 < \omega_i < \omega_0 : i = 1, 2, \cdots |\Omega|\}$ sampled uniformly between $(-\omega_0, \omega_0)$, we have with probability at least $1 - exp(-O(|\Omega|^2 \xi_m^2))$

$$\begin{aligned} &- \xi_{T_0} - \xi_D + (1 - \xi_{\omega_0}) E\left[|\mathbf{x}(t)^\top \boldsymbol{\beta} - y(t)|^2\right] \\ &\leq \frac{1}{2|\Omega| T_0} \sum_{\omega \in \Omega} E\left[\|\mathbf{X}_{T_0}(\omega)^\top \boldsymbol{\beta} - Y_{T_0}(\omega)\|^2\right] \\ &\leq E\left[|\mathbf{x}(t)^\top \boldsymbol{\beta} - y(t)|^2\right] + \xi_D + \xi_{T_0} \end{aligned} \tag{B.10}$$

In other words,

$$- \xi_{T_0} - \xi_D + (1 - \xi_{\omega_0}) \mathcal{L}(\boldsymbol{\beta}) \leq \hat{\mathcal{L}}(\boldsymbol{\beta}; \omega_0, T_0, \Omega) \leq \mathcal{L}(\boldsymbol{\beta}) + \xi_D + \xi_{T_0} \tag{B.11}$$

With some algebra, we have,

$$\begin{aligned} \mathcal{L}(\widehat{\boldsymbol{\beta}}) &< \left(\frac{1}{1 - \xi_{\omega_0}}\right) \hat{\mathcal{L}}(\boldsymbol{\beta}; \omega_0, T_0, \Omega) + \frac{1}{1 - \xi_{\omega_0}} (\xi_D + \xi_{T_0}) \\ &< \left(\frac{1}{1 - \xi_{\omega_0}}\right) \hat{\mathcal{L}}(\boldsymbol{\beta}^*; \omega_0, T_0, \Omega) + \frac{1}{1 - \xi_{\omega_0}} (\xi_D + \xi_{T_0}) \\ &< \left(\frac{1}{1 - \xi_{\omega_0}}\right) (\mathcal{L}(\boldsymbol{\beta}^*) + \xi_D + \xi_{T_0}) + \frac{1}{1 - \xi_{\omega_0}} (\xi_D + \xi_{T_0}) \\ &< \left(\frac{1}{1 - \xi_{\omega_0}}\right) \mathcal{L}(\boldsymbol{\beta}^*) + \frac{2}{1 - \xi_{\omega_0}} (\xi_D + \xi_{T_0}) \end{aligned}$$

where the first inequality is due to eq. (B.11), the second by definition of $\boldsymbol{\beta}^*$ and $\widehat{\boldsymbol{\beta}}$,and the final two by eq. (B.11). Therefore, we have

$$E\left[|\mathbf{x}(t)^\top \widehat{\boldsymbol{\beta}} - y(t)|^2\right] < \left(\frac{1}{1-\xi_{\omega_0}}\right) E\left[|\mathbf{x}(t)^\top \boldsymbol{\beta}^* - y(t)|^2\right] + \left(\frac{2}{1-\xi_{\omega_0}}\right)(\xi_D + \xi_{T_0}) \tag{B.12}$$

Choosing T_0, ω_0 and $|\Omega| = D$ such that $\xi_1 = \frac{\xi_{\omega_0}}{1-\xi_{\omega_0}}$, $\xi_2 = 2(\xi_D + \xi_{T_0})$ completes the proof.

■

B.2.2 Proof of Theorem 5.3.2

Note that a T_0-restricted finite Fourier Transform for a signal $z(t)$ is exactly identical to the full Fourier Transform of a T_0-restricted time-limited signal $z_{T_0}(t) = z(t)\mathbb{I}\{|t| < T_0\}$. Therefore, all the exposition in section 5.3.1 in Chapter 5 still hold. In particular, frequency domain representation for aggregated data still follows equation 5.12.

Proof: We require a few modifications to our lemmata to derive the proof of Theorem 5.3.2. In the subsequent analysis, all Fourier Transforms should be assumed to be finite Fourier Transforms, but we omit the T superscript for notational succinctness.[2] We also assume that for every $\omega \in \Omega$ below, we have $|sin(\omega)| > \tau$ for some $\tau > 0$. This will not affect our algorithm because for small enough τ, as long as ω_0

[2]We also omit subscripts from the sinc function notation in the interest of succinctness, they will be clear from context.

is small enough in comparison to $\frac{2\pi}{T}$, the probability of sampling ω which violates this assumption is vanishingly small. In particular, this will be true for $\omega_0 \ll \omega_s/2$ where $\omega_s = \frac{2\pi}{T_s}$ with $T_s = \max\{T_y, T_1, T_2, \cdots T_d\}$.

Denote the reconstructed Fourier Transforms as

$$\widehat{X}_i(\omega) = \frac{\overline{X}_i(\omega)}{U(\omega)},\ \widehat{Y}(\omega) = \frac{\overline{Y}(\omega)}{U(\omega)}$$

Let $\omega_y = \frac{2\pi}{T_y}$ and $\omega_i = \frac{2\pi}{T_i}$. We have

$$\widehat{X}_i(\omega) = X_i(\omega) + \Lambda_{X_i}(\omega|\omega_i) \tag{B.13}$$

$$\widehat{Y}(\omega) = Y(\omega) + \Lambda_Y(\omega|\omega_y) \tag{B.14}$$

where, using the notation of section 5.3.1 in Chapter 5,

$$\Lambda_{X_i}(\omega|\omega_i) = \frac{1}{T_i} \sum_{k\in\mathbb{Z}\setminus\{0\}} X_i(\omega - k\omega_i) \frac{U(\omega - k\omega_i)}{U(\omega)} \tag{B.15}$$

$$\Lambda_Y(\omega|\omega_y) = \frac{1}{T_y} \sum_{k\in\mathbb{Z}\setminus\{0\}} Y(\omega - k\omega_y) \frac{U(\omega - k\omega_y)}{U(\omega)} \tag{B.16}$$

Let $\widehat{\mathbf{x}}(t), \widehat{y}(t)$ and $\lambda_i(t), \lambda_y(t)$ be the corresponding time domain signals. Use the following notation

$$\varepsilon_\beta(t) = \mathbf{x}(t)\boldsymbol{\beta} - y(t) \tag{B.17}$$

$$\hat{\varepsilon}_\beta(t) = \widehat{\mathbf{x}}(t)\boldsymbol{\beta} - \widehat{y}(t) \tag{B.18}$$

$$\varepsilon_{\lambda,\beta}(t) = \lambda_x(t)\boldsymbol{\beta} - \lambda_y(t) \tag{B.19}$$

Clearly, $\hat{\varepsilon}_\beta(t) = \varepsilon_\beta(t) + \varepsilon_{\lambda,\beta}(t)$. Denote the corresponding power spectral densities as $\hat{P}_{\hat{\varepsilon}_\beta}, P_{\varepsilon_\beta}, P_{\varepsilon_{\lambda,\beta}}$. We now show the following result

---

Lemma B.2.4. *Suppose the power spectral densities of* $\mathbf{x}(t), y(t)$ *are finite for every* $\omega \in (-\omega_0, \omega_0)$*, and decay rapidly at a sub-Gaussian rate* $e^{-O((\omega-\omega_0)^2)}$ *beyond* $|\omega| > \omega_0$*.*

Then, we have, for any $\omega \in (-\omega_0, \omega_0)$

$$\hat{P}_{\hat{\varepsilon}_\beta}(\omega) - e^{-O((\omega_s - 2\omega_0)^2)} \leq P_{\varepsilon_\beta}(\omega) \leq \hat{P}_{\hat{\varepsilon}_\beta}(\omega) + e^{-O((\omega_s - 2\omega_0)^2)} \tag{B.20}$$

where $\omega_s = \frac{2\pi}{T_s}$ *with* $T_s = \max\{T_y, T_1, T_2, \cdots T_d\}$.

Proof. First, note that as a result of our assumptions, the power spectral densities $\hat{P}_{\hat{\varepsilon}_\beta}, P_{\varepsilon_\beta}, P_{\varepsilon_{\lambda,\beta}}$ are also finite for every $\omega \in (-\omega_0, \omega_0)$, and decays rapidly at a sub-Gaussian rate $e^{-O(\omega-\omega_0)^2}$ beyond $|\omega| > \omega_0$. Suppose $\hat{P}_{\hat{\varepsilon}_\beta}, P_{\varepsilon_\beta}, P_{\varepsilon_{\lambda,\beta}} < \gamma^2$ for some finite $\gamma > 0$.

The proof of this result requires two steps. First, suppose $g(t), h(t)$ are any two signals with corresponding (finite) power spectral densities P_g, P_h. Then, we have

$$P_{g+h} \leq P_g + P_h + 2\sqrt{P_g P_h} \tag{B.21}$$

The proof of this is easy, and proceeds by simply expanding the expression for power spectral density and using standard results from real analysis and probability

theory.

$$
\begin{aligned}
P_{g+h} &= \lim_{T\uparrow\infty} \frac{1}{T} E\left[|G_T(\omega) + H_T(\omega)|^2\right] \\
&\leq \lim_{T\uparrow\infty} \frac{1}{T} E\left[|G_T(\omega)|^2 + |H_T(\omega)|^2 + 2|G_T(\omega)H_T(\omega)|\right] \\
&\quad \text{(Triangle Inequality)} \\
&\leq \lim_{T\uparrow\infty} \frac{1}{T}\left[E|G_T(\omega)|^2 + E|H_T(\omega)|^2\right] \\
&+ \lim_{T\uparrow\infty} \frac{2}{T}\left[\sqrt{E[|G_T(\omega)H_T(\omega)|^2]}\right] \text{ (Jensen's Ineq.)} \\
&\leq \lim_{T\uparrow\infty} \frac{1}{T} E|G_T(\omega)|^2 + \lim_{T\uparrow\infty} \frac{1}{T} E||H_T(\omega)|^2 \\
&+ 2\sqrt{\lim_{T\uparrow\infty} \frac{1}{T} E[|G_T(\omega)|^2 \lim_{T\uparrow\infty} \frac{1}{T} E|H_T(\omega)|^2]} \\
&\text{(Cauchy-Schwartz, limit theorems)} \\
&= P_g + P_h + 2\sqrt{P_g P_h}
\end{aligned}
$$

Therefore, using this result, the definitions of $\hat{\varepsilon}_\beta(t), \varepsilon_\beta(t), \varepsilon_{\lambda,\beta}(t)$ and the fact that $P_{-z} = P_z$ for any signal z, we have,

$$
\begin{aligned}
\hat{P}_{\hat{\varepsilon}_\beta}(\omega) &- \left(P_{\varepsilon_{\lambda,\beta}}(\omega) + 2\gamma\sqrt{P_{\varepsilon_{\lambda,\beta}}(\omega)}\right) \\
&\leq P_{\varepsilon_\beta}(\omega) \\
&\leq \hat{P}_{\hat{\varepsilon}_\beta}(\omega) + \left(P_{\varepsilon_{\lambda,\beta}}(\omega) + 2\gamma\sqrt{P_{\varepsilon_{\lambda,\beta}}(\omega)}\right)
\end{aligned} \tag{B.22}
$$

We can easily extend equation (B.21) to the following standard result. Suppose $z_i(t) : i = 1, 2, \cdots$ are an arbitrary set of signals. Then,

$$
P_{\Sigma_i z_i} \leq \left(\sum_i \sqrt{P_{z_i}}\right)^2 \tag{B.23}
$$

This result works for infinite sums provided the right hand side exists finitely. The proof of this also proceeds by expanding the expression for power spectral density, and using standard limit theorems.

We shall use this to show that $P_{\varepsilon_{\lambda,\beta}}(\omega) \sim e^{-O(\omega_s - 2\omega_0)^2}$. Define the following quantities

$$\Lambda_{X_i,k}(\omega|\omega_i) = \frac{1}{T_i} X_i(\omega - k\omega_i) \frac{U(\omega - k\omega_i)}{U(\omega)}$$
$$\Lambda_{Y,k}(\omega|\omega_y) = \frac{1}{T_y} Y(\omega - k\omega_y) \frac{U(\omega - k\omega_y)}{U(\omega)}$$

Define $\lambda_{x_i,k}(t) = \mathcal{F}^{-1}\Lambda_{X_i,k}, \lambda_{y,k}(t) = \mathcal{F}^{-1}\Lambda_{Y,k}$. Clearly,

$$\Lambda_{X_i}(\omega|\omega_i) = \sum_{k\in\mathbb{Z}\setminus\{0\}} \Lambda_{X_i,k}(\omega|\omega_i) \tag{B.24}$$
$$\Lambda_{Y}(\omega|\omega_y) = \sum_{k\in\mathbb{Z}\setminus\{0\}} \Lambda_{Y,k}(\omega|\omega_y) \tag{B.25}$$
$$\lambda_i(t) = \sum_{k\in\mathbb{Z}\setminus\{0\}} \lambda_{x_i,k}(t) \tag{B.26}$$
$$\lambda_y(t) = \sum_{k\in\mathbb{Z}\setminus\{0\}} \lambda_{y,k}(t) \tag{B.27}$$

We note that for any signal $z(t)$, if $P_z(\omega) \sim e^{-O(\omega^2)}$ and $\tau(\omega)$ is a strictly bounded function of ω, then for $\lambda_z(t) = \mathcal{F}^{-1}Z(\omega)\tau(\omega)$, we have $P_\lambda(\omega) \sim e^{-O(\omega^2)}$.

By assumption, $P_{x_i}(\omega), P_y(\omega) \sim e^{-O(\omega-\omega_0)^2}$ and for the values of ω we use $\frac{U(\omega-k\omega_y)}{U(\omega)}$ is strictly bounded, therefore, we can show that $P_{\lambda_{x_i,k}}(\omega), P_{\lambda_{y,k}}(\omega) \sim e^{-O(\omega-\omega_0-k\omega_y)^2}$

We have, $\lambda_i(t) = \sum_k \lambda_{x_i,k}(t)$ and $\lambda_y(t) = \sum_k \lambda_{y,k}(t)$. Therefore, we have by equation B.23,

$$
\begin{aligned}
P_{\lambda_i}(\omega) &= \left(\sum_{k \in \mathbb{Z} \setminus \{0\}} \sqrt{P_{\lambda_{x_i,k}}} \right) \\
&= 2 \left(\sum_{k=1}^{\infty} \sqrt{P_{\lambda_{x_i,-k}}} \right) \text{ by symmetry around 0} \\
&= 2 \sum_{k=1}^{\infty} e^{-O(k\omega_y + \omega - \omega_0)^2} \\
&\sim e^{-O(\omega_i - \omega_0 + \omega)^2}
\end{aligned}
$$

Similarly, $P_{\lambda_y}(\omega) \sim e^{-O(\omega_y - \omega_0 + \omega)^2}$. The final step uses standard approximation techniques exploiting the fact that $\sum_n f(n) \sim \Theta(\int_x f(x)dx)$ for bounded, finite, monotonic functions f, and noting that $e^{-O(k\omega_y + \omega - \omega_0)^2}$ has Gaussian decay in terms of k, and the area under Gaussian functions over a subset of the positive real line is given by the complementary error function $erfc(\cdot)$. We also use the fact (Chang et al., 2011) that the complementary error function has a Gaussian decay $erfc(x) \sim e^{-O(x^2)}$.

If $\omega_s = \min\{\omega_y, \omega_1, \omega_2, \cdots, \omega_d\}$, and for $\omega \in (-\omega_0, \omega_0)$, we have in terms of ω_s the fact that $e^{-O(\omega - \omega_0 - \omega_y)^2} < e^{-O(\omega_s - 2\omega_0)^2}$. For $\omega_s > 2\omega_0$, these approximations can be written more succinctly as $P_{\lambda_i}(\omega), P_{\lambda_y}(\omega) \sim e^{-O(\omega_s - 2\omega_0)^2}$.

Finally, we note that by definition and using (B.23), we have $P_{\varepsilon_{\lambda,\beta}}(\omega) \leq (\beta_i \sum_{i=1}^{d} \sqrt{P_{\lambda_i}(\omega)} + \sqrt{P_{\lambda_y}(\omega)})^2$. For fixed d and since by assumption $|\boldsymbol{\beta}|$ is bounded, we have $P_{\varepsilon_{\lambda,\beta}}(\omega) \sim e^{-O(\omega_s - 2\omega_0)^2}$ and therefore, $\left(P_{\varepsilon_{\lambda,\beta}}(\omega) + 2\gamma \sqrt{P_{\varepsilon_{\lambda,\beta}}(\omega)} \right) \sim e^{-O(\omega_s - 2\omega_0)^2}$. This completes the proof for Lemma B.2.4.

□

The final piece of the proof is to approximate $E\|\widehat{\mathbf{X}}_{T_0}(\omega) - \widehat{Y}_{T_0}(\omega)\|^2$. By assumption, the individual processes at each location is strictly sub-Gaussian (Buldygin & Kozachenko, 2000; Mendelson, 2011). Simply put, this means that for each signal $z(t)$ at each time t, the logarithm of the moment generating function is quadratically bounded

$$\forall b > 0, \ln E[e^{b(z(t)-\mu)}] < \frac{b^2\sigma^2}{2}$$

for some constant σ, where $\mu = E[z(t)]$.

Since by assumption our random processes are bounded and almost surely finite, it can be shown by using results from calculus and probability theory that finite aggregation and Finite Fourier Transforms preserve sub-Gaussian property being linear operations[3]. In particular, note that most of our Fourier Transform computations can be estimated by discrete sums using the DTFT-DFT dual relationship, and linear sums preserve the sub-Gaussian property.

Now, we have that by using Hoeffding's inequality on sub-Gaussian random variables (Georgiou & Kyriakakis, 2006; Hsu et al., 2012), we can show that for independent observations $\{(\widehat{\mathbf{X}}^j(\omega), \widehat{Y}^j(\omega)) : j = 1, 2, \cdots N\}$ from N locations, for

[3]An easy way to prove it, for example, would be to represent integration as the limit of a Riemann sum using definition from first principles, and to use the bounded convergence theorem and continuity of the exponentiation operator with standard limit theorems

any small ξ, we have with probability $1 - exp(-O(N^2\xi_3^2))$,

$$E\|\widehat{\mathbf{X}}_{T_0}(\omega)\boldsymbol{\beta} - \widehat{Y}_{T_0}(\omega)\|^2 - \xi \tag{B.28}$$

$$< \frac{1}{N}\sum_{j\in[N]} \|\widehat{X}^j_{T_0}(\omega)^\top\boldsymbol{\beta} - \widehat{Y}^j_{T_0}(\omega)\|^2 \tag{B.29}$$

$$< E\|\widehat{\mathbf{X}}_{T_0}(\omega)\boldsymbol{\beta} - \widehat{Y}_{T_0}(\omega)\|^2 + \xi \tag{B.30}$$

Choose ξ such that $\xi_3 = (1 + \frac{1}{T_0})\xi$. Theorem 5.3.2 now follows in a manner exactly identical to the proof of Theorem 5.3.1, with the addition of two extra steps that incorporates Lemma B.2.4 and equation B.28. ■

Finally we note that Theorem 5.3.2 is only one of many possible results that can be obtained for estimation using our techniques. In particular, usage of different assumptions on the data distribution, and different decay rates on the power spectral densities can be used to derive alternative guarantees.

The proofs for results in the multidimensional case are exactly identical, except for the size of the sampled frequency set $|\Omega| = D$. As mentioned in chapter 5, D can grow exponentially in the ambient dimensionality p of the interaction space $\mathbb{R}^p$. This is because the sampled frequencies are expected to cover a certain volume, and volume grows exponentially with dimensionality. However, in most real life cases, p will be very small (for example $p \leq 4$ for spatio-temporal applications), hence the increase in required size is in and of itself no major impediment in application of our algorithmic framework.

B.3 Discussion: Decay Rates

Throughout this dissertation, we assume that the power spectral density and autocovariance function for every signal of interest exists finitely for each ω. We further assume that the autocovariance function decays rapidly with lag for all processes involved in our analysis. In essence this means that the value of the time series at any given point is highly correlated with values at points close to it in time, but the correlation decreases rapidly with values farther away in time.

In particular, we assume that $\rho_{(\cdot)}(\cdot)$ is a Schwartz function (Terzioğlu, 1969), that is $\rho(\cdot)$ and all its derivatives decay at least as fast as any inverse polynomial. That is, $\forall \alpha, \beta \in \mathbb{Z}_+^n$ we have

$$|\zeta^\alpha \frac{\partial^\beta \rho(\zeta)}{\partial \zeta^n}| \to 0 \quad as \quad |\zeta| \to \infty$$

Examples of Schwartz functions are exponential functions like $e^{-a\zeta^2}$ for $a > 0$, or any polynomial $\wp(\zeta)$ multiplied with an exponential function like $\wp(\zeta)e^{-a\zeta^2}$, or any smooth domain-restricted function $f(\zeta)$ which is 0 outside of a bounded compact subset $\zeta \in \Im \subset \mathbb{R}^n$ (e.g. time limited signals).

A key property of Schwartz functions is that the Fourier Transform of a Schwartz function is itself a Schwartz function (Gröchenig & Zimmermann, 2001; Strichartz, 2003). Therefore, if we assume that the covariance functions $\rho_{(\cdot)}(\tau)$ decays rapidly with τ for each of our signals, then their corresponding power spectral densities $P_{(\cdot)}(\omega)$ will decay rapidly with ω, since $P = \mathcal{F}\rho$. Therefore, most of the power for our signals will be concentrated around $\omega = 0$.

As seen earlier, the decay rates of the power spectral density and autocovariance function complement each other- e.g., if ρ exhibits a Gaussian decay, then P_{ε_β} also exhibits a Gaussian decay. Similarly, if ρ exhibits power law or Lorentzian decay, then P_{ε_β} exhibits exponential decay. The exact decay rates involved will vary on a case to case basis, but in essence, this means that we only need to care about a small set of frequencies around 0 to describe the signal up to a reasonable approximation.

We note that unlike traditional signal processing applications, we do not consider a flat power spectral density (e.g. white noise) for our noise process. This is because traditional signal processing applications assume band-limited signals of interest. Properties of the noise process outside the band are irrelevant since outputs are going to be filtered regardless, and analysis only needs to focus on effects of additive noise within the frequency band of interest. In our case, we can make no such assumption– signals need not be bandlimited and therefore we have to consider effects of noise throught the entire spectrum[4].

[4]Note that a true white noise process is unrealistic because it implies infinite variance for the noise process which renders any attempt at parameter learning futile.

Appendix C

Aggregation Paradigms and Learning with Sensitive Data: Appendix

C.1 Proof of Proposition 6.3.1

Proof. Let $\hat{p}$ be the probability of error of our algorithm (call it $\mathcal{A}$ on a randomly drawn datapoint. Let λ be the probability of error on the same data point of our baseline algorithm (call it $\mathcal{B}$). Then, the probability ϱ that our algorithm does at least as well as the baseline algorithm on that datapoint is:

$$\begin{aligned}
\varrho &= P(\mathcal{A} \text{ is correct}) + P(\text{Both } \mathcal{A} \text{ and } \mathcal{B} \text{ are incorrect}) \\
&= 1 - \hat{p} + \hat{p}\lambda \\
&\quad \text{since our algorithm is trained} \\
&\quad \text{independent of the baseline} \\
&= 1 - \hat{p}(1 - \lambda)
\end{aligned}$$

Therefore, the probability that our algorithm $\mathcal{A}$ does worse than the baseline is at most $\hat{p}(1 - \lambda)$.

Now, all we need to do is quantify $\hat{p}$ in terms of the number of learners M and the probability of error or misclassification probability p of each learner. Recall

that we use the median prediction on any data point for our final class estimate. Therefore, with M learners, the final prediction will be worse than the baseline algorithm if and only if at least $\frac{M}{2}$ of the learners do worse than the baseline.

Note that each learner is trained using a randomly selected sample, and each such sample is independently selected and is of fixed size. Therefore, the prediction of each sample can be seen as an independent $Bernoulli(p)$ random variable. Therefore, our final misclassification probability is the following:

$$P(\text{Error}) = \sum_{k \geq \frac{M}{2}} \binom{M}{k} p^k (1-p)^{M-k} \tag{C.1}$$

$$= (1-p)^M \sum_{k \geq \frac{M}{2}} \binom{M}{k} \left(\frac{p}{1-p}\right)^k \tag{C.2}$$

Let $\beta = \frac{p}{1-p}$.

The above probability can be approximated using Stirling's Approximation for binomial coefficients. Recall that using Stirling's Approximation, $\binom{M}{k}$ can be written as

$$log(\binom{M}{k}) \approx MH(\frac{k}{M}) - O(log(M)) + O(\frac{1}{M})$$

where $H(\alpha) = -\alpha log_2(\alpha) - (1-\alpha) log_2(1-\alpha)$ for any $\alpha \in (0,1)$ is the Binary Entropy function in nats.

Since for $H(\cdot) \leq \kappa = log2 \approx 0.693$, we have:

$$log(\binom{M}{k}) \leq \kappa M - O(log(M)) + O(\frac{1}{M})$$

Let $\beta = \frac{p}{1-p}$, and $\gamma = exp\left(\kappa - O(\frac{log(M)}{M}) + O(\frac{1}{M^2})\right)$

Therefore, we have:

$$\begin{aligned} P(Error) &\leq \gamma^M (1-p)^M \sum_{k \geq \frac{M}{2}} \beta^k \\ &= \gamma^M (1-p)^M \frac{1}{1-\beta} \beta^{M/2} \left(1 - \beta^{M/2}\right) \end{aligned}$$

Since by assumption, $p < 0.5$, we have $\beta < 1$, and therefore, the error probability $P(Error)$ is bounded above by as

$$\frac{1-p}{1-2p} \left[(1-p)p \; exp\left(2\kappa - O(\frac{log(M)}{M}) + O(\frac{1}{M^2})\right)\right]^{M/2}$$

This completes the proof.

□

C.2 Proof of Proposition 6.3.2

Claim: If $\mathbf{g}_{\boldsymbol{\phi}}^{-1}$ (equivalently g_ϕ) is a linear function[1], $\widehat{\boldsymbol{\theta}}$ is an unbiased estimator of $\boldsymbol{\theta}^*$

[1] gaussian, exponential, pareto, chi-squared, etc.

Proof. Omitting the subscript T for succinctness:

$$\begin{aligned}
E[\widehat{\boldsymbol{\theta}}_T] &= E[(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{g}_{\boldsymbol{\phi}}^{-1}(\mathbf{y})] \\
&= E[(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}E[\mathbf{g}_{\boldsymbol{\phi}}^{-1}(\mathbf{y})|\mathbf{X}]] \\
&= E[(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{g}_{\boldsymbol{\phi}}^{-1}(E[\mathbf{y}|\mathbf{X}])] \\
&\quad \text{since } g_\phi \text{ is linear} \\
&= E[(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{g}_{\boldsymbol{\phi}}^{-1}(g_\phi(\mathbf{X}^\top\boldsymbol{\theta}^*))] \\
&\quad \text{by definition of GLMs} \\
&= E[(\mathbf{X}\mathbf{X}^\top)^{-1}\mathbf{X}\mathbf{X}^\top\boldsymbol{\theta}^*] \\
&= \boldsymbol{\theta}^*
\end{aligned}$$

□

Bibliography

Yield monitor data for a corn field in Argentina with variable nitrogen. https://rdrr.io/cran/agridat/man/lasrosas.corn.html.

Acharyya, Sreangsu and Ghosh, Joydeep. Memr: A margin equipped monotone retargeting framework for ranking. In *UAI*, pp. 2–11, 2014.

Acharyya, Sreangsu, Koyejo, Oluwasanmi, and Ghosh, Joydeep. Learning to rank with bregman divergences and monotone retargeting. *arXiv preprint arXiv:1210.4851*, 2012.

Adams, William W and Loustaunau, Philippe. *An introduction to Gröbner bases*, volume 3. American Mathematical Society Providence, 1994.

AdWeek. U.S. Digital Advertising Will Make $ 83 Billion This YYear Says EMarketer, March 14, 2017. http://www.adweek.com/digital/u-s-digital-advertising-will-make-83-billion-this-yyear-says-emarketer/.

Aggarwal, Charu C and Philip, S Yu. A general survey of privacy-preserving data mining models and algorithms. In *Privacy-preserving data mining*, pp. 11–52. Springer, 2008.

Armstrong, Marc P, Rushton, Gerard, and Zimmerman, Dale L. Geographically masking health data to preserve confidentiality. *Statistics in Medicine*, 18(5): 497–525, 1999.

Banerjee, Arindam, Merugu, Srujana, Dhillon, Inderjit S, and Ghosh, Joydeep. Clustering with Bregman divergences. *The Journal of Machine Learning Research*, 6:1705–1749, 2005.

Bhowmik, Avradeep, Ghosh, Joydeep, and Koyejo, Oluwasanmi. Generalized Linear Models for Aggregated Data. In *Proceedings of the Eighteenth International Conference on Artificial Intelligence and Statistics*, pp. 93–101, 2015.

Bickel, Peter J, Hammel, Eugene A, and O'Connell, J William. Sex bias in graduate admissions: Data from berkeley. *Science*, 187(4175):398–404, 1975.

Bongiovanni, Rodolfo and Lowenberg-DeBoer, James. Nitrogen management in corn using site-specific crop response estimates from a spatial regression model. In *Proceedings of the Fifth International Conference on Precision Agriculture*, 2000.

Bottou, Léon. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pp. 177–186. Springer, 2010.

Breyer, Patrick. Telecommunications data retention and human rights: the compatibility of blanket traffic data retention with the echr. *European Law Journal*, 11 (3):365–375, 2005.

Brown, Ian. Communications data retention in an evolving internet. *International Journal of Law and Information Technology*, 19(2):95–109, 2010.

Brown, Philip J and Payne, Clive D. Aggregate data, ecological regression, and voting transitions. *Journal of the American Statistical Association*, 81(394): 452–460, 1986.

Buldygin, V. V. and Kozachenko, I. V. *Metric characterization of random variables and random processes*, volume 188. American Mathematical Soc., 2000.

Burrell, Jenna, Brooke, Tim, and Beckwith, Richard. Vineyard computing: Sensor networks in agricultural production. *IEEE Pervasive computing*, 3(1):38–45, 2004.

Cai, Deng, He, Xiaofei, and Han, Jiawei. Spectral regression for efficient regularized subspace learning. In *2007 IEEE 11th international conference on computer vision*, pp. 1–8. IEEE, 2007.

Cai, T Tony, Xu, Guangwu, and Zhang, Jun. On recovery of sparse signals via ℓ_1 minimization. *Information Theory, IEEE Transactions on*, 55(7):3388–3397, 2009.

Cai, T Tony, Wang, Lie, and Xu, Guangwu. Shifting inequality and recovery of sparse signals. *Signal Processing, IEEE Transactions on*, 58(3):1300–1308, 2010a.

Cai, Tony Tony, Wang, Lie, and Xu, Guangwu. Stable recovery of sparse signals and an oracle inequality. 2010b.

Candes, Emmanuel and Tao, Terence. The Dantzig selector: statistical estimation when p is much larger than n. *The Annals of Statistics*, pp. 2313–2351, 2007.

Candes, Emmanuel J. The restricted isometry property and its implications for compressed sensing. *Comptes Rendus Mathematique*, 346(9):589–592, 2008.

Candes, Emmanuel J and Plan, Yaniv. A probabilistic and ripless theory of compressed sensing. *Information Theory, IEEE Transactions on*, 57(11):7235–7254, 2011.

Candès, Emmanuel J and Recht, Benjamin. Exact matrix completion via convex optimization. *Foundations of Computational mathematics*, 9(6):717, 2009.

Candes, Emmanuel J and Tao, Terence. Decoding by linear programming. *Information Theory, IEEE Transactions on*, 51(12):4203–4215, 2005.

Candes, Emmanuel J and Tao, Terence. Near-optimal signal recovery from random projections: Universal encoding strategies? *Information Theory, IEEE Transactions on*, 52(12):5406–5425, 2006.

Candès, Emmanuel J, Romberg, Justin, and Tao, Terence. Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information. *Information Theory, IEEE Transactions on*, 52(2):489–509, 2006.

Candes, Emmanuel J, Romberg, Justin K, and Tao, Terence. Stable signal recovery from incomplete and inaccurate measurements. *Communications on pure and applied mathematics*, 59(8):1207–1223, 2006.

Chang, Seok-Ho, Cosman, Pamela C, and Milstein, Laurence B. Chernoff-type bounds for the gaussian error function. *IEEE Transactions on Communications*, 59(11):2939–2944, 2011.

Chi, Yuejie, Scharf, Louis L, Pezeshki, Ali, and Calderbank, A Robert. Sensitivity to basis mismatch in compressed sensing. *Signal Processing, IEEE Transactions on*, 59(5):2182–2195, 2011.

Cohen, Morris R and Nagel, Ernest. An introduction to logic and scientific method (new york, 1934). *Compare Norman R. Campbell:Measurement is the process of assigning numbers to represent qualities, Foundations of Science*, pp. 294, 1957.

Corbae, Dean, Ouliaris, Sam, and Phillips, Peter CB. Band spectral regression with trending data. *Econometrica*, 70(3):1067–1109, 2002.

Cortez, Paulo and Morais, Aníbal de Jesus Raimundo. A data mining approach to predict forest fires using meteorological data. 2007.

criteo.com. https://www.criteo.com/.

Da Xu, Li, He, Wu, and Li, Shancang. Internet of things in industries: A survey. *IEEE Transactions on Industrial Informatics*, 10(4):2233–2243, 2014.

Davidson, James EH, Hendry, David F, Srba, Frank, and Yeo, Stephen. Econometric modelling of the aggregate time-series relationship between consumers' expenditure and income in the united kingdom. *The Economic Journal*, pp. 661–692, 1978.

DESynPUF. Medicare Claims Synthetic Public Use Files (SynPUFs). *Centers for Medicare and Medicaid Services*, 2008. http://www.cms.gov/Research-Statistics-Data-and-Systems/Downloadable-Public-Use-Files/SynPUFs/index.html.

Dias, Sofia, Sutton, Alex J, Ades, AE, and Welton, Nicky J. Evidence synthesis for decision making 2 a generalized linear modeling framework for pairwise and network meta-analysis of randomized controlled trials. *Medical Decision Making*, 33(5):607–617, 2013.

Donoho, David L. For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Communications on pure and applied mathematics*, 59(6):797–829, 2006.

Donoho, David L and Elad, Michael. Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ_1 minimization. *Proceedings of the National Academy of Sciences*, 100(5):2197–2202, 2003.

Doucet, Arnaud, De Freitas, Nando, and Gordon, Neil. An introduction to sequential monte carlo methods. In *Sequential Monte Carlo methods in practice*, pp. 3–14. Springer, 2001.

Duncan, G, Elliot, M, and Salazar-González, JJ. Statistical confidentiality: Principles and practice 2011.

Dwork, Cynthia. Differential privacy: A survey of results. In *International Conference on Theory and Applications of Models of Computation*, pp. 1–19. Springer, 2008.

Dwork, Cynthia, Roth, Aaron, et al. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3–4): 211–407, 2014.

Dzhaparidze, Kacha. *Parameter estimation and hypothesis testing in spectral analysis of stationary time series*. Springer Science & Business Media, 2012.

Easton, Roger L. Multidimensional fourier transforms. *Fourier Methods in Imaging*, pp. 325–346.

Federal Trade Commission. *FACTA Disposal Rule Goes into Effect June 1*. 2005. https://www.ftc.gov/news-events/press-releases/2005/06/facta-disposal-rule-goes-effect-june-1.

Federal Trade Commission. *Under COPPA, data deletion isnt just a good idea. Its the law.* 2018. https://www.ftc.gov/news-events/blogs/business-blog/2018/05/under-coppa-data-deletion-isnt-just-good-idea-its-law.

Feige, Edgar L and Pearce, Douglas K. The causality relationship between money and income: A time series approach. *Quarterly Journal of Business and Economics*, 13(4):183, 1974.

Feldman, Jon, Henzinger, Monika, Korula, Nitish, Mirrokni, Vahab S, and Stein, Cliff. Online stochastic packing applied to display ad allocation. In *European Symposium on Algorithms*, pp. 182–194. Springer, 2010.

fivethirtyeight.com. The supreme court is allergic to math. 2017. https://fivethirtyeight.com/features/the-supreme-court-is-allergic-to-math/.

Foucart, Simon. A note on guaranteed sparse recovery via ℓ_1-minimization. *Applied and Computational Harmonic Analysis*, 29(1):97–103, 2010.

Freedman, David A, Klein, Stephen P, Sacks, Jerome, Smyth, Charles A, and Everett, Charles G. Ecological regression and voting rights. *Evaluation Review*, 15(6): 673–711, 1991a.

Freedman, David A, Klein, Stephen P, Sacks, Jerome, Smyth, Charles A, and Everett, Charles G. Ecological regression and voting rights. *Evaluation Review*, 15(6): 673–711, 1991b.

Gart, John J. The analysis of poisson regression with an application in virology. *Biometrika*, 51(3/4):pp. 517–521, 1964. ISSN 00063444.

Georgiou, Panayiotis G and Kyriakakis, Chris. Robust maximum likelihood source localization: the case for sub-gaussian versus gaussian. *Audio, Speech, and Language Processing, IEEE Transactions on*, 14(4):1470–1480, 2006.

Ghosh, Arpita, Rubinstein, Benjamin IP, Vassilvitskii, Sergei, and Zinkevich, Martin. Adaptive bidding for display advertising. In *Proceedings of the 18th international conference on World wide web*, pp. 251–260. ACM, 2009.

Goldfarb, Avi and Tucker, Catherine. Online display advertising: Targeting and obtrusiveness. *Marketing Science*, 30(3):389–404, 2011.

Good, Phillip I. *Permutation, parametric and bootstrap tests of hypotheses*, volume 3. Springer, 2005.

Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, and Bengio, Yoshua. Generative ad-

versarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Goodman, Leo A. Ecological regressions and behavior of individuals. *American Sociological Review*, 1953.

Google Product Taxonomy. Google product taxonomy. https://www.google.com/basepages/producttype/taxonomy.en-US.txt.

Grafakos, Loukas. *Classical and modern Fourier analysis*. Prentice Hall, 2004.

Granger, Clive William John and Newbold, Paul. *Forecasting economic time series*. Academic Press, 2014.

Gröchenig, Karlheinz and Zimmermann, Georg. Hardy's theorem and the short-time fourier transform of schwartz functions. *Journal of the London Mathematical Society*, 63(1):205–214, 2001.

Hardelid, P, Pebody, R, and Andrews, N. Mortality caused by influenza and respiratory syncytial virus by age group in England and Wales 1999–2010. *Influenza and other respiratory viruses*, 7(1):35–45, 2013.

Herman, Matthew A and Strohmer, Thomas. General deviants: An analysis of perturbations in compressed sensing. *Selected Topics in Signal Processing, IEEE Journal of*, 4(2):342–349, 2010.

Hibon, Michael and Makridakis, Spyros. Arma models and the box–jenkins methodology. 1997.

Ho, Joyce C, Park, Yubin, Carvalho, Carlos, and Ghosh, Joydeep. Dynacare: Dynamic cardiac arrest risk estimation. In *AISTATS*, pp. 333–341, 2013.

Hoeffding, Wassily. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

Hsu, Daniel, Kakade, Sham M, and Zhang, Tong. A tail inequality for quadratic forms of subgaussian random vectors. *Electron. Commun. Probab*, 17(52):1–6, 2012.

Hu, Yu, Shin, Jiwoong, and Tang, Zhulei. Pricing of online advertising: cost-per-click-through vs. cost-per-action. In *System Sciences (HICSS), 2010 43rd Hawaii International Conference on*, pp. 1–9. IEEE, 2010.

Hung, Shin-Yuan, Yen, David C, and Wang, Hsiu-Yu. Applying data mining to telecom churn management. *Expert Systems with Applications*, 31(3):515–524, 2006.

IBM TJ Watson. Using customer behavior data to improve customer retention. *IBM TJ Watson*. https://www.ibm.com/communities/analytics/watson-analytics-blog/predictive-insights-in-the-telco-customer-churn-data-set/.

Jamil, Tahira, Ozinga, Wim A, Kleyer, Michael, and ter Braak, Cajo JF. Selecting traits that explain species–environment relationships: a generalized linear mixed model approach. *Journal of Vegetation Science*, 24(6):988–1000, 2013.

Johnson, NL, Kotz, S, and Balakrishnan, N. Lognormal distributions. continuous univariate distributions (vol. 1), 1994.

Kaggle Churn in Telecom. Churn in telecom dataset. https://www.kaggle.com/becksddf/churn-in-telecoms-dataset.

Katz, Sheldon. *Enumerative geometry and string theory*. American Mathematical Soc., 2006.

Kievit, Rogier, Frankenhuis, Willem Eduard, Waldorp, Lourens, and Borsboom, Denny. Simpson's paradox in psychological science: a practical guide. *Frontiers in psychology*, 4:513, 2013.

King, Gary, Tanner, Martin A, and Rosen, Ori. *Ecological inference: New methodological strategies*. Cambridge University Press, 2004.

Kingma, Diederik P and Welling, Max. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

Koopmans, Lambert H. *The spectral analysis of time series*. Academic press, 1995.

Kramer, Gerald H. The ecological fallacy revisited: Aggregate-versus individual-level findings on economics and elections, and sociotropic voting. *American political science review*, 77(1):92–111, 1983.

Lambert, Dayton M, Lowenberg-Deboer, James, and Bongiovanni, Rodolfo. A comparison of four spatial regression models for yield monitor data: A case study from argentina. *Precision Agriculture*, 5(6):579–600, 2004.

Leskovec, Jure, Rajaraman, Anand, and Ullman, Jeffrey David. *Mining of massive datasets*. Cambridge university press, 2014.

Li, Shancang, Da Xu, Li, and Wang, Xinheng. Compressed sensing signal and data acquisition in wireless sensor networks and internet of things. *IEEE Transactions on Industrial Informatics*, 9(4):2177–2186, 2013.

Liberty, Edo. Simple and deterministic matrix sketching. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 581–588. ACM, 2013.

Lindell, Y. and Pinkas, B. Privacy preserving data mining. *LNCS*, 1880:36–77, 2000.

Liu, Yan, Niculescu-Mizil, Alexandru, Lozano, Aurelie C, and Lu, Yong. Learning temporal causal graphs for relational time-series analysis. In *Proceedings of the 27th International Conference on Machine Learning (ICML-10)*, pp. 687–694, 2010.

Lozano, Aurelie C, Li, Hongfei, Niculescu-Mizil, Alexandru, Liu, Yan, Perlich, Claudia, Hosking, Jonathan, and Abe, Naoki. Spatial-temporal causal modeling for climate change attribution. In *Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 587–596. ACM, 2009.

McCullagh, Peter and Nelder, John A. Generalized linear models. 1989.

Mendelson, Shahar. Discrepancy, chaining and subgaussian processes. *The Annals of Probability*, pp. 985–1026, 2011.

Merugu, S. and Ghosh, J. A distributed learning framework for heterogeneous data sources. In *Proc. KDD*, pp. 208–217, 2005.

Merugu, S. and Ghosh, J. Privacy perserving distributed clustering using generative models. In *Proc. ICDM*, pp. 211–218, Nov, 2003.

Montanari, Andrea et al. Computational implications of reducing data to sufficient statistics. *Electronic Journal of Statistics*, 9(2):2370–2390, 2015.

Nelder, J. A. and Wedderburn, R. W. M. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):pp. 370–384, 1972. ISSN 00359238.

Nelder, John A and Baker, RJ. *Generalized linear models*. Wiley Online Library, 1972.

Ng, Andrew Y and Jordan, Michael I. On discriminative vs. generative classifiers: A comparison of logistic regression and naive bayes. In *Advances in neural information processing systems*, pp. 841–848, 2002.

Nicholls, AO. How to make biological surveys go further with generalised linear models. *Biological Conservation*, 50(1):51–75, 1989.

NORC. *General Social Survey*. http://www3.norc.org/GSS+Website/.

NORC. *Google Shopping Ad Product*. https://developers.google.com/adwords/api/docs/appendix/reports/shopping-performance-report#averagecpc.

Oliver, Margaret A and Webster, Richard. Kriging: a method of interpolation for geographical information systems. *International Journal of Geographical Information System*, 4(3):313–332, 1990.

Park, Yubin and Ghosh, Joydeep. A probabilistic imputation framework for predictive analysis using variably aggregated, multi-source healthcare data. In *Proceedings of the 2nd ACM SIGHIT International Health Informatics Symposium*, pp. 445–454. ACM, 2012.

Park, Yubin and Ghosh, Joydeep. Ludia an aggregate-constrained low-rank reconstruction algorithm to leverage publicly released health data. In *Proceedings of the 20th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 55–64. ACM, 2014.

Patrini, Giorgio, Nock, Richard, Caetano, Tiberio, and Rivera, Paul. (almost) no label no cry. In *Advances in Neural Information Processing Systems*, pp. 190–198, 2014.

Paul, Grégory, Cardinale, Janick, and Sbalzarini, Ivo F. Coupling image restoration and segmentation: a generalized linear model/bregman perspective. *International Journal of Computer Vision*, 104(1):69–93, 2013.

Peligrad, Magda and Wu, Wei Biao. Central limit theorem for fourier transforms of stationary processes. *The Annals of Probability*, pp. 2009–2022, 2010.

Perlich, Claudia, Dalessandro, Brian, Raeder, Troy, Stitelman, Ori, and Provost,

Foster. Machine learning for targeted display advertising: Transfer learning in action. *Machine learning*, 95(1):103–127, 2014.

Phillips, Peter CB et al. *Spectral regression for cointegrated time series*. Cowles Foundation for Research in Economics at Yales University, 1988.

Quadrianto, Novi, Smola, Alex J, Caetano, Tiberio S, and Le, Quoc V. Estimating labels from label proportions. *The Journal of Machine Learning Research*, 10: 2349–2374, 2009.

Rahimi, Ali and Recht, Benjamin. Random features for large-scale kernel machines. In *Advances in neural information processing systems*, pp. 1177–1184, 2008.

Rahimi, Ali and Recht, Benjamin. Weighted sums of random kitchen sinks: Replacing minimization with randomization in learning. In *Advances in neural information processing systems*, pp. 1313–1320, 2009.

Recht, Benjamin. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(Dec):3413–3430, 2011.

Recht, Benjamin, Fazel, Maryam, and Parrilo, Pablo A. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Recht, Benjamin, Re, Christopher, Wright, Stephen, and Niu, Feng. Hogwild: A lock-free approach to parallelizing stochastic gradient descent. In *Advances in neural information processing systems*, pp. 693–701, 2011.

Robert, Christian P and Casella, George. Monte carlo integration. In *Monte Carlo Statistical Methods*, pp. 71–138. Springer, 1999.

Robinson, William S. Ecological correlations and the behavior of individuals. *International journal of epidemiology*, 38(2):337–341, 2009.

Rosenbaum, Mathieu, Tsybakov, Alexandre B, et al. Improved matrix uncertainty selector. In *From Probability to Statistics and Back: High-Dimensional Models and Processes–A Festschrift in Honor of Jon A. Wellner*, pp. 276–290. Institute of Mathematical Statistics, 2013.

Rudelson, Mark and Zhou, Shuheng. High dimensional errors-in-variables models with dependent measurements. *arXiv preprint arXiv:1502.02355*, 2015.

Ruiz, Jesús M. On hilbert's 17th problem and real nullstellensatz for global analytic functions. *Mathematische Zeitschrift*, 190(3):447–454, 1985.

Schneps, Leila and Colmez, Coralie. *Math on trial: how numbers get used and abused in the courtroom*. Wiley Online Library, 2013.

Scott, David W. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, 1979a.

Scott, David W. On optimal and data-based histograms. *Biometrika*, 66(3):605–610, 1979b.

Smith, Winthrop W and Smith, Joanne M. Handbook of real-time fast fourier transforms. *IEEE, New York*, 1995.

Song, Lin, Langfelder, Peter, and Horvath, Steve. Random generalized linear model: a highly accurate and interpretable ensemble predictor. *BMC bioinformatics*, 14(1):5, 2013.

Stein, Michael L. *Interpolation of spatial data: some theory for kriging*. Springer Science & Business Media, 2012.

Strichartz, Robert S. *A guide to distribution theory and Fourier transforms*. World Scientific, 2003.

Tangirala, Arun K. *Principles of System Identification: Theory and Practice*. CRC Press, 2014.

Taylor, Stephen J. Modelling financial time series. 2007.

Terzioğlu, T. On schwartz spaces. *Mathematische Annalen*, 182(3):236–242, 1969.

TxID. Texas Inpatient Public Use Data File. *Texas Department of State Health Services*, 2014. https://www.dshs.state.tx.us/thcic/hospitals/Inpatientpudf.shtm.

University of California, Irvine. *Forest Fires Data Set*. https://archive.ics.uci.edu/ml/datasets/Forest+Fires.

US Department of Commerce. *Bureau of Economic Analysis*. http://www.bea.gov/.

US Department of Labour. *Bureau of Labour Statistics*. http://www.bls.gov/.

Wagner, Clifford H. Simpson's paradox in real life. *The American Statistician*, 36 (1):46–48, 1982.

Wagner, David. Resilient aggregation in sensor networks. In *Proceedings of the 2nd ACM workshop on Security of ad hoc and sensor networks*, pp. 78–87. ACM, 2004.

Wiener, Norbert. *Extrapolation, interpolation, and smoothing of stationary time series*, volume 2. MIT press Cambridge, MA, 1949.

Wilks, Samuel S. Mathematical statistics. *New York*, pp. 644, 1962.

Wu, Wei. Fourier transforms of stationary processes. *Proceedings of the American Mathematical Society*, 133(1):285–293, 2005.

Yan, Jun, Liu, Ning, Wang, Gang, Zhang, Wen, Jiang, Yun, and Chen, Zheng. How much can behavioral targeting help online advertising? In *Proceedings of the 18th international conference on World wide web*, pp. 261–270. ACM, 2009.

Yuan, Yong, Wang, Feiyue, Li, Juanjuan, and Qin, Rui. A survey on real time bidding advertising. In *Service Operations and Logistics, and Informatics (SOLI), 2014 IEEE International Conference on*, pp. 418–423. IEEE, 2014.

Zeff, Robbin Lee and Aronson, Bradley. *Advertising on the Internet*. John Wiley & Sons, Inc., 1999.

Zhao, Jerry, Govindan, Ramesh, and Estrin, Deborah. Computing aggregates for monitoring wireless sensor networks. In *Sensor Network Protocols and Applications, 2003*, pp. 139–148. IEEE, 2003.

Zhao, Peng and Yu, Bin. On model selection consistency of lasso. *The Journal of Machine Learning Research*, 7:2541–2563, 2006.